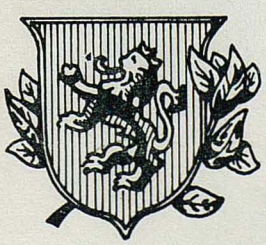
REVISTA
DE LA

ACADEMIA DE CIENCIAS

Exactas
Físicas
Químicas y
Naturales

DE
ZARAGOZA



Serie 2ª
Volumen 45

1990

INDICE DE MATERIAS

Rev. Acad. Ciencias. Zaragoza 45, (1990)

CASI-CONEXIONES RIEMANNIANAS

J. Peralta
Departamento de Algebra
Facultad de Ciencias Matemáticas
Universidad Complutense de Madrid
28040 - MADRID.

Abstract: In the present paper we introduce the concept of quasi-connection on the A-module M of vectors of a ring A by means of two equivalent ways, the quasi-derivations and the quasi-differentiations.

We define the torsion tensor and the curvature tensor of a quasi-connection and we prove that which verify the Bianchi identities.

An inner product on M is constructed and the metric quasi-connections and the Riemannian quasi-connections are studied. We conclude to proving that for every inversible (1,1) tensor H and for every inner product g on M, it exists an unique Riemannian quasi-connection associated to H and to g.

1. INTRODUCCION

Es bien sabido que las derivaciones en un álgebra tensorial adolecen - al contrario de lo que sucede con las derivadas - de unas estructuras de módulo y de álgebra de Lie para las operaciones definidas de modo natural en ellas.

Con objeto de subsanar tal dificultad, se han extendido estas nociones, por un lado a las pseudoconexiones y a las pseudoderivaciones y, por otro, a las casi-conexiones. Estas últimas suponen una particularización de las pseudoconexiones al caso en que el tensor de éstas sea no degenerado, y una generalización de las conexiones, que se obtendrían cuando se restringiese al tensor de Kronecker.

Del estudio de los conceptos anteriores venimos ocupándonos hace tiempo (Peralta [2], [3] y [4]), y últimamente, de modo especial, de las casi-conexiones (Peralta [5]).

Posiblemente las casi-conexiones hayan sido injustamente preteridas por ocupar un papel intermedio entre las conexiones y su máxima generalización, las pseudoderivaciones. Sin embargo, el hecho de que su tensor posea

inverso, las acerca en propiedades a las conexiones, y hace posible definir para ellas conceptos directamente extraídos de los de aquéllas, como toda la teoría concerniente a las formas de torsión y curvatura, que nos aparecen mucho más sofisticadas y de difícil solución en el caso de las pseudo-conexiones. Y lo mismo diríamos del problema de la métrica.

Como en los trabajos anteriormente reseñados, realizamos el problema en el A-módulo M de los vectores de un anillo A con una técnica similar a la usada en (Etayo [1]), que al concretarse en cada modelo particular, nos daría para él la correspondiente teoría; la más importante sería la que se refiere al módulo de los campos vectoriales sobre una variedad diferenciable.

* * *

En la sección 2 se establece el concepto de casi-conexión en un módulo partiendo de las nociones de casi-derivada y casi-derivación, y se indica otra vía paralela de llegar a aquéllas, que arranca de las ideas de casi-diferencial y casi-diferenciación.

En la sección 3 se introduce el tensor de torsión de una casi-conexión en un módulo, y en la 4, el tensor de curvatura.

Para poder estudiar relaciones entre ambos, se precisa ampliar la definición de casi-conexión a su actuación sobre tensores cualesquiera, lo que se hace en la sección siguiente. De alguna forma se sancionan favorablemente los conceptos presentados, ya que siguen siendo válidas las identidades de Bianchi.

Para finalizar, en la sección 6 se establece una definición de producto escalar en M, en analogía con el introducido en (Sikorski [6]), y se estudian las casi-conexiones métricas y Riemannianas. Se concluye demostrando la existencia y unicidad de una casi-conexión Riemanniana asociada a cada tensor (1,1) no degenerado H y a cada producto escalar g.

2. CASI-CONEXIONES EN UN MODULO

Sea A un anillo conmutativo y unitario. Un <u>vector</u> de A es una aplicación aditiva $X:A \longrightarrow A$ que cumple: $X(ab) = aX(b)+bX(a)$. Es trivial que se tiene, $X(1)=0$.

El conjunto V(A) de los vectores de A es un A-módulo y un álgebra de Lie para las operaciones usuales. En particular, el producto cruzado satis-

face:

$$[aX,bY] = ab[X,Y] - bY(a)X + aX(b)Y \quad ; \quad a,b \in A \quad ; \quad X,Y \in V(A).$$

Aunque las definiciones y muchos resultados que aparecen a continuación podrían extenderse a A-módulos cualesquiera, nos referiremos en lo sucesivo al A-módulo M=V(A). Denotaremos por $\mathfrak{T}^r_s$ al A-módulo de los tensores (r,s) de M.

Definición 2.1.- Llamaremos diferenciación exterior sobre M a la aplicación d que aumenta el grado de cada forma exterior sobre M en una unidad, y que está definida por:

$$(d_1)\ da(X) = X(a) \quad ; \quad a \in A,\ X \in M.$$

$$(d_2)\ d\omega(X_o,\ldots,X_r) = \sum_{i=o}^{r} (-1)^i X_i(\omega(X_o,\ldots,\hat{X}_i,\ldots,X_r)) +$$

$$+ \sum_{i<j} (-1)^{i+j} \omega([X_i,X_j],X_o,\ldots,\hat{X}_i,\ldots,\hat{X}_j,\ldots,X_r) \quad ; \quad \omega \in \mathfrak{T}^o_r,\ (X_o,\ldots,X_r) \in M^{r+1}.$$

Definición 2.2.- Sea $H \in \mathfrak{T}^1_1$ un tensor no degenerado y $K=H^{-1}$. El tensor H permite definir una diferenciación de grado uno ∂, sobre el álgebra de las formas exteriores sobre M (si $\omega \in \mathfrak{T}^o_r$, entonces $\partial\omega \in \mathfrak{T}^o_{r+1}$), que llamaremos diferenciación inducida por H, y que viene dada por:

$$(\partial_1)\ \partial a(X) = H(X)(a) \quad ; \quad a \in A,\ X \in M.$$

$$(\partial_2)\ \partial\omega(X_o,\ldots,X_r) = \sum_{i=o}^{r} (-1)^i H(X_i)(\omega(X_o,\ldots,\hat{X}_i,\ldots,X_r)) +$$

$$+ \sum_{i<j} (-1)^{i+j} \omega(K[H(X_i),H(X_j)],X_o,\ldots,\hat{X}_i,\ldots,\hat{X}_j,\ldots,X_r) \quad ; \quad \omega \in \mathfrak{T}^o_r,$$

$$(X_o,\ldots,X_r) \in M^{r+1}.$$

Proposición 2.3.- ∂ verifica las siguientes propiedades:

1) $\partial^2 a = 0$, $\forall a \in A$.

2) Si H es el tensor identidad, ∂ es la diferenciación exterior.

Demostración.-

1) $\partial^2 a(X,Y) = H(X)\partial a(Y) - H(Y)\partial a(X) - \partial a(K[H(X),H(Y)]) = H(X)H(Y)(a) -$
$- H(Y)H(X)(a) - H(K[H(X),H(Y)])(a) = 0.$

2) Es trivial. □

Definición 2.4.- Sea $H \in \mathfrak{T}^1_1$ un tensor no degenerado y $K=H^{-1}$. Para cada $X \in M$, llamaremos casi-derivada respecto de X de tensor H a toda aplicación

D_X: M $\longrightarrow$ M que verifique: (a) $D_X(Y+Z) = D_XY+D_XZ$, (b) $D_X(aY) = aD_XY +$ $+ H(X)(a)Y$; $a\in A$; $Y,Z\in M$.

En (Peralta [5]) pueden verse los siguientes resultados:

Proposición 2.5.- (i) Si D_X y D'_Y son dos casi-derivadas de tensores H y H', respectivamente, y es $K=H^{-1}$, $K'= H'^{-1}$; se cumple que $[D_X,D'_Y]$ es una casi-derivada de tensor H respecto de $K[H(X),H'(Y)]$ y de tensor H' respecto de $K'[H(X),H'(Y)]$.

(ii) En las hipótesis de (i), se tiene: $[aD_X,bD'_Y] = ab[D_X,D'_Y] +$ $+ aH(X)(b)D'_Y - bH'(Y)(a)D_X$; $a,b\in A$.

(iii) El conjunto $\mathfrak{D}(M;H)$ de casi-derivadas respecto de X, $\forall X\in M$, de tensor H, tiene estructura de A-módulo y de álgebra de Lie para las operaciones usuales.

Definición 2.6.- Una casi-derivación de tensor H es toda aplicación D: M $\longrightarrow \mathfrak{D}(M;H)$ que haga corresponder a cada $X\in M$ una casi-derivada de tensor H respecto de X, D_X. Si D es aditiva: (c) $D_{X+Y} = D_X+D_Y$, diremos que la casi-derivación es aditiva, y si además se cumple: (d) $D_{aX} = aD_X$, diremos que D es covariante o que es una casi-conexión de tensor H.

Así pues, una casi-conexión es una aplicación D: $X \longmapsto D_X$ que cumple (a), (b) (de la Def. 2.4), (c) y (d) (de la Def. 2.6). Así como esta definición se basa en la noción de casi-derivada, indicamos a continuación otra definición equivalente a partir del concepto de casi-diferencial.

Definición 2.7.- Sea $\mathcal{A}(M,M)$ el A-módulo de las aplicaciones de M en M. Para cada $X\in M$, la aplicación $\overline{D}X$: M $\longrightarrow$ M que hace corresponder a cada $Y\in M$ una casi-derivada de tensor H respecto de Y: $(\overline{D}X)(Y)=D_YX$, se llama casi-diferencial de tensor H del vector X. La aplicación $\overline{D}$: M $\longrightarrow \mathcal{A}(M,M)$, tal que $X \longmapsto \overline{D}X$, se llama casi-diferenciación de tensor H.

Proposición 2.8.- Si D es una casi-derivación de tensor H, su casi-diferenciación de tensor H, $\overline{D}$, verifica $\forall X,Y\in M$, $\forall a\in A$: (A) $\overline{D}(X+Y) = \overline{D}X+\overline{D}Y$; (B) $\overline{D}(aX) = a\overline{D}X + (X\otimes da)oH$; siendo d la diferenciación exterior. Esta última relación puede expresarse también: (B') $\overline{D}(aX) = a\overline{D}X + X\otimes\partial a$, donde ∂ es

la diferenciación de grado uno inducida por H.

Demostración.-

(A) es evidente.

$(a\overline{D}X + (X\hat{a}da)oH)Y = aD_YX + (X\hat{a}da)H(Y) = aD_YX + da(H(Y))X = aD_YX +$

$+ H(Y)(a)X = D_Y(aX) = \overline{D}(aX)Y$, con lo que se cumple (B).

(B´) se verifica en virtud de que $H(Y)(a) = \partial a(Y)$. □

Es sencillo probar las dos proposiciones siguientes.

Proposición 2.9.- Si una casi-derivación D de tensor H es aditivá, su casi-diferenciación satisface: (C) $(\overline{D}X)(Y+Z) = (\overline{D}X)Y+(\overline{D}X)Z$. Diremos entonces que $\overline{D}$ es aditiva.

Si D es covariante, se cumple además: (D) $(\overline{D}X)(aY) = a(\overline{D}X)Y$, en cuyo caso diremos que $\overline{D}$ es una casi-diferenciación covariante de tensor H.

Proposición 2.10.- Se puede definir una casi-conexión de tensor H, indistintamente, mediante una casi-derivación de tensor H, $D:M \longrightarrow \mathfrak{D}(M;H)$ que satisface los axiomas (a), (b), (c) y (d) o mediante una casi-diferenciación de tensor H, $\overline{D}:M \longrightarrow \mathfrak{T}^1_1$ que verifica (A), (B), (C) y (D). Ambas están relacionadas por $D_YX=(\overline{D}X)Y$.

En adelante, supondremos que todas las casi-derivadas, casi-derivaciones, casi-conexiones, casi-diferenciales y casi-diferenciaciones que vamos a usar son del mismo tensor H. Por tal motivo, cuando no exista posibilidad de equívoco, omitiremos hacer referencia al tensor y diremos simplemente, casi-derivada respecto de X, en vez de casi-derivada de tensor H respecto de X, etc.

3. TENSOR DE TORSION DE UNA CASI-CONEXION

Proposición 3.1.- Si D es una casi-conexión, la aplicación

$$\left.\begin{array}{l} T: M^2 \longrightarrow M \\ (X,Y) \longmapsto T(X,Y) = D_XY - D_YX - K[H(X),H(Y)] \end{array}\right\} \quad (1)$$

es un tensor (1,2) antisimétrico.

Demostración.-

A la biaditividad y antisimetría se llega sin dificultad.

$T(aX,Y) = D_{aX}Y - D_Y(aX) - K[aH(X),H(Y)] = aD_XY - aD_YX - H(Y)(a)X - aK[H(X),H(Y)] +$
$+ H(Y)(a)KH(X) = aT(X,Y)$.

$T(X,aY) = -T(aY,X) = -aT(Y,X) = aT(X,Y)$. □

Definición 3.2.- La aplicación T definida por (1) se llama tensor de torsión de la casi-conexión D.

Proposición 3.3.- (i) Si $S \in \mathfrak{T}_2^1$, D es una casi-derivación, y definimos $D'_XY = D_XY + S(X,Y)$, $\forall (X,Y) \in M^2$ (2), se verifica que D' es una casi-derivación. Recíprocamente, si D y D' son dos casi-derivaciones, existe un único $S \in \mathfrak{T}_2^1$ tal que $D'_XY = D_XY + S(X,Y)$, $\forall (X,Y) \in M^2$.

(ii) Si $S \in \mathfrak{T}_2^1$, D es una casi-conexión, y se define D' como en (2), entonces D' también es una casi-conexión. Además, si T y T' son los tensores de torsión de D y D', respectivamente, se tiene: $T'(X,Y) = T(X,Y) + S(X,Y) - S(Y,X)$.

Demostración.-

Es fácil probar que si D es una casi-derivación o una casi-conexión, también D' lo es.

Si D y D' son casi-derivaciones, basta con definir $S(X,Y) = D'_XY - D_XY$, $\forall (X,Y) \in M^2$ y demostrar que $S: M^2 \longrightarrow M$ es A-lineal.

Por último, $T'(X,Y) = D'_XY - D'_YX - K[H(X),H(Y)] = D_XY + S(X,Y) - D_YX -$
$- S(Y,X) - K[H(X),H(Y)] = T(X,Y) + S(X,Y) - S(Y,X)$. □

Definición 3.4.- Una casi-conexión es simétrica si su tensor de torsión es igual a cero.

Veamos que si imponemos una condición adicional al anillo conmutativo y unitario A, podremos construir una casi-conexión simétrica a partir de una casi-conexión cualquiera dada.

Definición 3.5.- El anillo A diremos que es dúplice si cualquiera que sea $a \in A$, existe un único $b \in A$, tal que $a = 2b$. En tal caso, escribiremos $b = \frac{1}{2}a$.

Proposición 3.6.- Supongamos que el anillo A es dúplice. Si D es una casi-conexión sobre el A-módulo M con tensor de torsión T, y construimos $\tilde{D}_XY = D_XY - \frac{1}{2}T(X,Y)$, $\forall (X,Y) \in M^2$, se cumple que $\tilde{D}$ es una casi-conexión simé-

trica.

Demostración.-

Es inmediato probar que $\tilde{D}$ cumple los axiomas (a), (c) y (d) de las casi-conexiones. Y también (b): $\tilde{D}_X(aY) = D_X(aY) - \frac{1}{2}T(aX,Y) = aD_XY + H(X)(a)Y - \frac{1}{2}aT(X,Y) = a\tilde{D}_XY + H(X)(a)Y$.

Su tensor de torsión $\tilde{T}$ verifica: $\tilde{T}(X,Y) = \tilde{D}_XY - \tilde{D}_YX - K[H(X),H(Y)] = D_XY - \frac{1}{2}T(X,Y) - D_YX + \frac{1}{2}T(Y,X) - K[H(X),H(Y)] = D_XY - D_YX - T(X,Y) - K[H(X),H(Y)] = 0$, $\forall (X,Y)\in M^2$. □

4. TENSOR DE CURVATURA DE UNA CASI-CONEXION

Proposición 4.1.- Si D es una casi-conexión, la aplicación de M en M definida como $R(X,Y) = [D_X,D_Y] - D_{K[H(X),H(Y)]}$, $\forall (X,Y)\in M^2$ es A-lineal.

Demostración.-

La aditividad no ofrece dificultades.

$R(X,Y)(aZ) = D_X(aD_YZ + H(Y)(a)Z) - D_Y(aD_XZ + H(X)(a)Z) - aD_{K[H(X),H(Y)]}Z - HK[H(X),H(Y)](a)Z = aD_XD_YZ + H(X)(a)D_YZ + H(Y)(a)D_XZ + H(X)H(Y)(a)Z - aD_YD_XZ - H(Y)(a)D_XZ - H(X)(a)D_YZ - H(Y)H(X)(a)Z - aD_{K[H(X),H(Y)]}Z - [H(X),H(Y)](a)Z = a[D_X,D_Y]Z + [H(X),H(Y)](a)Z - aD_{K[H(X),H(Y)]}Z - [H(X),H(Y)](a)Z = aR(X,Y)Z$. □

Análogamente se demuestra la siguiente

Proposición 4.2.- (i) La aplicación $R: M^2 \longrightarrow \mathfrak{T}_1^1$ tal que a cada (X,Y) le hace corresponder R(X,Y) es bilineal y antisimétrica.

(ii) R también puede considerarse como el tensor (1,3) siguiente:

$$\left.\begin{array}{l} R: M^3 \longrightarrow M \\ (X,Y,Z) \longmapsto R(X,Y,Z) = R(X,Y)Z \end{array}\right\}$$

Definición 4.3.- El tensor $R\in\mathfrak{T}_3^1$ definido como $R(X,Y,Z) = [D_X,D_Y]Z - D_{K[H(X),H(Y)]}Z$, $\forall X,Y,Z\in M$, se llama tensor de curvatura de la casi-conexión.

Proposición 4.4.- Si T y R son los tensores de torsión y de curvatura de una casi-conexión D, se verifica:

$$S_{cícl.}\{R(X,Y)Z\} = S_{cícl.}\{D_XT(Y,Z) + T(X,K[H(Y),H(Z)])\},\ \forall X,Y,Z\in M,$$

donde $S_{cícl.}$ denota la suma cíclica con respecto a X,Y,Z.

Demostración.-

$S_{cícl.}\{R(X,Y)Z\} = S_{cícl.}\{[D_X,D_Y]Z - D_{K[H(X),H(Y)]}Z\} = D_XD_YZ - D_YD_XZ +$
$+ D_YD_ZX - D_ZD_YX + D_ZD_XY - D_XD_ZY - D_{K[H(X),H(Y)]}Z - D_{K[H(Y),H(Z)]}X -$
$- D_{K[H(Z),H(X)]}Y = S_{cícl.}\{D_X(D_YZ-D_ZY) - D_{K[H(Y),H(Z)]}X\} = S_{cícl.}\{D_X(D_YZ-D_ZY-$
$-K[H(Y),H(Z)]) + D_XK[H(Y),H(Z)] - D_{K[H(Y),H(Z)]}X\} = S_{cícl.}\{D_XT(Y,Z) +$
$+ D_XK[H(Y),H(Z)] - D_{K[H(Y),H(Z)]}X\} = S_{cícl.}\{D_XT(Y,Z) + (D_XK[H(Y),H(Z)] -$
$- D_{K[H(Y),H(Z)]}X - K[H(X),HK[K(Y),K(Z)]]) + K[H(X),[H(Y),H(Z)]]\} =$
$= S_{cícl.}\{D_XT(Y,Z) + T(X,K[H(Y),H(Z)]) + K[H(X),[H(Y),H(Z)]]\} =$
$= S_{cícl.}\{D_XT(Y,Z) + T(X,K[H(Y),H(Z)])\}$; ya que por la Prop. 2.5 (iii), el conjunto $\mathfrak{D}(M;H)$ tiene estructura de álgebra de Lie y cumplirá la identidad de Jacobi: $S_{cícl.}\{K[H(X),[H(Y),H(Z)]]\} = 0$. □

5. CASI-CONEXIONES TENSORIALES

Definición 5.1.- Sea D_X una casi-derivada respecto de X, y $\mathcal{T} = \bigoplus_{r,s=o}^{\infty} \mathcal{T}^r_s$. Llamaremos casi-derivada tensorial respecto de X, extensión de la casi-derivada D_X, a una aplicación de $\mathcal{T}$ en $\mathcal{T}$, que seguiremos denotando por D_X, que verifique los siguientes axiomas:

(i) $D_X(\mathcal{T}^r_s) \subset \mathcal{T}^r_s$. $\forall r,s$.

(ii) $D_Xa = H(X)(a) = \partial a(X)$, $\forall a \in A$.

(iii) $D_X \circ c = c \circ D_X$, cualquiera que sea la contracción c.

(iv) $D_X(S \otimes S') = (D_XS) \otimes S' + S \otimes (D_XS')$, $\forall S,S' \in \mathcal{T}$.

(v) La casi-derivada D_X al actuar sobre cada $Y \in M$ coincide con la casi-derivada del vector Y, D_XY.

Denotaremos por $\mathfrak{D}(\mathcal{T};H)$ al conjunto de las casi-derivadas tensoriales de tensor H.

Definición 5.2.- Si D_X es una casi-derivada tensorial, llamaremos casi-derivación tensorial a la aplicación $D:M \longrightarrow \mathfrak{D}(\mathcal{T};H)$ tal que $X \longmapsto D_X$. Si la casi-derivación tensorial es extensión de una casi-derivación covariante, diremos que es una casi-derivación tensorial covariante o una casi-conexión tensorial.

Es sencillo probar la siguiente

Proposición 5.3.- Si D es una casi-derivación tensorial, $S\in\mathcal{T}^r_s$ y se define la aplicación $\overline{DS} : M^{s+1} \longrightarrow \mathcal{T}^r_o$

$$\left.\begin{array}{l}(X_1,\ldots,X_s,Y) \longmapsto (\overline{DS})(X_1,\ldots,X_s,Y)=(D_Y S)(X_1,\ldots,X_s)\end{array}\right\} \quad (3),$$

entonces $\overline{DS}$ pertenece al conjunto $\mathcal{L}(M^{s+1},\mathcal{T}^r_o)$ de las aplicaciones (s+1)-lineales de M^{s+1} en $\mathcal{T}^r_o$. (Se entiende que si s=0, $(\overline{DS})(Y)=D_Y S$).

Definición 5.4.- Si $S\in\mathcal{T}^r_s$, se llama casi-diferencial tensorial de S a la aplicación $\overline{DS}\in\mathcal{T}^r_{s+1}$ definida en (3), y casi-diferenciación tensorial a la aplicación $\overline{D}:\mathcal{T}\longrightarrow\mathcal{T}$ que hace corresponder a cada tensor S su casi-diferencial tensorial. Se dice que la casi-diferenciación tensorial $\overline{D}$ es covariante, si lo es la casi-derivación tensorial D.

Para no complicar la terminología, a las casi-derivadas tensoriales, casi-diferenciales tensoriales, etc., las llamaremos simplemente casi-derivadas, casi-diferenciales, etc., siempre que la omisión no dé lugar a confusión.

Proposición 5.5.- Sea D una casi-derivación tensorial y $\overline{D}$ su casi-diferenciación. Si $S\in\mathcal{T}^r_s$ y $(X_1,\ldots,X_s)\in M^{s+1}$, se cumple:

$$(\overline{DS})(X_1,\ldots,X_s,Y)=(D_Y S)(X_1,\ldots,X_s)=D_Y(S(X_1,\ldots,X_s))-\sum_{i=1}^{s} S(X_1,\ldots,D_Y X_i,\ldots,X_s).$$

Demostración.-

$$D_Y(S(X_1,\ldots,X_s))=D_Y(c_1\ldots c_s(X_1\otimes\ldots\otimes X_s\otimes S))=c_1\ldots c_s(D_Y(X_1\otimes\ldots\otimes X_s\otimes S))=$$

$$= c_1\ldots c_s\Big(\sum_{i=1}^{s} X_1\otimes\ldots\otimes D_Y X_i\otimes\ldots\otimes X_s\otimes S + X_1\otimes\ldots\otimes X_s\otimes D_Y S\Big)= \sum_{i=1}^{s} S(X_1,\ldots,D_Y X_i,\ldots,X_s)+$$

$$+ (D_Y S)(X_1,\ldots,X_s). \square$$

En las siguientes proposiciones se obtienen fórmulas relativas a la torsión y la curvatura de una casi-conexión. En concreto, se probará que la torsión y la curvatura de una casi-conexión en el A-módulo M cumplen, al igual que en el caso de una conexión lineal sobre una variedad diferenciable, las identidades de Bianchi.

Proposición 5.6 (Primera identidad de Bianchi).- Si D es una casi-conexión y T y R sus tensores de torsión y de curvatura, se deduce que:

$$S_{cícl.}\{R(X,Y)Z\} = S_{cícl.}\{T(T(X,Y),Z)+(D_X T)(Y,Z)\}, \forall X,Y,Z\in M \quad (4).$$

En particular, si D es simétrica, se tiene: $S_{cícl.}\{R(X,Y)Z\}=0$.

Demostración.-

$S_{cícl.}\{T(T(X,Y),Z)+(D_XT)(Y,Z)\}=S_{cícl.}\{D_{T(X,Y)}Z-D_ZT(X,Y)-K[HT(X,Y),H(Z)]+$
$+D_XT(Y,Z)-T(D_XY,Z)-T(Y,D_XZ)\}=S_{cícl.}\{D_{T(X,Y)}Z-K[HT(X,Y),H(Z)]-T(D_XY,Z)-$
$-T(Y,D_XZ)\}=S_{cícl.}\{D_{D_XY}Z-D_{D_YX}Z-D_{K[H(X),H(Y)]}Z-K[H(D_XY),H(Z)]+K[H(D_YX),H(Z)]+$
$+K[[H(X),H(Y)],H(Z)]-D_{D_XY}Z+D_ZD_XY+K[H(D_XY),H(Z)]-D_YD_XZ+D_{D_XZ}Y+K[H(Y),H(D_XZ)]$.

Reduciendo términos y teniendo en cuenta que $S_{cícl.}\{K[[H(X),H(Y)],H(Z)]\}=$
$=0$, se tiene: $S_{cícl.}\{T(T(X,Y),Z)+(D_XT)(Y,Z)\}=S_{cícl.}\{D_ZD_XY-D_YD_XZ-$
$-D_{K[H(X),H(Y)]}Z\}=S_{cícl.}\{D_XD_YZ-D_ZD_YX-D_{K[H(Y),H(Z)]}X\}=S_{cícl.}\{R(X,Y)Z\}$.

Si D es simétrica, T=0, y se anula el segundo miembro de (4). □

Proposición 5.7.- Si D es una casi-conexión de tensor de curvatura R, se cumple: $S_{cícl.}\{D_XR(Y,Z)+R(X,K[H(Y),H(Z)])\}=0$, $\forall X,Y,Z\in M$ (5).

Demostración.-

$(D_XR(Y,Z))V=D_X(R(Y,Z)V)-R(Y,Z)(D_XV)=D_X[D_Y,D_Z]V-D_XD_{K[H(Y),H(Z)]}V-$
$-[D_Y,D_Z]D_XV+D_{K[H(Y),H(Z)]}D_XV$, $\forall V,X,Y,Z\in M$, luego $D_XR(Y,Z)=[D_X,[D_Y,D_Z]]-$
$-[D_X,D_{K[H(Y),H(Z)]}]$.

Por lo tanto, $S_{cícl.}\{D_XR(Y,Z)+R(X,K[H(Y),H(Z)])\}=S_{cícl.}\{[D_X,[D_Y,D_Z]]-$
$-[D_X,D_{K[H(Y),H(Z)]}]+[D_X,D_{K[H(Y),H(Z)]}]-D_{K[H(X),HK[H(Y),H(Z)]]}=$
$=S_{cícl.}\{[D_X,[D_Y,D_Z]]-D_{K[H(X),[H(Y),H(Z)]]}\}=0$, ya que $S_{cícl.}\{[D_X,[D_Y,D_Z]]\}=0$
y $S_{cícl.}\{D_{K[H(X),[H(Y),H(Z)]]}\}=D_{S_{cícl.}\{K[H(X),[H(Y),H(Z)]]\}}=0$. □

Proposición 5.8 (Segunda identidad de Bianchi).-

Si T y R son los tensores de torsión y de curvatura de una casi-conexión D, se verifica: $S_{cícl.}\{(D_XR)(Y,Z)+R(T(X,Y),Z)\}=0$, $\forall X,Y,Z\in M$.

En particular, si D es simétrica, se tiene: $S_{cícl.}\{(D_XR)(Y,Z)\}=0$.

Demostración.-

Como $R\in\mathfrak{T}_3^1$, si $V,X,Y,Z\in M$ se tiene:

$$(D_XR)(Y,Z)V=D_X(R(Y,Z)V)-R(D_XY,Z)V-R(Y,D_XZ)V-R(Y,Z)D_XV \quad (6).$$

Ahora bien, como $R(Y,Z)\in\mathfrak{T}_1^1$, su casi-derivada respecto de X será:
$(D_XR(Y,Z))V=D_X(R(Y,Z)V)-R(Y,Z)D_XV$; luego $R(Y,Z)D_XV=D_X(R(Y,Z)V)-(D_XR(Y,Z))V$,
y sustituyendo en (6): $(D_XR)(Y,Z)=D_XR(Y,Z)-R(D_XY,Z)-R(Y,D_XZ)$.

De esta última relación y de (5) se tiene: $S_{cícl.}\{(D_XR)(Y,Z)\}=$
$=S_{cícl.}\{-R(X,K[H(Y),H(Z)])-R(D_XY,Z)-R(Y,D_XZ)\}=S_{cícl.}\{-R(X,K[H(Y),H(Z)])+$

$+R(Z,D_XY)-R(Y,D_XZ)\}=S_{cícl.}\{-R(X,K[H(Y),H(Z)])+R(X,D_YZ)-R(X,D_ZY)\}=$
$=S_{cícl.}\{R(X,T(Y,Z))\}=S_{cícl.}\{R(Z,T(X,Y))\}=-S_{cícl.}\{R(T(X,Y),Z)\}$. □

6. CASI-CONEXIONES RIEMANNIANAS

Antes de pasar a la definición de producto escalar sobre el A-módulo M, veamos una cuestión relativa a las formas de grado dos.

Proposición 6.1.- Si D es una casi-conexión, T su tensor de torsión y $g\in\mathcal{T}_2^0$ una forma de grado dos simétrica, se cumple:

$2g(D_XY,Z)=H(X)(g(Y,Z))+H(Y)(g(Z,X))-H(Z)(g(X,Y))-(D_Xg)(Y,Z)-(D_Yg)(Z,X)+(D_Zg)(X,Y)+g(K[H(X),H(Y)],Z)-g(K[H(Y),H(Z)],X)+g(K[H(Z),H(X)],Y)+g(T(X,Y),Z)-g(T(Y,Z),Z)+g(T(Z,X),Y)$, $\forall X,Y,Z\in M$ (7).

En particular, si la casi-conexión es simétrica, los tres últimos sumandos del segundo miembro de (7) son nulos.

Demostración.-

$-(D_Xg)(Y,Z)-(D_Yg)(Z,X)+(D_Zg)(X,Y)=-D_X(g(Y,Z))+g(D_XY,Z)+g(Y,D_XZ)-D_Y(g(Z,X))+g(D_YZ,X)+g(Z,D_YX)+D_Z(g(X,Y)-g(D_ZX,Y)-g(X,D_ZY)=-H(X)(g(Y,Z))-H(Y)(g(Z,X))+H(Z)(g(X,Y))+g(D_XY,Z)+g(Y,D_XZ)+g(D_YZ,X)+g(Z,D_YX)-g(D_ZX,Y)-g(X,D_ZY)=g(D_XY,Z)+g(Y,D_XZ)+g(D_YZ,X)+g(Z,D_YX)-g(D_ZX,Y)-g(X,D_ZY)+g(K[H(X),H(Y)],Z)-g(K[H(Y),H(Z)],X)+g(K[H(Z),H(X)],Y)+g(T(X,Y),Z)-g(T(Y,Z),X)+g(T(Z,X),Y)$.

Pero como g es simétrica, resulta:

$g(D_XY+D_YX+K[H(X),H(Y)],Z)+g(D_XZ-D_ZX-K[H(X),H(Z)],Y)+g(D_YZ-D_ZY-K[H(Y),H(Z)],X)+g(T(X,Y),Z)-g(T(Y,Z),X)+g(T(Z,X),Y)=g(D_XY+D_YX+K[H(X),H(Y)]+T(X,Y),Z)+g(D_XZ-D_ZX-K[H(X),H(Z)]+T(Z,X),Y)=g(2D_XY,Z)=2g(D_XY,Z)$. □

Si $g\in\mathcal{L}(M^2,A)$, para cada X M se puede definir la aplicación $g_X:M\longrightarrow A$, tal que $Y\longmapsto g_X(Y)=g(X,Y)$, que evidentemente es A-lineal. Por lo tanto, si denotamos por M^* el espacio dual de M, es posible asociar a cada $g\in\mathcal{L}(M^2,A)$ la aplicación A-lineal $\tilde{g}:M\longrightarrow M^*$, tal que $X\longmapsto g(X)=g_X$.

Definición 6.2.- Llamaremos producto escalar en el A-módulo M a una aplicación bilineal $g:M^2\longrightarrow A$ que verifique que es simétrica, y que para cada $\omega\in M^*$ exista un único $X\in M$ tal que $g(X,Y)=\omega(Y)$, $\forall Y\in M$.

Es sencillo probar la siguiente

Proposición 6.3.- Para cada forma A-bilineal simétrica $g:M^2 \longrightarrow A$, las dos condiciones siguientes son equivalentes:

1) g es un producto escalar.

2) La aplicación $\tilde{g}:M \longrightarrow M^*$ asociada a g es biyectiva (en cuyo caso, será un isomorfismo de A-módulos).

En consecuencia, si g es un producto escalar en M, y para cada $Y\in M$ se conoce g(X,Y), entonces el vector $X\in M$ queda unívocamente determinado; es decir, si $g(X,Y)=g(X',Y)$, $\forall Y\in M$, se deduce necesariamente que $X=X'$.

Definición 6.4.- Diremos que una casi-conexión D es una casi-conexión métrica si hay definido un producto escalar g sobre M, tal que $D_X g=0$, $\forall X\in M$. En tal caso, escribiremos Dg=0. Diremos que una casi-conexión D es una casi-conexión Riemanniana si es métrica y simétrica.

Proposición 6.5.- Si A es un anillo dúplice, para cada tensor no degenerado $H\in\mathcal{T}^1_1$ y para cada producto escalar g sobre el A-módulo M, existe una casi-conexión Riemanniana de tensor H cuyo producto escalar es g.

Demostración.-

- Existencia:

Para cada $X,Y\in M$ definimos

$$g(D_X Y,Z)=\frac{1}{2}\{H(X)(g(Y,Z))+H(Y)(g(Z,X))-H(Z)(g(X,Y))+g(K[H(X),H(Y)],Z)-$$
$$-g(K[H(Y),H(Z)],X)+g(K[H(Z),H(X)],Y)\}\ ,\ \forall Z\in M \qquad (8).$$

De esta forma, $\forall X,Y\in M$ queda definido de forma única $D_X Y$.

Veamos que D es una casi-conexión:

Evidentemente se cumplen los axiomas (a) y (c), ya que $g(D_X(Y+Y'),Z)=$
$=g(D_X Y+D_X Y',Z)$ y $g(D_{X+X'}Y,Z)=g(D_X Y+D_{X'}Y,Z)$, $\forall Z\in M$.

$$g(D_X(aY),Z)=\frac{1}{2}\{H(X)(ag(Y,Z))+aH(Y)(g(Z,X))-H(Z)(ag(X,Y))+$$
$$+g(K(a[H(X),H(Y)]+H(X)(a)H(Y)),Z)-g(K(a[H(Y),H(Z)]-H(Z)(a)H(Y)),X)+$$
$$+ag(K[H(Z),H(X)],Y)\}=\frac{1}{2}\{aH(X)(g(Y,Z))+H(X)(a)g(Y,Z)+aH(Y)(g(Z,X))-$$
$$-aH(Z)(g(X,Y))-H(Z)(a)g(X,Y)+ag(K[H(X),H(Y)],Z)+H(X)(a)g(Y,Z)-$$
$$-ag(K[H(Y),H(Z)],X)+H(Z)(a)g(Y,X)+ag(K[H(Z),H(X)],Y)\}=a\frac{1}{2}\{H(X)(g(Y,Z))+$$
$$+H(Y)(g(Z,X))-H(Z)(g(X,Y))+g(K[H(X),H(Y)],Z)-g(K[H(Y),H(Z)],X)+$$
$$+g(K[H(Z),H(X)],Y)\}+H(X)(a)g(Y,Z)=ag(D_X Y,Z)+H(X)(a)g(Y,Z).$$

Por tanto, $g(D_X(aY),Z)=g(aD_XY+H(X)(a)Y,Z)$, $\forall Z\in M$, con lo que se verifica (b). Y de forma parecida se prueba (d).

La casi-conexión D es simétrica:

$g(T(X,Y),Z)=g(D_XY,Z)-g(D_YX,Z)-g(K[H(X),H(Y)],Z)=\frac{1}{2}\{H(X)(g(Y,Z))+$
$+H(Y)(g(Z,X))-H(Z)(g(X,Y))+g(K[H(X),H(Y)],Z)-g(K[H(Y),H(Z)],X)+$
$+g(K[H(Z),H(X)],Y)-H(Y)(g(X,Z))-H(X)(g(Z,Y))+H(Z)(g(Y,X))-$
$-g(K[H(Y),H(X)],Z)+g(K[H(X),H(Z)],Y)-g(K[H(Z),H(Y)],X)-g(K[H(X),H(Y)],Z)\}=0$,
$\forall Z\in M$. Por lo tanto, $\tilde{g}(T(X,Y))(Z)=0$, $\forall Z\in M$; y como $\tilde{g}$ es isomorfismo, $T=0$.

Por último, D es métrica:

$(D_Xg)(Y,Z)=D_X(g(Y,Z))-g(D_XY,Z)-g(Y,D_XZ)=H(X)(g(Y,Z))-g(D_XY,Z)-$
$-g(D_XZ,Y)=H(X)(g(Y,Z))-\frac{1}{2}\{H(X)(g(Y,Z))+H(Y)(g(Z,X)-H(Z)(g(X,Y))+$
$+g(K[H(X),H(Y)],Z)-g(K[H(Y),H(Z)],X)+g(K[H(Z),H(X)],Y)\}-\frac{1}{2}\{H(X)(g(Z,Y))+$
$+H(Z)(g(Y,X))-H(Y)(g(X,Z))+g(K[H(X),H(Z)],Y)-g(K[H(Z),H(Y)],X)+$
$+g(K[H(Y),H(X)],Z)\}=0$, $\forall X,Y,Z\in M$.

- Unicidad:

Toda casi-conexión Riemanniana D de tensor H y cuyo producto escalar es g, viene definida por (8). En efecto, por ser simétrica cumplirá (7) con los tres últimos sumandos del segundo miembro iguales a cero, y por ser métrica: $(D_Xg)(Y,Z)=(D_Xg)(Z,X)=(D_Zg)(X,Y)=0$. Llevando estas igualdades a (7), se llega a (8). □

En la demostración de la Proposición 6.5 se ha obtenido el siguiente

Corolario 6.6.- Sea A un anillo dúplice. Toda casi-conexión Riemanniana D de tensor H y producto escalar g, definida sobre el A-módulo M, cumple (8).

BIBLIOGRAFIA

[1] ETAYO, J.J. "Pseudoderivaciones". Revista Matemática Hispano- Americana. 4ª serie, 35 (1975), 81-98.

[2] PERALTA, J. "On the graduated derivatives". Coll. Math. 35, Fascículo 2°, (1984), 189-205.

[3] PERALTA, J. "Derivaciones, pseudoderivaciones y casi-derivaciones de grado superior y su comportamiento algebraico". Tesis. Edit. Univ. Compl. 75 (1985).

[4] PERALTA, J. "Pseudoderivadas primarias y de Bompiani". X Jornadas

Hispano-Lusas. Sección V. Univ. Murcia (1985), 1-9.

[5] PERALTA, J. "Casi-conexiones en un módulo y algunas formas de generarlas". XIV Jornadas Hispano-Lusas, Univ. de la Laguna (1989). En prensa.

[6] SIKORSKI, R. "Abstract covariant derivative". Colloquium mathematicum. XVIII (1967), 251-272.

Rev. Acad. Ciencias. Zaragoza 45, (1990)

A MESH INDEPENDENCE PRINCIPLE FOR NONLINEAR EQUATIONS USING NEWTON'S METHOD AND NONLINEAR PROJECTIONS.

by

IOANNIS K. ARGYROS
Department of Mathematics
New Mexico State University
Las Cruces, NM 88003

Abstract. We consider the nonlinear operator equation in a Banach space. We make use of nonlinear projections on finite dimensional spaces to produce the finite dimensional discretization of the nonlinear equation. Using Newton's method we then prove the mesh-independence principle for this problem. Our results cover and extend previous results involving linear projections on finite dimensional spaces.

Key words and phrases: Newton's method, mesh independence, nonlinear equation.

A.M.S. (1980) classification codes: 65J15, 65B05.

Introduction. In this paper we extend the validity of the mesh independence principle for discretizations of operators. We consider the equation

$$F(z) = 0 \qquad (1)$$

where F is a nonlinear operator on a Banach space X. Newton's iteration for (1) is given by

$$z_{n+1} = z_n - F'(z_n)^{-1}F(z_n) . \qquad (2)$$

The above iteration, under certain assumptions, converges quadratically to a solution z^* of (1). However, it is not at all easy in general to compute the iterates in (2). That is why we consider the discretized family of equations

$$T_h(x) = 0 \qquad (3)$$

indexed by some $h > 0$ to solve (1), where T_h is a nonlinear operator on a finite dimensional space X_h. Under certain assumptions we show that the equations (3) have solutions

$$x_h^* = P_h(z^*) + O(h^p)$$

which are the limit of the Newton sequence applied to (3) such that:

$$x_0^h = P_h(z_0),\ x_{n+1}^h = x_n^h - T_h'(x_n^h)^{-1}T_h(x_n^h),\quad n = 0, 1, 2, \cdots . \qquad (4)$$

To achieve this, we define the discretization on X by the bounded nonlinear operators $P_h : X \to X_h$.

It has been observed in many computations that for sufficiently small h there is at most a difference of one between the number of steps required by the iterations (2) and (4) to converge to within a given tolerance $\epsilon > 0$. This is one aspect of the mesh-independence principle for Newton's method. Here we actually show that the number of steps in both iterations is the same.

The above results have already been proved in [3] but for linear projections from X to X_h. For special classes of boundary value problems for mesh-independence principle was proved in [1], [5]. It was used to construct certain mesh-refinement strategies.

Most of our results carry over immediately to Newton like methods [7].

I. Preliminaries

The norms in all spaces will be denoted by the same symbol $\| \ \|$.

Let $F : D \subset X \to X$ be a nonlinear operator with Lipschitz continuous Fréchet derivative on the open domain, that is

$$\|F'(x) - F'(y)\| \leq \gamma\|x - y\|, \; x, y \in D, \; \gamma > 0. \tag{5}$$

Let us assume that (1) has a simple solution and set $d = \|F'(z^*)^{-1}\|$.

We will need the following theorem [8].

Theorem 1. Let $F : D \subset X \to X$ satisfy the stated hypotheses and with

$$r^* = \frac{2}{3d\gamma}, \tag{6}$$

suppose that $U = U(z^*, r^*) = \{z \in X \mid \|z - z^*\| < r^*\}$. Then for any $z_0 \in U$, iteration (2) converges to z^* and the iterates satisfy

$$\|z_{n+1} - z^*\| \leq \frac{d\gamma\|z_n - z^*\|^2}{2(1 - d\gamma\|z_n - z^*\|)}, \; n = 0, 1, 2, \cdots . \tag{7}$$

Consider the family

$$\{T_h, P_h\}, \; h > 0 \tag{8}$$

where

$$T_h : D_h \subset X_h \to X_h, \; h > 0$$

are nonlinear operators and

$$P_h : X \to X_h, \; h > 0$$

are nonlinear operators which are twice continuously Fréchet- differentiable at least in some ball $U(z^*, \rho)$ for some $\rho > 0$ and $P_h(U) \subset D_h$.

The discretization (8) is called Lipschitz uniform if there exist scalars $\rho > 0$, $L > 0$ such that

$$\bar{U}(P_h(z^*), \rho) \subset D_h, \; h > 0 \tag{9}$$

and

$$\|T_h'(w) - T_h'(v)\| \leq L\|w - v\|, \; h > 0, \; w, v \in \bar{U}(P_h(z^*), \rho) . \tag{10}$$

Moreover, the family (8) is called "bounded" if there exist constants $q_1, q_2 > 0$ such that

$$\|P_h(u) - P_h(z^*)\| \leq q_1 \|u - z^*\| + q_2\|u - z^*\|^2, \; u \in U . \qquad (11)$$

The motivation for the above condition is due to the identity

$$P_h(u) = P_h(z^*) + P_h'(z^*)(u - z^*) + \frac{1}{2}P_h''(\bar{z}^*)(u - z^*)^2 , \qquad (12)$$

where P_h', P_h'' represent the first and second Fréchet derivatives of P_h and $\bar{z}^* \in U$.

Note that the identity (12) is not true, in general, for any nonlinear operator P_h that is twice Fréchet-differentiable on X. However, it is certainly true for P_h being a polynomial operator [4], [7].

From now on we restrict ourselves only to that subclass of twice Fréchet-differentiable nonlinear operators on X that can satisfy the identity (12).

Our proofs will indicate that the same results can be obtained for any nonlinear sufficiently Fréchet-differentiable operator P_h, under similar assumptions. However, due to clarity and to the fact that the main results obtained here can be achieved using the above mentioned subclass of nonlinear operators P_h, we do not pursue this goal here.

We will use the following well known estimate [7]:

$$\|P_h(u) - P_h(z^*) - P_h'(z^*)(u - z^*)\| \leq q_2\|u - z^*\|^2 . \qquad (13)$$

Let us assume from now on that: The operators P_h are continuously Fréchet differentiable at a closed ball U^* such that

$$0 \in U^* \subset U \qquad (14)$$

and

$$P_h(F(z^*)) = P_h(0) = 0, \; h > 0 . \qquad (15)$$

The family (8) is called <u>stable</u> if there is $\sigma > 0$ such that

$$\|T_h'(P_h(u))^{-1}\| \leq \sigma, \; u \in U, \; h > 0 \qquad (16)$$

and <u>consistent of order p</u> if there are two constants $c_0, c_1 > 0$ such that

$$\|P_h(F(z)) - T_h(P_h(z))\| \leq c_0 h^p, \; z \in U, \; h > 0 , \qquad (17)$$

and

$$\|P_h(F'(u))(v) - T_h'(P_h(u))P_h(v)\| \leq c_1 h^p, \ u, v \in U, \ h > 0 \ . \quad (18)$$

II. Main results

We will need the following definition.

Definition. Let us define the numbers r, c and the real functions f_1 and f_2 by:

$$r = \|z_0 - z^*\| \quad (19)$$

$$c = \max(c_0, c_1) \quad (20)$$

$$f_1(r) = 3Lq_2\sigma r^2 + 3Lq_1\sigma r + 16L\sigma^2 c \cdot h^p - 2 \quad (21)$$

$$f_2(r) = q_2 r^2 + q_1 r + 4\sigma c h^p - \rho \ . \quad (22)$$

Let h_1, r_1 be such that:

$$h_1 = \min\left(\frac{\rho}{4\sigma c} , \frac{1}{8L\sigma^2 c} \right)^{\frac{1}{p}} , \quad (23)$$

$$r_1 = \min(r_1^+, r_2^+, r^*, \rho) \quad (24)$$

where, r_1^+, r_2^+ are the positive solutions of (20) and (21) respectively, provided that:

$$h \in [0, h_1) \ . \quad (25)$$

It is easy to check that if

$$r \in [0, r_1) \ , \quad (26)$$

then

$$f_1(r) \leq 0 \ , \quad (27)$$

and

$$f_2(r) \leq 0 \ . \quad (28)$$

Define the real function f_3 and the number b by

$$f_3(s) = As^2 + Bs + C \,, \tag{29}$$

$$b = \|z^*\| \,, \tag{30}$$

where,

$$A = \frac{3}{2}L\sigma \,,$$

$$B = B(r) = B_3 r^3 + B_2 r^2 + B_1 r + B_0 \,,$$

$$C = C(r) = C_2 r^2 + C_1 r + C_0 \,,$$

$$B_3 = 16L\sigma q_2 \,,$$

$$B_2 = L\sigma(4q_1 + 8bq_2 - 6q_2) \,,$$

$$B_1 = 2L\sigma q_1 b + 2L\sigma q_2 b^2 - 4q_1 - 4bq_2 \,,$$

$$B_0 = -(L\sigma q_1 b + Lq_2\sigma + 1) \,,$$

and

$$C_2 = 6q_2 \,,$$

$$C_1 = 4(q_1 + bq_2) \,,$$

$$C_0 = 2ch^p\sigma + q_1 b + q_2 b^2 \,.$$

We now observe the following:

(a) $B_0 < 0, \; B_3 > 0$. (31)

(b) The function $B^2 - 4AC$ is a sixth degree polynomial with the coefficient of the highest power being positive and the constant given by

$$B_0^2 - 4AC_0 = E_1 q_1^2 + E_2 q_2^2 + E_3 q_1 q_2 + E_4 q_1 + E_5 q_2 + E_6 \tag{32}$$

where,

$$E_1 = (L\sigma b)^2$$

$$E_2 = (L\sigma)^2$$

$$E_3 = 2(L\sigma)^2 b$$

$$E_4 = -4L\sigma b$$

$$E_5 = 2L\sigma(1 - 4b^2)$$

$$E_6 = 1 - 12L\sigma^2 h^p \,.$$

(c) Define the function g by

$$g(k) = A(4AB^2 - C')k^2 + 4AB\sigma ck + (\sigma c)^2, \; C' = C - 2\sigma ch^p \,. \tag{33}$$

The function $4AB^2 - C'$ is a sixth degree polynomial with the coefficient of the highest power being positive and the constant given by

$$4AB_0^2 - q_1 b - q_2 b^2 = = E_1^1 q_1^2 + E_2^1 q_2^2 + E_3^1 q_1 q_2 + E_4^1 q_1 + E_5 q_2^1 + E_6^1 \qquad (34)$$

where,

$$E_1^1 = 6(L\sigma)^3 b^2$$

$$E_2^1 = 6(L\sigma)^3$$

$$E_3^1 = 12(L\sigma)^3 b$$

$$E_q^1 = [12(L\sigma)^2 - 1]b$$

$$E_5 = 12(L\sigma)^2 - b^2$$

$$E_6 = 6L\sigma \ .$$

$$\text{Let } q_1^* = \max(-\frac{E_4}{E_1}, -\frac{E_4^1}{E_1^1}), \ q_2^* = \max(-\frac{E_5}{E_2}, -\frac{E_5^1}{E_2^1}) \qquad (35)$$

and

$$k^* = \max(-\frac{\sigma c}{2AB}, k_\ell) \qquad (36)$$

where k_ℓ is the large solution of the equation

$$g(k) = 0 \ .$$

Choose q_1, q_2, k, h and h_2 such that:

$$q_1 \geq q_1^* \ , \qquad (37)$$

$$q_2 \geq q_2^* \ , \qquad (38)$$

$$k \geq k^* \ , \qquad (39)$$

and

$$h < h_2 = \min\left((\frac{1}{12L\sigma^2})^{\frac{1}{p}}, (\frac{1}{Lk\sigma})^{\frac{1}{p}}, h_1\right) \ . \qquad (40)$$

It is easy to check that we can find $r_2 > 0$ such that if $r \in [0, r_3)$ where

$$r_3 = \min(r_1, r_2) \tag{41}$$

then the following are true:

$$B < 0 ,$$

$$B^2 - 4AC \geq 0 ,$$

$$\sigma c + 2AkB \leq 0 ,$$

$$g(k) \geq 0 ,$$

and

$$0 < s \leq kh^p < \frac{1}{L\sigma} , \tag{42}$$

where s is the small solution of the equation

$$f_3(s) = 0 . \tag{43}$$

Let $\epsilon > 0$, since $z_n \to z^*$ as $n \to \infty$ there exists $N = N(\epsilon)$ such that

$$\|z_n - z^*\| \leq [\frac{1}{q_2} k_1 h^p]^{\frac{1}{2}} \quad \text{for some} \quad k_1 > 0 . \tag{44}$$

Finally, define the numbers k_2, β, r_4, h_4 and the function f_4 by

$$k_2 = k + 2\sigma c + k_1 \tag{45}$$

$\beta = \min(2[\bar{q}_1 - (q_1 + q_2)], \frac{1}{q_1+q_2})$ for some fixed number $\bar{q}_1$ with

$\bar{q}_1 > q_1 + q_2 .$ (46)

$$h_3 = \min\{h_2, (\frac{\beta}{(1+2\beta)L\sigma^2 c})^{\frac{1}{p}}, (\frac{(q_1+q_2)\epsilon}{k_2})^{\frac{1}{p}}\} \tag{47}$$

$$f_4(r) = L\sigma q_2(1 + \beta)r^2 + L\sigma q_1(1 + \beta)r + (1 + 2\beta)h^p - \beta \tag{48}$$

and r_4 to be the positive solution of the equation

$$f_4(r) = 0 \tag{49}$$

guaranteed to exist by the choice of h_3 and β.

Let h, r be such that:

$$h \in (0, h_3) \tag{50}$$

and

$$r \in [0, r_4) \tag{51}$$

then it can easily be checked that:

$$f_4(r) \leq 0 \ . \tag{51}$$

We can now prove the first result.

Theorem 2. Let $F : DCX \to X$ be a nonlinear operator satisfying the hypotheses of theorem 1 and consider a Lipschitz uniform discretization (8) which is "bounded," stable and consistent of order p.

Assume:

$$q_1 \geq q_1^* \ ,$$
$$q_2 \geq q_2^* \ ,$$
$$k \geq k^* \ ,$$

and that (14) and (15) are satisfied.

Then equation (3) has a locally unique solution x_h^* such that:

$$x_h^* = P_h(z^*) + O(h^p) \tag{53}$$

for all $h \in (0,h_1]$.

There exist constants $h \in (0,h_2]$ and $r \in (0,r_3)$ such that the iteration (4) converges to x_h^*.

Moreover, the following are true:

$$x_n^h = P_h(z_n) + O(h^p) \ , \tag{54}$$

$$T_h(x_n^h) = P_h(F(z_n)) + O(h^p) \ , \tag{55}$$

and

$$x_n^h - x_h^* = P_h(z_n - z^*) + O(h^p) \tag{56}$$

for $n = 0, 1, 2 \cdots$ and all $z_0 \in U(z^*, r_3)$.

Proof. Using the Newton-Kantorovich theorem [7], (15) we show exactly as in [3] that (3) has a solution $x_h^* \in \bar{U}(P_h(z^*), r(h))$ which is unique in $U(P_h(z^*), \bar{r}(h))$ where

$$\alpha(h) = 2\sigma L \|T_h'(P_h(z^*))^{-1} T_h(P_h(z^*))\| \tag{57}$$

$$r(h) = \frac{1}{\sigma L}\left(1 - \sqrt{1-\alpha(h)}\right) \leq \rho \tag{58}$$

$$\bar{r}(h) = \frac{1}{\sigma L}\left(1 + \sqrt{1-\alpha(h)}\right) . \tag{59}$$

Moreover,

$$\|x_h^* - P_h(z^*)\| \leq r(h) \leq 2\sigma c_0 h^p \tag{60}$$

which proves (53).

As in [3] by applying theorem 1 to (3) we see that (4) converges to x_h^* if

$$\|P_h(z_0) - x_h^*\| < \frac{2}{3\|T_h'(x_h^*)^{-1}\|} \tag{61}$$

and

$$\bar{U}(x_h^*, \|P_h(z_0) - x_h^*\|) \subset \bar{U}(P_h(z^*), \rho) \tag{62}$$

that is if

$$q_2\|z_0 - z^*\|^2 + q_1\|z_0 - z^*\| + 4\sigma c_0 h^p < \frac{2-4\sigma^2 L c_0 h^p}{3L\sigma} \tag{63}$$

and

$$q_2\|z_0 - z^*\|^2 + q_1\|z_0 - z^*\| + 4\sigma c_0 h^p \leq \rho \tag{64}$$

hold respectively.

But (63) and (64) can now be written as

$$f_1(r) \leq 0$$

and

$$f_2(r) \leq 0 ,$$

which are true by the choice of h, r, h_1 and r_1.

We now show, using induction as in [3] that for $h \in (0, h_3)$, $z_0 \in U(z^*, r_3)$

$$\|x_n^h - P_h(z_n)\| \leq s \ , \quad n = 0, 1, 2 \cdots . \tag{65}$$

By rearranging the sequences in (65) we can start the induction from $n = 0$ if necessary. The above inequality is true for $n = 0$. Let us assume that it is true for $n = 0, 1, 2, \cdots, i$.

Then,

$$\begin{aligned} x_{i+1}^h - P_h(z_{i+1}) &= T_h'(x_i^h)^{-1}[[T_h'(x_i^h)(x_i^h - P_h(z_i)) - T_h(x_i^h) + T_h(P_h(z_i))] \\ &+ [(T_h'(x_i^h) - T_h'(P_h(z_i)))P_h(F'(z_i)^{-1}F(z_i))] + [T_h'(P_h(z_i)P_h(F'(z_i)^{-1}F(z_i) - P_h(F(z_i))] \\ &+ [P_h(F(z_i)) - T_h(P_h(z_i))]] + [P_h(z^*) - P_h(z_i) + P_h'(z^*)(z_i - z^*)] \\ &- P_h'(z^*)(F'(z_i)^{-1}F(z_i)) + P_h''(\bar{z}^*)(z_{i+1} - z^*)^2 + P_h(F'(z_i)^{-1}F(z_i)) \end{aligned} \tag{66}$$

We now have, using (10), (13).

$$\|P_h(z^*) - P_h(z_i) + P_h'(z^*)(z_i - z^*)\| \leq q_2\|z_i - z^*\|^2 \ ,$$

$$\|P_h'(z^*)(z_i - z_{i+1})\| \leq q_1\|z_i - z_{i+1}\| \ ,$$

$$\|P_h''(\bar{z}^*)(z_{i+1} - z^*)^2\| \leq q_2\|z_{i+1} - z^*\|^2 \ ,$$

$$\|P_h(z_i - z_{i+1})\| \leq \|P_h(z^*)\| + q_1\|z_i - z_{i+1} - z^*\| + q_2\|z_i - z_{i+1} - z^*\|^2 \ .$$

By theorem 1, $\|z_{i+1} - z^*\| \leq z_i - z^*\| \leq \cdots \leq \|z_0 - z^*\|$ and the result on the norm of the braces above in [3] we have that the right hand side of (66) is bounded by

$$\frac{\sigma}{1-L\sigma s}\left[\tfrac{1}{2}Ls^2 + 2L(q_1(2r + b) + q_2(2r + b)^2)rs + 2ch^p\right]$$

$$+ q_2r^2 + 2q_1r + q_2r^2 + 2q_1r + q_1b + 4q_2r^2 + q_2b^2 + 4bq_2r = s \ .$$

That is,

$$\|x_h^h - P(z_h)\| \leq s \leq kh^p \tag{67}$$

which proves (54).

Equation (55) can be proved exactly as in [3].

Finally, for some $\bar{z} \in U^*$

$$\|(x_n^h - x_h^*) - P_h(z_n - z^*)\| \le \|x_n^h - P_h(z_n)\| + \|x_h^* - P_h(z^*)\|$$

$$+ \|[P_h(0) + (P_h(z^*) - P_h(z_n) + P_h'(\bar{z})(z_n - z^*))]\|$$

$$\le kh^p + 2\sigma ch^p + q_2\|z_n - z^*\|^2 \le k_2 h^p ,$$

by (13), (14), (15), (44), (45), (66), (67) and (68) which proves (56) and completes the proof of the theorem.

We now prove the last result.

Theorem 3. Assume:

(1) The hypotheses of theorem 2 hold;

(2) there exist constants $\bar{q}_1$, $\bar{q}_2$ such that

$$\liminf_{n>0} \|P_h(z)\| \ge \bar{q}_1\|z\| + \bar{q}_2\|z\|^2 \quad \text{for} \quad z \in \bar{U} \quad \text{and} \quad \bar{q}_1 > q_1 + q_2 . \tag{69}$$

Then for $r \in (0,r_4)$, and for any fixed $0 < \epsilon \le 1$ and $z_0 \in \bar{U}(z^*,r)$ there exists $h_4 = h_4(\epsilon,z_0) \in (0,h_6)$ where,

$$h_6 = \min(h_3,h_5)$$

and $h_5 = h(z_0)$ is such that

$$\|P_h(z_i - z^*)\| \ge \bar{q}_1\|z_i - z^*\| + \bar{q}_2\|z_i - z^*\|^2 , \quad h \in (0,h_5) \tag{70}$$

and i is the unique integer defined by

$$\|z_{i+1} - z^*\| < \epsilon \le \|z_i - z^*\| \tag{71}$$

such that

$$|\min\{n \ge 0, \|z_n - z^*) < \epsilon\} - \min\{n \ge 0, \|x_n^h - x_h^*\| < \epsilon\} \le 1 \tag{72}$$

for all $h \in (0,h_4]$.

Proof. By (68) and the choice of h, β we have

$$\|x_{i+1}^h - x_h^*\| \le \|P_h(z_{i+1} - z^*)\| + k_2 h^p \le q_1\|z_{i+1} - z^*\| + q_2\|z_{i+1} - z^*\|^2 + k_2 h^p$$

$$\le (q_1 + q_2)\epsilon + k_2 h^p < 2(q_1 + q_2)\epsilon \ . \tag{73}$$

Using theorem 1, and the choice of h, β, r we get

$$\|x_{i+2}^h - x_h^*\| \le \frac{L\sigma\|x_{i+1}^h - x_h^*\|^2}{2(1-L\sigma\|x_{i+1}^h - x_h^*\|)}$$

$$\le \frac{L\sigma\|x_0^h - P_h(z^*) + P_h(z^*) - x_h^*\|}{2(1-L\sigma\|x_0^h - P_h(z^*) + P_h(z^*) - x_h^*\|)} \|x_{i+1} - x_h^*\|$$

$$< \frac{L\sigma[q_1\|z_0 - z^*\| + q_2\|z_0 - z^*\|^2 + 2\sigma c h^p]}{[1 - L\sigma(q_1\|z_0 - z^*\| + q_2\|z_0 - z^*\|^2 + 2\sigma c h^p)]} (q_1 + q_2)\epsilon$$

$$\le \beta(q_1 + q_2)\epsilon < \epsilon \tag{74}$$

Moreover, by (68), (69) and (70),

$$\bar{q}_1\epsilon + \bar{q}_2\epsilon^2 \le q_1\|z_i - z^*\| + q_2\|z_i - z^*\|^2$$

$$\le \|P_h(z_i - z^*)\| \le (\|x_i^h - x_h^*\| + k_2 h^p)$$

or

$$\|x_i^h - x_h^*\| \ge \bar{q}_1\epsilon + \bar{q}_2\epsilon^2 - c_2 h^p \ . \tag{75}$$

If $\|x_{i-1}^h - x_h^*\| < \epsilon$ then as in (74) we get

$$\|x_i^h - x_h^*\| < \tfrac{1}{2}\beta\epsilon \ . \tag{76}$$

By, (75) and (76) we must also have

$$\bar{q}_1\epsilon + \bar{q}_2\epsilon^2 - k_2 h^p < \tfrac{1}{2}\beta\epsilon$$

or

$$\bar{q}_1\epsilon + \bar{q}_2\epsilon^2 - (q_1 + q_2)\epsilon < \tfrac{1}{2}\beta\epsilon$$

or

$$\bar{q}_2 < \frac{1}{\epsilon}(\tfrac{1}{2}\beta + q_1 + q_2 - \bar{q}_1) < 0 \ ,$$

that is

$$\bar{q}_2 < 0 \ ,$$

which is a contradiction. Therefore, we have

$$\|x_{i-1}^h - x_h^*\| \geq \epsilon \ . \tag{77}$$

The result is now obtained by (71), (74) and (77).

We can do better sometimes. Let us assume that:

$$0 < q_1 < 1 \tag{78}$$

and

$$\bar{q}_1 + \bar{q}_2 > \max(\frac{1}{2}, \frac{q_1+2q_2}{4}, \frac{q_2}{2(1-q_1)}) \ . \tag{79}$$

Define $\beta_1 > 0$ by

$$\beta_1 = \min(\frac{1}{R}, \beta, 2\bar{q}_2, 2(\bar{q}_1 + \bar{q}_2) - 1, \frac{4(\bar{q}_1+\bar{q}_2)-(q_1+2q_2)}{2}, \frac{2(1-q_1)(\bar{q}_1+\bar{q}_2)-q_2}{1-q_1}) \ .$$

Choosing R, $h_7 > 0$ such that:

$$\max(\frac{1}{2}, \frac{q_1+2q_2}{4}, \frac{q_2}{2(1-q_1)}) < R \leq \bar{q}_1 + \bar{q}_2 - \frac{1}{2}\beta^*, \text{ for some } \beta^* \in (0,\beta_1) \ ,$$

$$h_7 = \min(h_6, (\frac{R\epsilon}{k_2}^2)^{\frac{1}{p}}, (\frac{8R-rq_2-q_1}{k_2})^{\frac{1}{p}}, (\frac{(2R-q_2)(2R-q_2-2Rq_1)}{4R^2(2R-q_2)k_2})^{\frac{1}{p}}).$$

Here, $\epsilon \in [\epsilon^+, \frac{1}{2R}]$ and ϵ^+ is the positive solution of the equation

$$(2R - q_2)\epsilon^2 - q_1\epsilon - k_2h^p = 0 \ ,$$

guaranteed to exist by the choice of h and R.

Define r_5 to be the positive solution of the equation

$$f_5(r) = L\sigma q_2(1 + \beta^*)r^2 + L\sigma q_1(1 + \beta^*)r + (1 + 2\beta^*)h^p - \beta^* = 0 \ ,$$

guaranteed to exist by the choice of h, R and b^*.

Set

$$r_6 = \min(r_4, r_5) \ .$$

We can now easily check that by (78), (79) and the choice of h, R, β^*

$$0 < \epsilon^+ < \frac{1}{2R} < 1 \ .$$

We can show exactly as in theorem 3 that:

$$\|x_{i+1}^h - x_h^*\| \leq q_2\epsilon^2 + q_1\epsilon + k_2h^p < 2R\epsilon^2$$

$$\|x_{i+2}^h - x_h^*\| \leq \beta^* R\epsilon^2 < \epsilon^2$$

$$\|x_{i-1}^h - x_h^*\| \geq \epsilon^2 .$$

Then by theorem 3

$$|\min\{n \geq 0, \|z_n - z^*\| < \epsilon\} - \min\{n \geq 0, \|x_n^h - x_h^*\| < \epsilon^2 < \epsilon\}| \leq 1 . \quad (80)$$

Moreover,

$$\|x_{i+1}^h - x_h^*\| \leq 2R\epsilon^2 < \epsilon . \quad (81)$$

Therefore (80), because of (81), becomes

$$I = \min\{n \geq 0, \|x_n^h - x_h^*\| < \epsilon\} - \min\{n \geq 0, \|z_n - z^*\| < \epsilon\}, \; I = -1, 0.$$

We have now proved the following:

Corollary. Assume that the hypotheses of theorem 3, (78) and (79) hold.

Then for $r \in (0,r_6)$, and for any $\epsilon \in [\epsilon^+, \frac{1}{2R}]$ and $z_0 \in \bar{U}(z^*,r)$ there exists $h_8 = h_8(\epsilon,z_0) \in (0,h_7)$ such that:

$$I = \min\{n \geq 0, \|x_n^h - x_h^*\| < \epsilon\} - \min\{n \geq 0, \|z_n - z^*\| < \epsilon\}, \text{ with } I = -1, 0 \quad (82)$$

for all $h \in (0,h_8]$.

The above result improves and extends the corresponding one in [3].

Note that (82) implies that:

$$\min\{n \geq 0, \|x_n^h - x_h^*\| < \epsilon\} \leq \min\{n \geq 0, \|z_n - z^*\| < \epsilon\} . \quad (83)$$

Inequality (83) arises the question of the actual equality being achieved in (83). The proof of (83) was based primarily on (7), (56) and (69).

We note:

$$x_n^h - x_h^* = P_h(z_n - z^*) + 0(h^p)$$

$$= P_h(0) + P_h'(\bar{z})(z_n - z^*) + 0(h^p), \text{ for some } \bar{z} \in U^*$$

$$= P_h'(\bar{z})(z_n - z^*) + 0(h^p) . \qquad (84)$$

Assume that the inverse of $P_h'(\bar{z})$ exists and is bounded for all sufficiently small h and $\bar{z} \in U^*$. Set $L_h = (P_h'(\bar{z}))^{-1}$ and $\bar{q}_1^* \geq \|L_h\|$.

Then the equation (84) can be written as

$$z_n - z^* = L_h(x_n^h - x_h^*) + 0(h^p) \text{ , } n = 0, 1, 2, \cdots, h \text{ sufficiently small. } (85)$$

By theorem 1 we can have,

$$\|x_{n+1}^h - x_h^*\| \leq \frac{L\sigma\|x_n^h - x_h^*\|^2}{2(1-L\sigma\|x_n^h - x_h^*\|)} \text{ , } n = 0, 1, 2, \cdots . \qquad (86)$$

Moreover, note that the assumption on P_h given by (11), (13), (14), and (15) hold for the linear operator L_h with $q_1 = \bar{q}_1^*$, $q_2 = 0$.

If the rest of the assumptions made for P_h hold for L_h then the proofs of theorem 2 will go through for L_h.

Finally, assume that:

$$\liminf_{h>0} \|L_h(x)\| \geq \bar{q}_1^*\|x - x_h^*\| + \bar{q}_2^*\|x - x_h^*\|^2, \; x \in \bar{U}(P_h(z^*),\rho) \qquad (87)$$

$(\bar{q}_2^* = 0)$.

If we now interchange the role of the z's with the x's in theorems 1, 2, 3 and the Corollary, under similar assumptions on h, $\rho_h^* = |x_0^h - x_h^*\|$, $\bar{q}_1^*$, $\bar{q}_2^*$, q_1^* the proofs of both theorem 3 and the Corollary will go through. Therefore, we can show that:

$$\min\{n \geq 0, \|x_n^h - x_h^*\| < \epsilon\} \geq \min\{n \geq 0, \|z_n - z^*\| < \epsilon\} , \qquad (88)$$

for sufficiently small h and ρ_h^*.

By (83) and (88) we finally obtain

$$\min\{n \geq 0, \|x_n^h - x_h^*\| < \epsilon\} = \min\{n \geq 0, \|z_n - z^*\| < \epsilon\} , \qquad (89)$$

where $h \in (0,h_9]$, $z_0 \in \bar{U}(z^*,r)$, $x_0^h \in \bar{U}(P_h(z^*), \rho)$ and any $\epsilon' \in [\epsilon_1^+, \frac{1}{2R_1}] \cap [\epsilon_1, \frac{1}{2R}]$ provided that the intersection is nonempty.

Here, $h_9 = \min(h_8, h_8^*)$, $r \in (0,r_6)$, $\rho \in (0,\rho_6)$. Finally, the h_8^*, ρ_6, ρ_1^+ and R_1 were found as the h_8, r_6, ϵ^+ and R respectively.

The result given by equation (89) improves and extends the corresponding result in [3].

III. Remarks - Examples.

The condition (15) does not hold in general for any nonlinear projections P_h, $h > 0$. However, it holds for linear projections [3] and in fact for multilinear projections [4].

In particular, it has been shown in [4] that when F is a quadratic operator and the P_h are quadratic projections on $\mathbb{R}^n$(or $\mathbb{C}^n$) then (3) can be replaced with a quadratic system on $\mathbb{R}^n$ (or $\mathbb{C}^n$).

REFERENCES

[1] Allgower, E.L. and McCormick, S.F. Newton's method with mesh refinements for numerical solution of nonlinear two-point boundary value problems. Numer. Math. 29 (1978), 237-260.

[2] Allgower, E.L. McCormick, S.F. and Pryor, D.V. A general mesh independence principle for Newton's method applied to second order boundary value problems. Computing 23 (1979), 233-246.

[3] Allgower, E.L., Böhmer, K. Potra, F.A. and Rheinboldt, W.C. A mesh-independence principle for operator equations and their discretizations. SIAM J. Num. Anal. Vol. 23, 1, (1986), 160-169.

[4] Argyros, I.K. Quadratic equations and applications to Chandrasekhar's and related equations. Bull. Austral. Math. Soc. Vol. 32 (1985), 275-292.

[5] McCormick, S.F. A revised mesh reforcement strategy for Newton's method applied to nonlinear two-point boundary value problems. Lecture Notes in Mathematics 679, Springer-Verlag Publ. 1978, 15-23.

[6] Ortega, J.M. and Rheinboldt, W.C. On discretization and differentiation of operators with application to Newton's method. SIAM J. Num. Anal. Vol. 3, (1966), 143-156.

[7] _____. Iterative solutions of nonlinear equations in several variables. Academic Press, New York, 1970.

[8] Rheinboldt, W.C. An adaptive continuation process for solving systems of nonlinear equations. Polish Academy of Science. Banach Ctr. Publ. 3 (1977), 129-142.

Rev. Acad. Ciencias. Zaragoza 45, (1990)

A NOTE ON NEWTON'S METHOD

by

IOANNIS K. ARGYROS
Department of Mathematics
New Mexico State University
Las Cruces, NM 88003

Abstract. We use a Newton-like iteration to solve the nonlinear operator equation in a Banach space. The basic assumption is that the Fréchet-derivative of the nonlinear operator is Hölder continuous on some open ball centered at the initial guess. Under natural assumptions, we prove linear convergence of the iteration to a locally unique solution of the nonlinear equation.

Key words and phrases: Newton-like iteration, Banach space, Fréchet-derivative.

(1980) A.M.S. classification codes: 65J15, 65H10.

Introduction. We introduce the Newton-like iteration

$$x_{n+1} = x_n - F'(y_n)^{-1}F(x_n),\ n = 0, 1, 2, \cdots \qquad (1)$$

to solve the equation

$$F(x) = 0, \qquad (2)$$

where F is a nonlinear operator on a Banach space X. Here we assume that the Fréchet derivative $F'(x)$ of F is Hölder continuous on some ball

$$U(x_0, r_0) = \{x \in X \mid \|x - x_0\| < r_0\},\ r_0 > 0\ .$$

The point $x_0 \in X$ and the arbitrary points y_n, $n = 0, 1, 2, \cdots$ are chosen sufficiently close to the desired solution. Then under natural assumptions we show that (1) converges linearly to a unique solution x^* of (2) in $U(x_0, r)$ for some $r > 0$.

Note that for $y_n = x_n$ we obtain Newton's iteration, whereas for $y_n = x_0$ we obtain the modified one. The computer will determine the y_n's as to minimize the effort each time.

Here is an incomplete list of the usual assumptions for the convergence of (1) to a solution of (2):

(a) the existence and boundedness of the second Fréchet derivative of F (see, ex. [2], [3] and [4]).

(b) The assumption of analyticity which eliminates explicit mention of the second derivative [7].

(c) The case when the first Fréchet derivative of F satisfies a Lipschitz condition. (See, ex. [5], [6] and the references there.)

Finally,

(d) various other assumptions mainly based on the possibility of replacing $F'(x_n)^{-1}$ with a sequence of linear operators which are "close" in some sense to $F'(x_n)^{-1}$, $n = 0, 1, 2, \cdots$. (See, ex. [5], [6] and the references there.)

We will need the following definition.

Definition. We assume that F is once Fréchet-differentiable [2] and $F'(x)$ is the first Fréchet-derivative at a point $x \in X$. It is well known that $F'(x) \in L(X)$, the space of bounded linear operators from X to X. We say that the Fréchet-derivative $F'(x)$ is Hölder continuous over a domain $D \subset X$ if for some $c > 0$, $p \in [0, 1]$, and all $x, y \in D$,

$$\|F'(x) - F'(y)\| \leq c\|x - y\|^p . \qquad (3)$$

From now on we will find it more convenient to assume that $D = U(x_0, r_0)$ for some fixed $x_0 \in X$ and $r_0 > 0$.

Lemma. Let $F'(x)$ be Hölder continuous on $U(x_0, r_0)$ for some x_0, r_0, p such that $x_0 \in X$, $r_0 > 0$ and $p \in [0, 1]$. Suppose that $F'(x_0)$ has a bounded inverse. For any $b_0 > \|F'(x_0)^{-1}\|$, there is a number $r_3 \leq \min\{1, r_0\}$ such that:

(a) If $x \in U(x_0, r_3)$, then the linear operator $F'(x)$ has bounded inverse and

$$\|F'(x)^{-1}\| < b_0 ; \tag{4}$$

(b) if $x_i \in U(x_0, r_3)$, $i = 1, 2, 3$ then

$$\|F(x_1) - F(x_2) - F'(x_3)(x_1 - x_2)\| \leq \frac{1}{2b_0} \|x_1 - x_2\| . \tag{5}$$

Proof. (a) If $x \in U(x_0, r_0)$, then

$$\|F'(x) - F'(x_0)\| \leq c\|x - x_0\|^p .$$

Choose $r_1 > 0$ such that

$$0 < r_2 \leq \min(r_0, (\frac{1}{4b_0c})^{1/p}) ,$$

then if $x \in U(x_0, r_1)$

$$\|F'(x) - F'(x_0)\| \leq c\|x - x_0\|^p \leq cr_2^p \leq \frac{1}{4b_0} .$$

Since,

$$\|F'(x_0)^{-1}\| \cdot \|F'(x) - F'(x_0)\| < b_0 \cdot \frac{1}{4b_0} < 1 ,$$

by the Banach lemma $F'(x)$ has a bounded inverse for $x \in U(x_0, r_1)$. Therefore, there exists an $r_2 > 0$, with

$$0 < r_2 \leq r_1$$

such that if $x \in U(x_0, r_2)$, then $\|F'(x)^{-1}\| < b_0$.

That proves (4).

(b) If $r_3 = \min\{1, r_2\}$ and $x_i \in U(x_0, r_3)$, $i = 1, 2, 3$, we first have

$$F(x_1)-F(x_2)-F'(x_3)(x_1-x_2) = (F(x_1)-F(x_2)-F'(x_0)(x_1-x_2)) + (F'(x_0)-F'(x_3))(x_1-x_2)$$

$$= \int_0^1 (F'[tx_1+(1-t)x_2]-F'(x_0))(x_1-x_2)dt+(F'(x_0)-F'(x_3))(x_1-x_2) .$$

But,

$$\|\int_0^1 (F'[tx_1 + (1 - t)x_2] - F'(x_0))(x_1 - x_2)dt\| \leq cr_3^p\|x_1 - x_2\|$$

$$\leq \frac{1}{4b_0}\|x_1 - x_2\| \qquad (6)$$

and

$$\|(F'(x_0) - F'(x_3))(x_1 - x_2)\| \leq cr_3^p\|x_1 - x_2\|$$

$$\leq \frac{1}{4b_0}\|x_1 - x_2\| .$$

Therefore,

$$\|F(x_1) - F(x_2) - F'(x_3)(x_1 - x_2)\| \leq \frac{1}{2b_0}\|x_1 - x_2\| .$$

That proves (5) and completes the proof of the lemma.

We now state and prove the main result.

Theorem. Let $F'(x)$ be Hölder continuous on $U(x_0, r_3)$, where r_3 is defined in the lemma. Suppose that $x_0 \in X$ is such that $F'(x_0)$ has an inverse satisfying $\|F'(x_0)^{-1}\| < b_0 < \infty$ and

$$\|F(x_0)\| < \frac{r_3}{2b_0} .$$

Let y_n be arbitrary points such that $y_n \in U(x_0, r_3)$, $n = 0, 1, 2, \cdots$.

Then the iteration $\{x_n\}$, given by

$$x_{n+1} = x_n - (F'(y_n))^{-1}F(x_n), \quad n = 0, 1, 2, \cdots$$

converges to a unique solution x^* of (2) in $U(x_0, r_3)$.

Moreover, the following estimate holds

$$\|x_n - x^*\| < 2^{-n}r_3, \quad n = 0, 1, 2, \cdots . \qquad (7)$$

Proof. Using (1) for $y_0 = x_0$ we obtain

$$\|x_1 - x_0\| = \|F'(x_0)^{-1}F(x_0)\| \le \|F'(x_0)^{-1}\| \cdot \|F(x_0)\| \le b_0\|F(x_0)\| < \frac{r_3}{2} .$$

By (5) and the identity

$$F(x_1) = F(x_1) - F(x_0) - F'(y_0)(x_1 - x_0)$$

we get

$$\|F(x_1)\| \le \frac{1}{2b_0} \|x_1 - x_0\| .$$

Claim. Suppose that x_i, $i = 1, 2, \cdots, n$ have been chosen such that

$$\|x_i - x_0\| < r_3 , \tag{8}$$

$$\|x_i - x_{i-1}\| \le b_0\|F(x_{i-1})\| , \tag{9}$$

and

$$\|F(x_i)\| \le \frac{1}{2b_0} \|x_i - x_{i-1}\| . \tag{10}$$

Then, (8), (9) and (10) hold for $i = 1, 2, 3, \cdots, n, \cdots$.

By (4), we have

$$\|x_{n+1} - x_n\| \le b_0\|F(x_n)\| < \frac{1}{2} \|x_n - x_{n-1}\| \tag{11}$$

that proves (9). Also, by (11)

$$\|x_{n+1} - x_0\| \le \{ \sum_{i=0}^{n} 2^{-i} \} \|x_1 - x_0\|$$

$$< \{1 - 2^{-(n+1)}\} r_3 < r_3$$

which proves (8).

Moreover, using (5) and

$$F(x_{n+1}) = F(x_{n+1}) - F(x_n) - F'(y_n)(x_{n+1} - x_n) ,$$

we obtain

$$\|F(x_{n+1})\| \leq \frac{1}{2b_0} \|x_{n+1} - x_n\| ,$$

which proves (10). The claim is now proved.

Let n, q be two integers, then

$$\|x_{n+p} - x_n\| \leq \sum_{j=1}^{q} \|x_{n+j} - x_{n+j-1}\|$$

$$\leq b_0\|F(x_0)\| \cdot 2^{-n}\{\sum_{j=0}^{q-1} 2^{-j}\} < 2^{-n} r_3 .$$

Therefore, $\{x_n\}$, $n = 0, 1, 2, \cdots$ constitutes a Cauchy sequence in a Banach space and as such it converges to some $x^* \in X$. By (8) and (10) respectively, we get

$$\|x^* - x_0\| \leq r_3$$

and

$$F(x^*) = 0 .$$

Finally, to show that x^* is the unique solution of (2) in $U(x_0, r_3)$ let us assume that x_1^* is another solution of (2) in $U(x_0, r_3)$, with $x^* \neq x_1^*$. Then

$$\|x^* - x_1^*\| = \|F'(x_0)^{-1}F'(x_0)(x^* - x_1^*)\| \leq b_0\|F'(x_0)(x^* - x_1^*)\|$$

$$\leq \frac{1}{2} \|x^* - x_1^*\| ,$$

which contradicts the assumption $x^* \neq x_1^*$.

Therefore, $x^* = x_1^*$. Letting $q \to \infty$ in (12), we obtain (7) and that completes the proof of the theorem.

If $y_n = x_0$, $n = 0, 1, 2, \cdots$, (1) reduces to the modified Newton's iteration which requires the evaluation of the same inverse $F'(x_0)^{-1}$ at each step of the iteration.

However, if $y_n \neq x_0$, (1) requires the evaluation of the inverse operators $F'(y_n)^{-1}$, $n = 0, 1, 2, \cdots$ at each step, which constitutes a difficult task in general.

A usual alternative is then to find a sequence of bounded linear operators L_n, $n = 0, 1, 2, \cdots$, such that

$$\|L_n - F'(x_0)\| < \frac{1}{4b_0}$$

and

$$\|L_n^{-1}\| < b_0 .$$

Following the proof of the above theorem, we can then easily show that the iteration

$$x_{n+1} = x_n - L_n^{-1}F(x_n), \quad n = 0, 1, 2, \cdots \tag{13}$$

converges to a solution x^* of (2).

One can refer to [5], [6] and the references there for an extensive analysis of iterations like (13).

Some of the results in [1] (especially Theorems 1 and 2) are similar to ours for $p = 1$ only. However, the results there cannot be applied here for $p \neq 1$.

The motivation for the introduction of an interation like (1) when $F'(x)$ is Hölder continuous on some open ball is due to the existence of problems like the one illustrated in the example.

Example. Consider the differential equation

$$x'' + x^{1+p} = 0, \quad p \in [0, 1] \tag{14}$$

$$x(0) = x(1) = 0 .$$

We divide the interval $[0, 1]$ into n subintervals and we set $h = \frac{1}{n}$. Let $\{v_k\}$ be the points of subdivision with

$$0 = v_0 < v_1 < \cdots < v_n = 1 .$$

A standard approximation for the second derivative is given by

$$x_i'' = \frac{x_{i-1} - 2x_i + x_{i+1}}{h^2}, \quad x_i = x(v_i), \quad i = 1, 2, \cdots, n - 1 .$$

Take $x_0 = x_n = 0$ and define the operator $F : \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ by

$$F(x) = H(x) + h^2\varphi(x) \tag{15}$$

$$H = \begin{bmatrix} 2 & -1 & & & 0 \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ 0 & & & -1 & 2 \end{bmatrix}, \qquad \varphi(x) = \begin{bmatrix} x_1^{1+p} \\ x_2^{1+p} \\ \vdots \\ x_{n-1}^{1+p} \end{bmatrix},$$

and

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \end{bmatrix}.$$

Then

$$F'(x) = H + h^2(p+1)\begin{bmatrix} x_1^p & & & 0 \\ & x_2^p & & \\ & & \ddots & \\ 0 & & & x_{n-1}^p \end{bmatrix}.$$

Newton's method cannot be applied to the equation

$$F(x) = 0 . \tag{16}$$

We may not be able to evaluate the second Fréchet-derivative since it would involve the evaluation of quantities of the form x_i^{-p} and they may not exist.

Let $x \in \mathbb{R}^{n-1}$, $H \in \mathbb{R}^{n-1} \times \mathbb{R}^{n-1}$ and define the norms of x and H by

$$\|x\| = \max_{1 \le j \le n-1} |x_j|$$

$$\|H\| = \max_{1 \le j \le n-1} \sum_{k=1}^{n-1} |h_{jk}| .$$

For all $x, z \in \mathbb{R}^{n-1}$ for which $|x_i| > 0$, $|z_i| > 0$, $i = 1, 2, \cdots, n-1$ we obtain, for $p = \frac{1}{2}$ say,

$$\|F'(x) - F'(z)\| = \|\mathrm{diag}\{(1 + \tfrac{1}{2})h^2(x_j^{1/2} - z_j^{1/2})\}\|$$

$$= \tfrac{3}{2} h^2 \max_{1 \le j \le n-1} |x_j^{1/2} - z_j^{1/2}| \le \tfrac{3}{2} h^2 [\max|x_j - z_j|]^{1/2}$$

$$= \tfrac{3}{2} h^2 \|x - z\|^{1/2} .$$

Therefore, under the assumptions of the theorem, iteration (1) will converge to the solution x^* of (16).

REFERENCES

[1] Graves, L.M. Some mapping theorems. Duke Math. J., Vol 17 (1950), pp. 111–114.

[2] Kantorovich, L.V. On Newton's method. Math. Reviews, Vol. 12 (1951), p. 419.

[3] Kantorovich, L.V. and Akilov, G.P. Functional analysis in normed spaces. Oxford Publ. Pergamon Press, 1964.

[4] Mysovkih, I.P. On the convergence of Newton's method. Math. Reviews, Vol. 12 (1951), p.419.

[5] Rall, L.B. Nonlinear functional analysis and applications. (Article by J. Dennis.) Academic Press, 1971.

[6] Rheinboldt, W.C. Numerical analysis of parametrized nonlinear equations. John Wiley, Publ. 1986.

[7] Stein, L.M. Sufficient conditions for the convergence of Newton's method in complex Banach spaces. Proc. Amer. Math. Soc. Vol. 3 (1952), pp. 858–863.

Rev. Acad. Ciencias. Zaragoza 45, (1990)

ON THE SOLUTION OF COMPACT LINEAR AND QUADRATIC OPERATOR EQUATIONS IN A HILBERT SPACE.

by

IOANNIS K. ARGYROS

Department of Mathematics

New Mexico State University

Las Cruces, NM 88003

Abstract. Some improvements by iteration for the approximate solutions of compact linear and quadratic operator equations in a separable Hilbert space are provided.

Key words and phrases: Compact operator, Hilbert space.

(1980) A.M.S. classification codes: 46B15, 45L05, 65R05.

Introduction.

Consider the problem of approximating a solution x^* of the linear equation

$$x = y + L(x) , \tag{1}$$

defined in a separable Hilbert space H with y given and $y, x \in H$. Here L is considered to be a compact linear operator in H.

There are two popular ways of approximating a solution x^* of equation (1). The first one consists of considering a suitable complete basis $\{z_i\}$, $i = 1, 2, \cdots, n$ of elements of H and choosing constants $\{c_{ij}\}$ such that the element

$$\bar{x}_n = \sum_{i=1}^{n} c_{ni} z_i \tag{2}$$

is an approximate solution of equation (1). It is well known [4] that Galerkin methods, least squares methods et al. make use of approximate solutions of the form (2).

The second way consists of considering the iterated approximation

$$\tilde{x}_n = y + \sum_{i=1}^{n} d_{ni} L(z_i) \ . \qquad (3)$$

In this paper we answer to the following question: is it true that $\bar{x}_n$ is improved by an iteration if n is chosen sufficiently large? The answer is yes for the Galerkin method.

The iteration of Galerkin's solution on the one hand is very cheap since the elements $L(z_i)$ have already been computed and on the other hand practical experience often recommends it (see ex [4]).

Finally we extend these results to include the approximate solutions of the quadratic equation

$$w = y + B(w, w) \qquad (4)$$

where B is a bilinear operator in H [2], [3], $y \in H$ is fixed and w is the unknown element in H.

II. Main results

We will need the lemmas:

Lemma 1. Let Z_n be a subspace of H spanned by a complete sequence $\{z_i\}$, $i = 1, 2, \cdots$ and P_n be the orthogonal projection onto Z_n. Then the following properties are true

$$P_n^2 = P_n \ ,$$

$$P_n = P_n^* \ ,$$

$$\|P_n\| = 1$$

and

$$P_n(x) \to x \quad \text{for all} \quad x \in H \ .$$

Lemma 2. If L is a compact linear operator in H and P_n is the orthogonal projection onto Z_n, then

$$\|L - LP_n\| \to 0 \quad \text{as} \quad n \to \infty . \tag{5}$$

Proof. Let us assume that

$$\|L - LP_n\| \not\to 0 \quad \text{as} \quad n \to \infty . \tag{6}$$

Then there exists $\epsilon > 0$ and a unit norm sequence $\{x_n\}$, $n = 0, 1, 2, \cdots$ in H such that

$$\|L^*(x_n) - P_n L^*(x_n)\| > \epsilon .$$

Since L^* is a compact operator there exists an element $v \in H$ such that

$$L^*(x_n) \to v \quad \text{as} \quad n \to \infty .$$

But

$$\|L - LP_n\| = \|L^* - P_n L^*\| = \sup_{\|x\|=1} \|L^*(x) - P_n L^*(x)\|$$

and

$$\|L^*(x_n) - P_n L^*(x_n)\| \leq \|(I - P_n)(L^*(x_n) - v)\| + \|v - P_n(v)\| \to 0$$

as $n \to \infty$, contradicting (6).

The result now follows.

We can now prove the main result:

Theorem 1. Let $\bar{x}_n^*$, $\tilde{x}_n^*$ denote the best possible approximations of the forms (2) and (3) respectively.

If

$$\tilde{x}_n = y + L(\bar{x}_n^*) , \tag{7}$$

then

$$\|x^* - \tilde{x}_n^*\| \leq \|x^* - \tilde{x}_n\| \leq e_n \|x^* - \bar{x}_n^*\| \tag{8}$$

where

$$e_n = \|L - LP_n\| \to 0 \quad \text{as} \quad n \to \infty . \tag{9}$$

Proof. The first inequality in (8) follows from the definition of $\tilde{x}_n^*$. To show the second inequality, using (1) and (7) we get

$$x^* - \tilde{x}_n^* = L(x^* - \bar{x}_n^*) . \tag{10}$$

But the element that minimizes $\|x^* - \bar{x}_n\|$ is [[1], Ch. 1]

$$\bar{x}_n^* = P_n(x^*) . \tag{11}$$

Using Lemma 1 and (11) the estimate (10) becomes

$$\begin{aligned} \|x^* - \tilde{x}_n\| &= \|(L - LP_n)(x^* - P_n(x^*))\| \\ &\leq \|L - LP_n\| \cdot \|x^* - \bar{x}_n^*\| \\ &= e_n \|x^* - \bar{x}_n^*\| . \end{aligned} \tag{12}$$

The result now follows immediately from (12) and Lemma 2.

We will now extend our results to include quadratic equations of the form (4).

We first need the definition.

Definition. An operator $B : H \times H \to H$ sending $(x, y) \in H \times H$ to $B(x, y) \in H$ is called bilinear if it is linear in each variable separately and symmetric if

$$B(x, y) = B(y, x) .$$

Let us also define the linear operator $B(x) : H \to H$ by

$$B(x)(y) = B(x, y), \text{ for fixed } x \in H \text{ and all } y \in H . \tag{13}$$

From now on we will assume that the bilinear operator B is symmetric.

As before, let us consider two ways of approximating a solution w^* of equation (4). The first one consists of considering a suitable basis $\{z_i\}$, $i = 1, 2, \cdots, n$ of elements of H and choosing constants $\{r_{ij}\}$ such that the element

$$\bar{w}_n = \sum_{i=1}^{n} r_{ni} z_i \tag{14}$$

is an approximate solution of equation (4).

The second way consists of considering the approximation

$$\tilde{w}_n = y + \sum_{i=1}^{n} q_{ni} z_i \qquad (15)$$

where the constant $\{q_{ni}\}$, $i = 1, 2, \cdots, n$ are to be found by solving a quadratic system in $\mathbb{R}^n$ (or $\mathbb{C}^n$) with the techniques developed in [2], [3].

We can now prove the following extension of the main result (thm. 1) for quadratic equations:

Theorem 2. Let $\bar{w}_n^*$, $\tilde{w}_n^*$ denote the best possible approximations of the forms (14) and (15) respectively.

Assume:

(a) For each fixed $x \in H$ the linear operator $B(x) : H \to H$ defined by (13) is compact in H;

(b) the following estimate is true

$$\tilde{w}_n = y + B(\bar{w}_n^*, \bar{w}_n^*) . \qquad (16)$$

Then

$$\|\bar{w}^* - \tilde{w}_n^*\| \le \|w^* - \tilde{w}_n\| \le s_n \|w^* - \bar{w}_n^*\| \qquad (17)$$

where

$$s_n = s_n(w^*) = \|B(w^*)(I - P_n) + B(P_n(w^*))(I - P_n)\| \to 0 \qquad (18)$$

as $n \to \infty$.

Proof. We follow the proof of Theorem 1. The first inequality in (17) follows from the definition of $\tilde{w}_n^*$. To show the second inequality, using (4), (16) and the estimate

$$\bar{w}_n^* = P_n(w^*)$$

we get

$$\begin{aligned} w^* - \tilde{w}_n &= B(w^*, w^*) - B(\bar{w}_n^*, \bar{w}_n^*) \\ &= B(w^*, w^*) - B(P_n(w^*), P_n(w^*)) \\ &= [B(w^* + P_n(w^*)) - B(w^* + P_n(w^*))P_n](w^* - P_n(w^*)) \\ &= [B(w^*)(I - P_n) + B(P_n(w^*))(I - P_n)](w^* - P_n(w^*)) . \end{aligned} \qquad (19)$$

For fixed $w^* \in H$ the linear operators $B(w^*)$ and $B(P_n(W^*))$ are compact by definition. That is the results of Lemma 2 apply to immediately obtain that

$$s_n \leq \|B(w^*)(I - P_n)\| + \|B(P_n(w^*))(I - P_n)\| \to 0 + 0 = 0$$

as $n \to \infty$.

By taking norms in (19) we get

$$\|w^* - \tilde{w}_n\| \leq s_n \|w^* - P_n(w^*)\| = s_n \|w^* - \bar{w}_n^*\| .$$

The proof of the theorem is now complete.

Finally we complete this paper with the note that when H is an L^2 space our results have immediate applications to linear and quadratic integral equations with square-integrable kernels, say (see [2] and there references there).

REFERENCES

[1] Ahiezer, N.I., Glazma, I.M. Theory of linear operators in Hilbert space. Ungrar Publ. New York, 1961, (Vol. 1, Ch. 1).

[2] Argyros, I.K. Quadratic equations and applications to Chandrasekhar's and related equations. Bull. Austral. Math. Soc. Vol. 32, 2, (1985), 275-292.

[3] _____. Quadratic finite rank operator equations in Banach space. Tamkang J. Math. Vol. 18, 4, (1987), 8-19.

[4] Baker C.T.H. Expansion methods. In Numerical solution of integral equations (J. Walsh ed.) Clarendon Press, Oxford publ. 1974 (Ch. 7).

Rev. Acad. Ciencias. Zaragoza 45, (1990)

Fourier Hermite–Bessel Series for Meijer's G-Function

S.D. Bajpai

Department of Mathematics. University of Bahrain
P.O. Box 32038 Isa Town, BAHRAIN

Abstract

In this paper, we present a new class of Fourier Hermite–Bessel series for Meijer's G-function.

1 Introduction

The object of this paper is to introduce a new class of Fourier Hermite–Bessel for Meijer's G-function [3, pp. 206-222] and establish a Fourier Hermite–Bessel series of this class.

The following formulæ are required in the proof:

The integral [1, p.9, (2.1)]

$$\int_{-\infty}^{\infty} e^{-x^2} x^{u+2p} H_u(x) G_{p,q}^{m,n}\left[zx^{2\delta} \middle| \begin{matrix} a_p \\ b_p \end{matrix} \right] dx$$
$$= (2\pi)^{\frac{1}{2}(1-\delta)} 2^u \delta^{u+\rho} G_{p+2\delta,q+\delta}^{m,n+2\delta}\left[z\delta^{\delta} \middle| \begin{matrix} \Delta(2\delta, -2\rho-u), a_p \\ b_p, \Delta(\delta, -\rho) \end{matrix} \right], \tag{1}$$

where δ is a positive integer, $p+q < 2(m+n)$, $|\arg z| < (m+n-\frac{1}{2}p-\frac{1}{2}q)\pi$, $\rho = 0, 1, 2, \ldots$.

The integral [2, p. 37, (2.1)] :

$$\int_0^{\infty} y^{\sigma-1} J_\nu(y) G_{p,q}^{m,n}\left[zy^{2\lambda} \middle| \begin{matrix} a_p \\ b_p \end{matrix} \right] dy$$
$$= \frac{2^{\sigma-1}\lambda^{2(\sigma-1)}}{(2\pi)^{1-\lambda}} G_{p+5\lambda,q+\lambda}^{m,n+2\lambda}\left[\begin{matrix} \Delta(\lambda, \frac{2-\sigma-\mu-\nu}{2}), \Delta(\lambda, \frac{2-\sigma+\mu-\nu}{2}), a_p \\ \Delta(\lambda, \frac{3-\nu+\mu-\sigma}{2}), \Delta(\lambda, \frac{\nu-\mu-\sigma+2}{2}), \Delta(\lambda, \frac{\nu+\mu-\sigma+2}{2}) \\ b_q, \Delta(\lambda, \frac{3-\nu+\mu-\sigma}{2}) \end{matrix} \right] \tag{2}$$

where λ is a positive integer, $p+q < (m+n)$, $|\arg z| < (m+n-\frac{1}{2}p-\frac{1}{2}q)\pi$.

$$\mathrm{Re}(\sigma+\nu\pm\mu+2\lambda b_j) > 0, \quad (j = 1, \ldots, m) \qquad \mathrm{Re}(\sigma+2\lambda(a_j-1)) < 1, \quad (j = 1, \ldots, n).$$

2 Fourier Laguerre–Bessel series

The Fourier Laguerre–Bessel series to be established is

$$x^{u+2\rho}y^{\sigma}Y_u(y)G_{p,q}^{m,n}\left[zx^{2\delta}y^{2\lambda}\middle|\begin{matrix}a_p\\ b_p\end{matrix}\right]$$

$$=\frac{\delta^p 2^{\sigma}\lambda^{2(\sigma-1)}}{\sqrt{\pi}(2\pi)^{\frac{1}{2}(\delta+1)-\lambda}}\sum_{r=0}^{\infty}\sum_{t=0}^{\infty}\frac{\delta^r(\nu+2t+1)}{r!}H_r(x)J_{\nu+2t+1}(y)$$

$$G_{p+2\delta+5\lambda,q+\delta+\lambda}^{m,n+2\delta+2\lambda}\left[z\delta^{\delta}2^{2\lambda}\lambda^{4\lambda}\middle|\begin{matrix}\Delta(2\delta,-2\rho-r),\ \Delta(\lambda,\frac{1-\sigma-\mu-\nu-2t}{2})\\ \Delta(\lambda,\frac{1-\sigma+\mu-\nu-2t}{2}),\ a_p,\ \Delta(\lambda,\frac{2-\nu-2t+\mu-\sigma}{2}),\\ \Delta(\lambda,\frac{\nu+2t-\mu-\sigma+3}{2}),\ \Delta(\lambda,\frac{\nu+2t+\mu-\sigma+3}{2})\\ b_q,\ \Delta(\lambda,\frac{2-\nu-2t+\mu-\sigma}{2}),\Delta(\delta,-\rho)\end{matrix}\right], \tag{3}$$

valid under the conditions of (1) and (2) together with $\mathrm{Re}\nu > -1$.

PROOF. Let

$$\begin{aligned} f(x,y) &= x^{u+2\rho}y^{\sigma}Y_u(y)G_{p,q}^{m,n}\left[zx^{2\delta}y^{2\lambda}\middle|\begin{matrix}a_p\\ b_p\end{matrix}\right] \\ &= \sum_{r=0}^{\infty}\sum_{t=0}^{\infty}C_{r,t}H_r(x)J_{\nu+2t+1}(y). \end{aligned} \tag{4}$$

Equation (4) is valid, since $f(x,y)$ is continuous and of bounded variation in the region $-\infty < x < \infty$, $0 < y < \infty$.

Multiplying both sides of (4) by $y^{-1}J_{\nu+2v+1}(y)$, integrating with respect to y from 0 to ∞, and using (2) and the orthogonality property of the Bessel functions [4, p. 291, (6)]. Then multiplying both sides of the resulting expression by $e^{-x^2}H_u(x)$, integrating with respect to x from $-\infty$ to ∞, and using (1) and the orthogonality property of the Hermite polynomials [5, pp. 192-193, (5)], the value of $C_{r,t}$ is obtained. Substituting this value of $C_{r,t}$ in (4), the Fourier Hermite–Bessel series (3) is established.

Note: On applying the same procedure as above, we can establish three other forms of two–dimensional expansions of this class with the help of alternative forms of (1) and (2).

Since on specializing the parameters Meijer's G–function yields almost all special functions appearing in applied mathematics and physical sciences. Therefore, the result presented in the paper are of a general character and hence may encompass several cases of interest.

References

[1] Bajpai, S.D. (1970), An expansion formula for Meijer's G-function involving Hermite polynomials *Labdev J. Sci. Tech.*, **8A**, 9–11.

[2] Bajpai, S.D. (1974), Expansion formulæ for the products of Meijer's G-function and Bessel functions. *Portugal Math.*, **33**, 35–41.

[3] Erdélyi, A. (1953), *Higher transcendental functions* Vol 1. McGraw–Hill, New York

[4] Luke, Y.L. (1962), *Integrals of Bessel functions.* McGraw–Hill, New York

[5] Rainville, E.D.: 1960, *Special functions.* Chelsea Pub. Co. New York

Rev. Acad. Ciencias. Zaragoza 45, (1990)

TANGENT OF REFERENCES EMBEDS IN REFERENCES OF TANGENT: A GEOMETRICAL EXPLANATION

Fernando ETAYO GORDEJUELA

Departamento de Geometria y Topologia. Facultad de Ciencias Matematicas
Universidad Complutense de Madrid
28040 Madrid (SPAIN)

Abstract.

We study some properties of the embedding $J:TFM \longrightarrow FTM$, from the tangent bundle of linear frames bundle of a manifold M into the linear frames bundle of the tangent bundle of the manifold. First of all, we give a new definition of the embedding, which is intrinsic. After that, we study the situation when M is a Riemannian manifold. In particular, we prove that J is an affine mapping when M is flat and the manifolds TFM and FTM are endowed with suitable metrics; in this case, TFM is a totally geodesic submanifold of FTM and we also find Kahler structures for TFM and FTM.

0. Let M be a smooth, real, n-dimensional manifold and let $\pi_M:TM \longrightarrow M$ and $\tilde{\pi}_M:FM \longrightarrow M$ denote its tangent and linear frames bundles. Then, there exists a canonical embedding J from the $2(n+n^2)$-manifold TFM into the $2n+(2n)^2$-manifold FTM. Using local coordinates (x^i) in M, we call (x^i,X^i_j) the induced coordinates in FM, where X^i_j denotes a non-singular matrix of $GL(n,\mathbb{R})$ for each point of FM, and (x^i,y^i) will be the induced coordinates in TM. In both cases, i,j run through $\{1,\ldots,n\}$. Then, the embedding J is given by the following expression:

$$J(x^i,X^i_j,y^i,Y^i_j) = (x^i,y^i,\begin{pmatrix} X^i_j & 0 \\ Y^i_j & X^i_j \end{pmatrix}) \qquad (1)$$

where

$$\begin{pmatrix} X^i_j & 0 \\ Y^i_j & X^i_j \end{pmatrix}$$

denotes an element of the linear group $GL(2n,\mathbb{R})$, for each point. This construction is briefly indicated in [M], [Y-I].

In this work we want to give a geometrical definition of J, using the notions of vertical lift of vector fields from M to TM (1) and the canonical automorphism of TTM (2). We give a geometrical description of these constructions and, in 3, we obtain J.

Then, we study some properties of J. If (M,g) is a Riemannian manifold, TM and FM are endowed with "natural" metrics: The Sasaki and the Sasaki-Mok ones. In 4 we summarize the main properties of both metrics. In 5 we obtain a first theorem of J, which relates geodesics of these both metrics. And in 6,we show that J(TFM) is a totally geodesic submanifold of FTM, when they are endowed with suitable metrics and (M,g) is flat. This result is a corollary of proposition 4: in the above conditions, J is an affine mapping. And we show TFM and FTM as flat Kahler manifolds.

The notation is the usual in Differential Geometry Books. In particular, all the manifolds are real, paracompact and smooth. If $f:A \longrightarrow B$ is a map, f_* and f^* denote the tangent and cotangent induced maps.

Acknowledgments: I wish express my thanks to prof. Javier Lafuente for his useful geometric point of view.

I. GEOMETRICAL DEFINITION OF THE EMBEDDING

1. Vertical lifts of vector fields from M to TM.

This construction is given in [Y-I]: Let X denote a vector field on M, with local expression $X = (\partial/_{\partial x^i})\cdot X^i$. Then, its vertical lift to TM is the vector field $X^v = (\partial/_{\partial y^i})\cdot X^i$. The geometrical meaning X^v is the following: If $e \in T_pM$, then $(X^v)_e$ is the translation of the vector X_p from p to e in the vector space T_pM.

2. Canonical automorphism of TTM.

TTM has two vector bundle structures over TM, given by $\pi_{TM}:TTM \longrightarrow TM$ and by $T\pi_M:TTM \longrightarrow TM$. The second one is the tangent map of the tangent bundle $\pi_M:TM \longrightarrow M$. There exists a canonical automorphism $\sigma:TTM \longrightarrow TTM$, which

interchanges both structures [K], [G]: using local coordinates σ is given by $\sigma(x^i,y^i,\dot{x}^i,\dot{y}^i) = (x^i,\dot{x}^i,y^i,\dot{y}^i)$. Obviously, $\sigma\circ\sigma = id$.

The geometrical meaning of σ is the following:

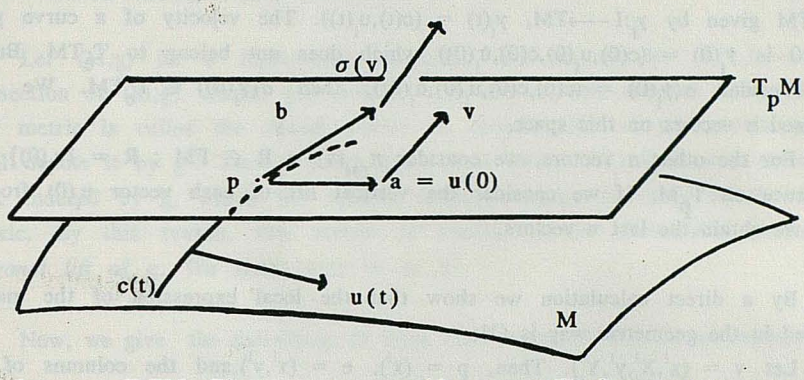


fig.1

Let $v \in T_aTM$. Then, we can consider $v = \gamma(0)$ for a curve γ on TM, $\gamma:I \longrightarrow TM$, $\gamma(t) = (c(t),u(t))$. So, we have $a = (c(0),u(0)) \in TM$ and $v = (c(0),u(0),\dot{c}(0),\dot{u}(0))$. Let $T\pi_M(v) = (c(0),\dot{c}(0)) = b$. Then, $\sigma(v) \in T_b,TM$ is the vector given by $\sigma(v) = (c(0),\dot{c}(0),u(0),\dot{u}(0))$.

Remark: If $a = b$, then $v = \sigma(v)$. The set of σ-invariant vector fields can be identified with the set of sections of the 2-jet bundle $T_2M \longrightarrow M$, i.e., the acceleration bundle of curves on M (see [B], [D-R]).

3. Definition of the embedding.

We want to obtain the embedding $J:TFM \longrightarrow FTM$ by a geometric definition. If $v \in TFM$ we must find a linear reference on a point of TM, naturally induced by the vector v. Let $v \in T_RFM$, where R denotes a linear frame on T_pM for some $p \in M$. Observe that $p = \tilde{\pi}_M(R)$. Then, we are going to define a linear frame on T_eTM, where $e = T\tilde{\pi}_M(v) \in T_pM$:

We have to find 2n linearly independent vectors on T_eTM.

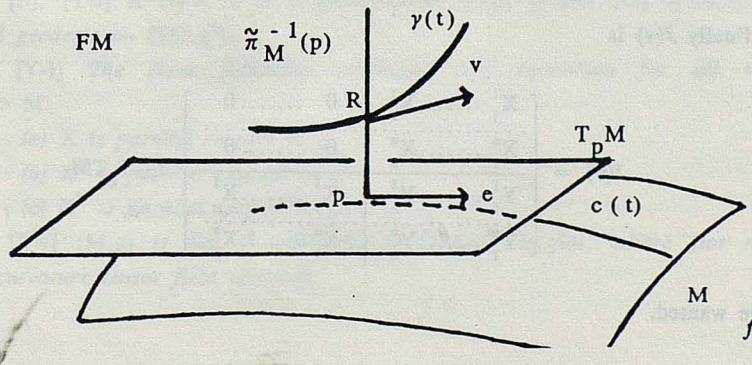


fig.2

As $v \in TFM$, we can consider it as the velocity of a smooth curve $\gamma: I \longrightarrow FM$, given by $\gamma(t) = (c(t), u_1(t), \ldots, u_n(t))$, $c(t)$ being a curve on M and $\{u_i(t)\}$ being a moving frame along that curve. Then, we can define some curves on TM given by $\gamma_i: I \longrightarrow TM$, $\gamma_i(t) = (c(t), u_i(t))$. The velocity of a curve γ_i at $t = 0$ is $\dot{\gamma}_i(0) = (c(0), u_i(0), \dot{c}(0), \dot{u}_i(0))$ which does not belong to T_eTM. But if we consider $\sigma(\dot{\gamma}_i(0) = (c(0), \dot{c}(0), u_i(0), \dot{u}_i(0))$, then $\sigma(\dot{\gamma}_i(0)) \in T_eTM$. We have obtained n vectors on this space.

For the other n vectors, we consider $\pi_{FM}(v) = R \in FM$; $R = \{u_i(0)\}$ is a reference on T_pM. If we consider the vertical lift of each vector $u_i(0)$ from p to e we obtain the last n vectors.

By a direct calculation we show that the local expression of the map J defined in the geometric way is (1):

Let $v = (x^i, X^i_j, y^i, Y^i_j)$. Then, $p = (x^i)$, $e = (x^i, y^i)$, and the columns of the matrix X^i_j are the coordinates of the vectors $u_i(0)$. We obtain:

$$\dot{\gamma}_i(t) = \left(x^i, \begin{bmatrix} X^1_i \\ \ldots \\ X^n_i \end{bmatrix}, y^i, \begin{bmatrix} Y^1_i \\ \ldots \\ Y^n_i \end{bmatrix}\right), \quad \text{and then,} \sigma(\dot{\gamma}_i(t)) = \begin{bmatrix} X^1_i \\ \ldots \\ X^n_i \\ Y^1_i \\ \ldots \\ Y^n_i \end{bmatrix} \in T_eTM$$

On the other hand, $(u_i(0))^v$ is

$$\begin{bmatrix} 0 \\ \ldots \\ 0 \\ X^1_i \\ \ldots \\ X^n_i \end{bmatrix} \in T_eTM$$

and finally $J(v)$ is

$$J(v) = \left[\begin{array}{ccc|ccc} X^1_1 & \ldots & X^1_n & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ X^n_1 & \ldots & X^n_n & 0 & \ldots & 0 \\ \hline Y^1_1 & \ldots & Y^1_n & X^1_1 & \ldots & X^1_n \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ Y^n_1 & \ldots & Y^n_n & X^n_1 & \ldots & X^n_n \end{array}\right] \in T_eTM$$

as we wanted.

II. RIEMANNIAN MANIFOLDS.

4. Sasaki and Sasaki-Mok metrics.

Let (M,g) be a Riemannian manifold and let ω be the Levi-Civita connection of (M,g). Sasaki [S] defined a metric on TM induced by g. Today, this metric is called the *Sasaki metric* or *diagonal lift* of the metric g. We shall denote it by g^d. Twenty years later, Mok [M1], [M2] defined a metric on FM, induced by g, which has several properties close to those of the Sasaki metric. By this reason, this metric is known as the *Sasaki-Mok metric* or *diagonal lift* of g. We shall write it as g^D.

Now, we give the definitions of these metrics and their main properties:

Definition 1: Let (M,g) be a Riemannian manifold. We call *Sasaki metric* g^d to the unique metric on TM which verifies:

(1) $g^d(X^V,Y^V) = [g(X,Y)]^V$

(2) $g^d(X^H,Y^H) = [g(X,Y)]^V$

(3) $g^d(X^V,Y^H) = 0$

for all vector fields X, Y $\in \mathcal{T}_0^1(M)$, where V is the vertical lift defined in 1 and H is the horizontal lift respect to ω.

Proposition 1. *With the above notation,*

(1) [S], [Y-I] (TM,g^d) *is a Riemannian manifold.*

(2) [M-T] g^d *is the unique metric on* TM *such that:*

(a) vertical and horizontal distributions are orthogonal.

(b) the induced metric on the fibers is euclidean.

(c) $\pi_M:(TM,g^d) \longrightarrow (M,g)$ *is a Riemannian submersion.*

(3) [B] *Fibers are totally geodesics.*

(4) [B], [Y-I] *A curve* c *is a geodesic on* (M,g) *if and only if* $(c,\dot{c})$ *is a horizontal geodesic on* (TM,g^d).

(5) [Y-I] *The three following conditions are equivalent for all vector field* X *on* M*:*

(a) X *is parallel respect to* g.

(b) X^V *is parallel respect to* g^d.

(c) X^H *is parallel respect to* g^d.

(6) [Kw] (M,g) *is flat if and only if* (TM,g^d) *is flat, where flat means that the curvature tensor field vanishes.*

Definition 2. Let (M,g) be a Riemannian manifold. We call *Sasaki-Mok metric* g^D to the unique metric on FM which verifies:

(1) $g^D(X^{(\alpha)},Y^{(\beta)}) = \delta^{\alpha\beta}\ [g(X,Y)]^V$

(2) $g^D(X^H,Y^H) = [g(X,Y)]^V$

(3) $g^d(X^{(\alpha)},Y^H) = 0$

for all vector fields $X, Y \in \mathcal{T}^1_0(M)$, where V s the vertical lift defined in 1, (α) and (β) denote the α- and β-lifts to FM (see [C-D-L]) and H is the horizontal lift respect to ω.

We have used this definition because it seems to definition 1. There are other equivalent definitions (see [C-D-L], [C-L], [M1])

Proposition 2. *With the above notation:*

(1) [M1] (FM,g^D) *is a Riemannian manifold.*

(2) [M1] $\tilde{\pi}_M:(FM,g^D)\longrightarrow(M,g)$ *is a Riemannian submersion.*

(3) [M1] *A curve* c *is a geodesic on* (M,g) *if and only if its horizontal lift* $\tilde{c}$ *is a geodesic on* (FM,g^D).

(4) [M1], [C-D-L] (M,g) *is flat if and only if* (FM,g^D) *is flat.*

Observe that condition (3) is equivalent to the following [K-N]: *every horizontal lift of* c *is an integral curve of a standard horizontal vector field.*

Remark. Definitions of both metrics can be given for a linear connection different from the Levi-Civita connection of the metric on M (see [Do], [C-D-L]), but we use this restrictive meaning, which is enough for our purposes.

5. The embedding *J* when M is a Riemannian manifold.

In this section we prove a theorem which relates both horizontal lifts of a geodesic on a Riemannian manifold (M,g) to the tangent and frame bundles.

Proposition 3. *Let* (M,g) *be a Riemannian manifold and let* $c:I\longrightarrow M$ *be a geodesic. Consider its horizontal lift* $\tilde{c}$ *to* FM *and let* v(t) *denote the tangent vector of* $\tilde{c}$ *at the point* $\tilde{c}(t) \in FM$. *Then* $\{J(v(t)) / t \in I\}$ *is the parallel respect to* g^d *moving frame along the geodesic* $(c,\dot{c})$ *of* (TM,g^d) *defined by the vertical and horizontal lifts to* TM *of the vectors of the frame* $\tilde{c}(t)$ *on the point* $c(t) \in M$.

Proof.

The curve $\{\tilde{c}(t) / t \in I\}$ represents a parallel respect to g moving frame along the curve $\{c(t) / t \in I\}$. Moreover, by propositions 1 and 2 we know that $\{\tilde{c}(t) / t \in I\}$ is a geodesic on (FM,g^D) and $\{(c(t),\dot{c}(t)) / t \in I\}$ is a geodesic on (TM,g^d).

As $(T\tilde{\pi}_M)(v(t)) = (c(t),\dot{c}(t))$, then $J(v(t))$ is a frame on $T_{(c(t),\dot{c}(t))}TM$, and when t runs through I, $J(v(t))$ is a moving frame along $(c(t),\dot{c}(t))$. This frame has 2n vectors: the last n ones are the vertical lift of the n vectors of the frame $\tilde{c}(t)$ on the point c(t), as we have seen in 3. And by (5) of proposition 1 we know that these are parallel respect to g^d.

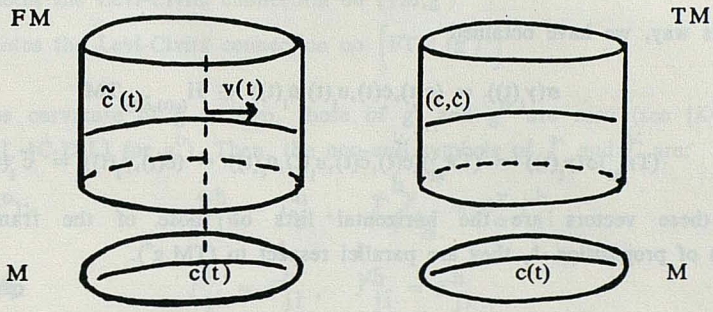


fig.3

For the other n vectors the idea is the following: With the notation of 3, we have

$$\gamma(t) = \tilde{c}(t) = (c(t),u_1(t),\ldots,u_n(t))$$

which is a parallel respect to g moving frame along the curve $\{c(t) / t \in I\}$, and

$$\gamma_i(t) = \tilde{c}_i(t) = (c(t),u_i(t)) \in T_{c(t)}M$$

which is a parallel respect to g vector field along the curve $\{c(t) / t \in I\}$. In this way, the frame $\tilde{c}(t)$ on c(t) has n vectors: $\tilde{c}_1(t),\ldots,\tilde{c}_n(t)$

We want to prove that the last n vectors, $\sigma(\gamma_i(t))$, of the frame $J(v(t))$ are the horizontal lifts of the n vectors $\tilde{c}_i(t)$ of the frame on c(t), from this point c(t) to the point $(c(t),\dot{c}(t)) \in TM$.

Let V be the vertical projector defined by the Levi-Civita connection ω on (M,g): as it is well known, $V:TTM \longrightarrow TTM$, and its kernel is the horizontal distribution given by the connection. If $\Gamma^{\alpha}_{\beta\gamma}$ denotes the Christoffel symbols of ω, the horizontal lift of the vector $(c(t),u_i(t))$ from the point $c(t) \in M$ to the point $(c(t),u_i(t))$ is the horizontal vector

$$(c(t),u_i(t),\dot{c}(t),\dot{u}_i(t)) \in T_{(c(t),u_i(t))}TM$$

and then,

$$V(c(t),u_i(t),\dot{c}(t),\dot{u}_i(t)) = \frac{\partial}{\partial x^\alpha}\cdot 0 + \frac{\partial}{\partial y^\alpha}\cdot\left[\dot{u}_i^\alpha(t) + \Gamma^\alpha_{\beta\gamma}(c(t))\cdot u_i^\gamma(t)\cdot\dot{c}^\beta(t)\right] =$$

$$= 0 \in T_{(c(t),u_i(t))}TM$$

Now,

$$V(c(t),\dot{c}(t),u_i(t),\dot{u}_i(t)) = \frac{\partial}{\partial x^\alpha}\cdot 0 + \frac{\partial}{\partial y^\alpha}\cdot\left[\dot{u}_i^\alpha(t) + \Gamma^\alpha_{\beta\gamma}(c(t))\cdot\dot{c}^\gamma(t)\cdot\dot{u}_i^\beta(t)\right] =$$

$$= 0 \in T_{(c(t),\dot{c}(t))}TM$$

In this way, we have obtained

$$\sigma(\gamma_i(t)) = (c(t),\dot{c}(t),u_i(t),\dot{u}_i(t)) \in H_{(c(t),\dot{c}(t))}TM$$

$$(T\pi_M)\sigma(\gamma_i(t)) = (T\pi_M)(c(t),\dot{c}(t),u_i(t),\dot{u}_i(t)) = (c(t),u_i(t)) = \tilde{c}_i(t)$$

i.e., these vectors are the horizontal lifts of those of the frame. Finally, by (5) of proposition 1, they are parallel respect to (TM,g^d).

QED

Caution. In general, σ:TTM $\longrightarrow$TTM does not map horizontal (resp. vertical) vectors on horizontal (resp. vertical) vectors. It is true in the case of the proof of proposition 3: if we lift a parallel vector field u(t) along the geodesic c(t) to the points (c(t),u(t)) and (c(t),ċ(t)) we obtain two horizontal vectors such that σ map each of them on the other one.

6. The embedding *J* when M is a flat Riemannian manifold.

The following theorem proves that *J* is an affine mapping, when (M,g) has vanishing curvature tensor field and we consider TFM and FTM endowed with these metrics: TFM with the Sasaki metric of the Sasaki-Mok metric of g, i. e., $\left[TFM,(g^D)^d\right]$, and FTM with the Sasaki-Mok metric of the Saski metric of g, i. e., $\left[FTM,(g^d)^D\right]$. As a consequence, we shall obtain that *J*(TFM) is a totally geodesic submanifold of FTM

Proposition 4. *Let* (M,g) *be a flat Riemannian manifold. Then J is an affine mapping, i.e., J maps geodesics of* $\left(TFM,(g^D)^d\right)$ *on geodesics of* $\left(FTM,(g^d)^D\right)$.

Proof.

We introduce the following notation:

Γ denotes the Levi-Civita connection on (M,g)

$\overset{*}{\Gamma}$ denotes the Levi-Civita connection on (FM,g^D)

Γ denotes the Levi-Civita connection on $\left(TFM,(g^D)^d\right)$

$\overset{*}{\Gamma}$ denotes the Levi-Civita connection on (TM,g^d)

$\hat{\Gamma}$ denotes the Levi-Civita connection on $\left(FTM,(g^d)^D\right)$

As the curvature of g is zero, those of g^d and g^D are zero (see [Kw] for g^d and [M1], [C-D-L] for g^D). Then, the non-null symbols of $\overset{*}{\Gamma}$ and $\overset{*}{\Gamma}$ are:

(FM,g^D): $\quad \overset{*}{\Gamma}{}^{h}_{ji} = \Gamma^{h}_{ji}\,, \quad \overset{*}{\Gamma}{}^{h\ \gamma}_{j\ i_\alpha} = \delta^\gamma_\alpha\, \Gamma^{h}_{ji}$

(TM,g^d): $\quad \overset{*}{\Gamma}{}^{h}_{ji} = \Gamma^{h}_{ji}\,, \quad \overset{*}{\Gamma}{}^{\bar{h}}_{j\bar{i}} = \Gamma^{h}_{ji}$

where i,j,h,α,β,γ run through $\{1,\ldots,n\}$ and $\bar{i},\bar{j},\bar{h}$ run through $\{n+1,\ldots,2n\}$, n being the dimension of M.

Now, it is easy to find the non-null symbols of the connections on TFM and FTM. We obtain:

$\left(TFM,(g^D)^d\right)$: $\quad \Gamma^{h}_{ji} = \overset{*}{\Gamma}{}^{h}_{ji} = \Gamma^{h}_{ji}\ ;\ \Gamma^{\bar{h}}_{j\bar{i}} = \overset{*}{\Gamma}{}^{h}_{ji} = \Gamma^{h}_{ji}\ ;$

$$\Gamma^{h\ \gamma}_{j\ i_\alpha} = \overset{*}{\Gamma}{}^{h\ \gamma}_{j\ i_\alpha} = \delta^\gamma_\alpha\, \Gamma^{h}_{ji}\ ;\ \Gamma^{\bar{h}\ \gamma}_{j\ \bar{i}_\alpha} = \overset{*}{\Gamma}{}^{h\ \gamma}_{j\ i_\alpha} = \delta^\gamma_\alpha\, \Gamma^{h}_{ji}$$

$\left(FTM,(g^d)^D\right)$: $\quad \hat{\Gamma}^{h}_{ji} = \overset{*}{\Gamma}{}^{h}_{ji} = \Gamma^{h}_{ji}\ ;\ \hat{\Gamma}^{\bar{h}}_{j\bar{i}} = \overset{*}{\Gamma}{}^{\bar{h}}_{j\bar{i}} = \Gamma^{h}_{ji}\ ;$

$$\hat{\Gamma}^{h\ \gamma}_{j\ i_\alpha} = \delta^\gamma_\alpha\, \overset{*}{\Gamma}{}^{h}_{ji} = \delta^\gamma_\alpha\, \Gamma^{h}_{ji}\ ;\ \hat{\Gamma}^{\bar{h}\ \gamma}_{j\ \bar{i}} = \delta^\gamma_\alpha\, \overset{*}{\Gamma}{}^{\bar{h}}_{j\bar{i}} = \delta^\gamma_\alpha\, \Gamma^{h}_{ji}\ ;$$

$$\hat{\Gamma}^{h\ \bar{\gamma}}_{j\ i_{\bar{\alpha}}} = \delta^{\bar{\gamma}}_{\alpha}\, \overset{*}{\Gamma}{}^{h}_{ji} = \delta^\gamma_\alpha\, \Gamma^{h}_{ji}\ ;\ \hat{\Gamma}^{\bar{h}\ \bar{\gamma}}_{j\ \bar{i}_{\alpha}} = \delta^{\bar{\gamma}}_{\alpha}\, \overset{*}{\Gamma}{}^{\bar{h}}_{j\bar{i}} = \delta^\gamma_\alpha\, \Gamma^{h}_{ji}\ ;$$

Using the above constructions, we can prove the theorem. We must show that if $c:I \longrightarrow TFM$ is a geodesic, then $(J \circ c):I \longrightarrow FTM$ is also a geodesic, i. e., if c satisfies the geodesic equations on $\left[TFM,(g^D)^d\right]$, $(J \circ c)$ verifies those of geodesic on $\left[FTM,(g^d)^D\right]$.

Geodesic equations on $\left[TFM,(g^D)^d\right]$:

Let $c(t) = (x^h(t),X^h_\gamma(t),y^h(t),Y^h_\gamma(t))$. It is a geodesic if and only if it verifies all the equations of the following four types:

$$(1)\quad \frac{\partial^2 x^h(t)}{\partial t^2} + \Gamma^h_{ji} \frac{\partial x^j(t)}{\partial t} \frac{\partial x^i(t)}{\partial t} = 0$$

$$(2)\quad \frac{\partial^2 X^h_\gamma(t)}{\partial t^2} + \Gamma^{h\gamma}_{j\ i_\alpha} \frac{\partial x^j(t)}{\partial t} \frac{\partial X^i_\alpha(t)}{\partial t} = 0$$

$$(3)\quad \frac{\partial^2 y^h(t)}{\partial t^2} + \Gamma^{\bar{h}}_{j\bar{i}} \frac{\partial x^j(t)}{\partial t} \frac{\partial y^i(t)}{\partial t} = 0$$

$$(4)\quad \frac{\partial^2 Y^h_\gamma(t)}{\partial t^2} + \Gamma^{\bar{h}\gamma}_{j\ \bar{i}_\alpha} \frac{\partial x^j(t)}{\partial t} \frac{\partial Y^i_\alpha(t)}{\partial t} = 0$$

Geodesic equations on $\left[FTM,(g^d)^D\right]$:

Let $\bar{c}(t) = (x^h(t),y^h(t),\begin{bmatrix} X^h_\gamma(t) & X^{\bar{h}}_\gamma(t) \\ X^{\bar{h}}_\gamma(t) & X^{\bar{h}}_{\bar{\gamma}}(t) \end{bmatrix})$ be a curve on FTM. It is a geodesic if and only if it verifies all the equations of the following six types:

$$(I)\quad \frac{\partial^2 x^h(t)}{\partial t^2} + \hat{\Gamma}^h_{ji} \frac{\partial x^j(t)}{\partial t} \frac{\partial x^i(t)}{\partial t} = 0$$

$$(II)\quad \frac{\partial^2 y^h(t)}{\partial t^2} + \hat{\Gamma}^{\bar{h}}_{j\bar{i}} \frac{\partial x^j(t)}{\partial t} \frac{\partial y^i(t)}{\partial t} = 0$$

$$(III)\quad \frac{\partial^2 X^h_\gamma(t)}{\partial t^2} + \hat{\Gamma}^{h\gamma}_{j\ i_\alpha} \frac{\partial x^j(t)}{\partial t} \frac{\partial X^i_\alpha(t)}{\partial t} = 0$$

$$\text{(IV)} \qquad \frac{\partial^2 X^{\bar{h}}_{\bar{\gamma}}(t)}{\partial t^2} + \hat{\Gamma}^{\bar{h}\bar{\gamma}}_{j\,\bar{i}\bar{\alpha}} \frac{\partial x^j(t)}{\partial t} \frac{\partial X^{\bar{i}}_{\bar{\alpha}}(t)}{\partial t} = 0$$

$$\text{(V)} \qquad \frac{\partial^2 X^{\bar{h}}_{\gamma}(t)}{\partial t^2} + \hat{\Gamma}^{\bar{h}\gamma}_{j\,\bar{i}\alpha} \frac{\partial x^j(t)}{\partial t} \frac{\partial X^{\bar{i}}_{\alpha}(t)}{\partial t} = 0$$

$$\text{(VI)} \qquad \frac{\partial^2 X^{\bar{h}}_{\bar{\gamma}}(t)}{\partial t^2} + \hat{\Gamma}^{\bar{h}\bar{\gamma}}_{j\,\bar{i}\bar{\alpha}} \frac{\partial x^j(t)}{\partial t} \frac{\partial X^{\bar{i}}_{\bar{\alpha}}(t)}{\partial t} = 0$$

Assume that c is a geodesic. Then, it is easy to check that $J \circ c$ verifies these equations: (I) by using (1), (II) by (3), (III) and (VI), which are equal, by (2), (IV) is the equation $0 = 0$, and (V) by (4).

QED

Corollary 1. *With the above assumptions,* $\left[\text{TFM},(g^D)^{\bar{d}}\right]$ *is a totally geodesic submanifold of* $\left[\text{FTM},(g^{\bar{d}})^D\right]$.

Proof.

Let $\bar{v} \in T_{\bar{p}}(J\text{TFM})$ and let $\bar{\gamma}$ be the geodesic on $\left[\text{FTM},(g^{\bar{d}})^D\right]$ beginning on $\bar{p}$ with velocity $\bar{v}$. We have to prove that $\bar{\gamma}(t)$ belongs to JTFM for small t.

As J is a diffeomorphism between TFM and JTFM, there exists a unique $p \in \text{TFM}$ and a unique vector $v \in T_p\text{TFM}$ such that $J(p) = \bar{p}$ and $J_*(v) = \bar{v}$. Let γ be the geodesic on $\left[\text{TFM},(g^D)^{\bar{d}}\right]$ beginning on p with velocity v. Then, by the proposition, $J \circ \gamma$ is a geodesic on $\left[\text{FTM},(g^{\bar{d}})^D\right]$ which lies on JTFM and $\bar{\gamma} = J \circ \gamma$.

QED

Corollary 2. *With the above assumptions, the Riemannian connection of* $\left[\text{TFM},(g^D)^{\bar{d}}\right]$ *and that of* $\left[\text{TFM},J^*(g^{\bar{d}})^D\right]$ *coincide.*

Proof.

It follows from proposition 4 and the following theorem [K-N, prop. VII 8.8]: "If B is a Riemannian manifold and A is a totally geodesic submanifold of B, the Riemannian connection of A with respect to the induced Riemannian metric coincides with the induced connection".

We apply the above result to $A = \left[\text{TFM},(g^D)^{\bar{d}}\right]$ and $B = \left[\text{FTM},(g^{\bar{d}})^D\right]$.

QED

Caution. This condition is necessary, but not sufficient, to prove that $J\left[TFM,(g^D)^d\right]$ is a Riemannian submanifold of $\left[FTM,(g^d)^D\right]$, i.e., $(g^D)^d = J^*\left[(g^d)^D\right]$.

We study now the complex structures which can be defined on TFM and FTM induced by the Riemannian connection ω of (M,g). First of all, we show the known results in this topic:

(a) [Do] If (M,g) is a flat Riemannian manifold, g^d is the Riemannian metric associated to a hermitian metric h_g on TM. Moreover, (TM,h_g) is a complex and Kahler manifold, that can be considered as the natural complexification of the given manifold.

(b) [C-D-L] If (M,h_g) is a flat Kahler manifold with associated Riemannian metric g, then the Sasaki-Mok metric g^D is the associated Riemannian metric of a hermitian metric H on FM. Moreover, (FM,H) is a flat Kahler manifold.

Suppose that (M,g) is a flat Riemannian manifold. Then, by (4) of proposition 2, (FM,g^D) is flat. And by (a) $\left[TFM,(g^D)^d\right]$ is a Kahler manifold, which is flat by (6) of proposition 1.

On the other hand, by (a), (TM,g^d) is a flat Kahler manifold. Now, by (b), $\left[FTM,(g^d)^D\right]$ is a flat Kahler manifold.

Then $J\left[TFM,(g^D)^d\right]$ is a totally geodesic submanifold of $\left[FTM,(g^d)^D\right]$, and J is a real embedding between two Kahler manifolds. But this is not sufficient to prove that the first manifold is a complex submanifold of the second one.

Open problems.

(1) Prove that $J\left[TFM,(g^D)^d\right]$ is a complex submanifold of $\left[FTM,(g^d)^D\right]$ when (M,g) is flat.

(2) Prove that $J\left[TFM,(g^D)^d\right]$ is a totally geodesic submanifold of $\left[FTM,(g^d)^D\right]$ when (M,g) is not flat. Moreover, it is a Riemannian submanifold.

These two problems can be studied by using local coordinates, but the calculation of all the needed symbols is disgusting.

References

[B] Besse, A.: *Manifolds all of whose geodesics are closed.* Springer, Berlin, Heildelbrg, 1978.

[C-D-L] Cordero, L.A. ; Dodson, C.T.J. & de León, M.: *Differential Geometry of Frame Bundles*. Kluwer, Dordrecht, 1989.

[C-L] Cordero, L.A. & de León, M.: Lifts of Tensor Fields to the Frame Bundle. *Rend. Circ. Mat. Palermo*, **32** (1983) 236-271.

[D-R] Dodson, C.T.J. & Radivoiovici, M.S.: Second Order Tangent Structures. *International Journal of Theoretical Physics*, **21**, N.2 (1982) 151-161.

[Do] Dombrowski, P.: On the Geometry of the Tangent Bundle . *Jour. reine Angew. Math.* **210** (1962) 73-88.

[G] Godbillon, C: *Geometrie Differentielle et Mecanique Analytique*. Hermann, Paris, 1969.

[K] Kobayashi, S: Theory of Connections. *Ann. Mat. Pura Appl.* **43** (1957) 119-194.

[K-N] Kobayashi, S. & Nomizu, K.: *Foundations of Differential Geometry I & II.* J. Wiley, N. York, 1963 and 1969.

[Kw] Kowalski, O.: Curvature of the induced Riemannian metric on the Tangent Bundle of a Riemannian manifold. *Jour. reine Angew. Math.* **250** (1971) 124-129.

[M1] Mok, K.P.: On the Differential Geometry of Frame Bundle of Riemannian Manifolds. *Jour. reine Angew. Math.* **302** (1978) 16-31.

[M2] Mok, K.P.: Complete Lifts of Tensor Fields and Connections to the Frame Bundle. *Proc. London. Math. Soc.* **38** (1979) 72-88.

[M] Morimoto, A: *Prolongation of Geometric Structures*. Math. Inst. Nagoya Univ., 1969.

[M-T] Musso, E & Tricerri, F.: Riemannian Metrics on Tangent Bundles. *Ann. Mat. Pura Appl.* **150** (1988) 1-19.

[O'N] O'Neill, B.: *Semi-Riemannian Geometry with applications to Relativity*. Ac. Press, N. York, 1983.

[S] Sasaki, S.: On the Differential Geometry of Tangent Bundles of Riemannian Manifolds. *Tohoku Math. Jour.* **10** (1958) 338-354.

[Y-I] Yano, K. & Ishihara, S.: *Tangent and Cotangent Bundles. Differential Geometry*. Dekker, N. York, 1973.

Rev. Acad. Ciencias. Zaragoza 45, (1990)

NEW CONTINUOUS EXTENSIONS FOR FIFTH–ORDER RK FORMULAS

M. CALVO, J. I. MONTIJANO, L. RANDEZ

Departamento de Matemática Aplicada
Universidad de Zaragoza. 50009 Zaragoza. Spain

Abstract. In this paper the construction of continuous extensions of order five for fifth order Runge–Kutta methods is considered. A technique for the construction of general families of interpolants is proposed. This technique is applied to the derivation of general families of interpolants for the fifth order solutions of the pairs DOPRI5(4)7M of Dormand and Prince [2] and RKF4(5)#2 of Fehlberg [5]. It is shown that these families contain the interpolants given by Calvo–Montijano–Rández[1], Shampine[9] and Enright et al [3]. Finally new interpolants for the above pairs are determined by choosing the free parameters so as to minimize some measure of the local error of the continuous solution.

Key words. Runge–Kutta methods, continuous extensions, interpolation

AMS(MOS) subject classifications. 65L07, CR517

1.Introduction.

Consider the initial value problem for a system of ordinary differential equations

$$\begin{aligned} y'(t) &= f(t, y(t)), \qquad t \in [t_0, t_0 + T] \\ y(t_0) &= y_0, \end{aligned} \tag{1.1}$$

where the function f is assumed to be as smooth as necessary. It is well known that most explicit Runge–Kutta (RK) codes for the numerical solution of (1.1) are designed to produce approximations y_n, $n = 0, \ldots, N$ to the solution of (1.1) at a discrete set of points t_n, $n = 0, \ldots, N$ in the integration interval $[t_0, t_0 + T]$.The code proceeds step by step selecting automatically the step sizes $h_n = t_{n+1} - t_n$ of the grid (t_n) so that an estimate of the local error is smaller than a given tolerance. However, in many applications (e.g. problems with dense output, discontinuous IVPs[4], differential equations with deviating arguments) a continuous approximation in the entire interval, or some part of it, is required. For this reason there has been recently much interest

Partially supported by C.Y.C.I.T. PS87/0060.

in the problem of obtaining, in an inexpensive way, accurate approximations to the solution between grid points and many authors (see e.g.[1][3][6][7][9][10]) have proposed techniques and interpolants for several pairs of RK formulas. The aim of this paper is to propose a general approach to the problem of constructing interpolants for given RK formulas of order five with minimum computational cost. It has been proved by Owren and Zennaro [8] that a continuous RK method of order five requires at least 8 stages and since the usual fifth–order RK formulas have at least six stages, the interpolants for these formulas need at least two additional function evaluations per step. In this paper we have considered continuous extensions with three additional stages but one of the new function evaluations coincides with the first of the next step, therefore our continuous methods have actually two effective additional stages. The proposed approach is applied to obtain families of order five for the widely used pairs DOPRI5(4)7M of Dormand and Prince[2] and RKF4(5) of Fehlberg [5].Then, optimal methods in both families have been selected by choosing the free parameters so as to minimize a certain norm of the leading coefficient of the local truncation error of the continuous solution. Finally, graphs of the local error of the new interpolants are compared with those of the interpolants proposed by Shampine [9] and Calvo–Montijano–Rández [1] for the DOPRI5(4)7M and Enright et al [3] for the RKF4(5)#2. It is shown that between mesh points the local error of the new interpolants behaves better than the local error of the interpolants given by the above authors.

2. The construction of interpolants for RK formulas of order five.

Let (y_n), $n = 0, \ldots, N$ be approximations to the solution of (1.1) at the grid–points t_n, $n = 0, \ldots, N$ obtained by using the explicit m–stage RK formula (A, b) of order five. Then, the formula advances the numerical solution from the grid–point t_n to t_{n+1} by the formulas

$$\begin{aligned} y_{n+1} &= y_n + h_n \sum_{j=1}^{m} b_j f_{n,j}, \\ f_{n,1} &= f(t_n, y_n), \\ f_{n,j} &= f(t_n + c_j h_n, y_n + h_n \sum_{k=1}^{j-1} a_{jk} f_{n,j}) \quad j = 2, \ldots, m, \end{aligned} \tag{2.1}$$

where $c_j = \sum_{k=1}^{j-1} a_{jk}$.

It should be noted that most codes based on RK formulas of order five actually employ a pair of orders 4 and 5, but the lower order formula is used only to estimate the local error and the integration advances with the formula of higher order, hence our goal is to give a continuous extension of the fifth–order solution.

We are interested in $\mathcal{C}^1$ continuous extensions $y_h(t)$ of the grid solution (y_n) with uniform order five, so that its restriction to each interval $[t_n, t_{n+1}]$ is a polinomyal function. Moreover, as the integration proceeds in a step by step mode, we want to determine the interpolant $y_h(t)$ in the interval $[t_n, t_{n+1}]$ just after the step from t_n to t_{n+1} has been completed, performing some additional function evaluations. In this way, the extra work required to get the continuous solution in the interval $[t_n, t_{n+1}]$ will be done only if we need the continuous solution in this interval.

Firstly, the continuity assumption on $y_h(t)$ implies that in each interval $[t_n, t_{n+1}]$ we have

$$\begin{aligned} y_h(t_n) &= y_n, & y_h(t_{n+1}) &= y_{n+1}, \\ y'_h(t_n) &= f(t_n, y_n), & y'_h(t_{n+1}) &= f(t_{n+1}, y_{n+1}). \end{aligned} \tag{2.2}$$

The global solution $y_h(t)$ has uniform order five if the local error between two consecutive steps satisfies

$$|e(\theta h_n)| = |u_n(t_n + \theta h_n) - y_h(t_n + \theta h_n)| = \mathcal{O}(h_n^6), \tag{2.3}$$

uniformly in $\theta \in [0,1]$, where $u_n(t)$ is the local solution at (t_n, y_n) i.e. the solution of the IVP

$$\begin{aligned} u'_n(t) &= f(t, u_n(t)), \qquad t \geq t_n \\ u_n(t_n) &= y_n. \end{aligned} \tag{2.4}$$

To construct an interpolant in the interval $[t_n, t_{n+1}]$, several approaches have been proposed. In the interpolant DPS for the DOPRI5(4) pair given by Shampine [9], this author takes a value $\sigma \in (0,1)$ and adds a new stage to (2.1).Then he shows that it is possible to choose the free parameters a_{8j}, $j = 1, \ldots, 8$ and $\hat{b}_j$, $j = 1, \ldots, 7$ so that

$$y_{n,2} = y_n + h_n \sum_{j=1}^{8} \hat{b}_j f_{n,j} \tag{2.5}$$

is a solution of order five at the point $t_{n,2} = t_n + \sigma h$. Moreover it is clear that

$$y'_{n,2} = f(t_{n,2}, y_{n,2}) \tag{2.6}$$

is also an approximation of the same order to $y'_h(t_{n,2})$. Now, setting $t_{n,0} = t_n$, $t_{n,1} = t_{n+1}$, Shampine constructs the interpolatory polynomial for the data (2.2), (2.5) and (2.6) ,that may be written in the form

$$y_h(t_n + \theta h_n) = \sum_{i=0}^{2} \Phi_{0,i}(\theta) y_{n,i} + h_n \Phi_{1,i}(\theta) y'_{n,i}, \tag{2.7}$$

with $\theta \in [0,1]$, where $\Phi_{0,i}$ and $\Phi_{1,i}$ are the corresponding polynomials of the Hermite basis. Finally he shows that (2.7) has uniform order five and therefore its interpolant is determined at the price of two extra function evaluations per step because $f(t_{n+1}, y_{n+1})$ coincides with the first evaluation of the following step.

The same idea is used by the present authors in [1] to derive a family of fifth order interpolants for the DOPRI5(4) method. However, our family of interpolants has more free parameters and these degrees of freedom allow to obtain a new interpolant whose local error is smaller than the local error of DPS interpolant.

It should be noted that substituting (2.1), (2.5) and (2.6) in the right hand side of (2.7) all the above interpolants may be written in the form

$$y_h(t_n + \theta h_n) = y_n + h_n \sum_{j=1}^{9} b_j(\theta) f_{n,j}, \tag{2.8}$$

where $b_j(\theta)$ are polynomials of degree ≤ 5 and $f_{n,j}$ can be considered as the function evaluations of a 9–stage RK formula whose tableau of coefficients is

$$\begin{array}{c|c} c & A \\ 1 & b^T \\ c_8 & a_8^T \\ \sigma & \hat{b}^T \\ \hline & b(\theta)^T. \end{array}$$

where $a_8^T = (a_{81}, \ldots, a_{87})$, $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_9)^T$.

On the other hand, Enright et al [3] have developped another procedure for the construction of interpolants of any order. We briefly describe it for our case of order five. The procedure starts with an interpolant of order four (note that in the pairs DOPRI5(4) and RKF4(5) this can be obtained without additional function evaluations). Taking σ_1 and σ_2, $(0 < \sigma_1 < \sigma_2 < 1)$ it is possible to get fourth order approximations

$$\begin{aligned} y_{n+\sigma_1} &= y_n + h_n \sum_j \alpha_{1j} f_{n,j}, \\ y_{n+\sigma_2} &= y_n + h_n \sum_j \alpha_{2j} f_{n,j}. \end{aligned}$$

at the points $t_{n+\sigma_1} = t_n + \sigma_1 h_n$ and $t_{n+\sigma_2} = t_n + \sigma_2 h_n$ and therefore

$$\begin{aligned} y'_{n+\sigma_1} &= f(t_{n+\sigma_1}, y_{n+\sigma_1}), \\ y'_{n+\sigma_2} &= f(t_{n+\sigma_2}, y_{n+\sigma_2}). \end{aligned}$$

are also fourth–order accurate approximations to the derivatives at $t_{n+\sigma_1}$ and $t_{n+\sigma_2}$ respectively. Now, we may construct the interpolatory polynomial with the data y_n, y'_n at t_n, $y'_{n+\sigma_1}$ at $t_{n+\sigma_1}$, $y'_{n+\sigma_2}$ at $t_{n+\sigma_2}$ and y_{n+1}, y'_{n+1} at t_{n+1}. Since the derivative data enter in the interpolation polynomial in the form hy', the interpolatory solution is again of order five. Notice that this procedure can be written also in the form (2.8), (2.9) where the last two stages of (2.9) are the coefficients α_{1j} and α_{2j} respectively.

Clearly, a more general technique to construct interpolants of order five consists in adding three additional stages to the original formula (which is assumed to have six stages) where one of the additional stages is the first evaluation of the next step. Now, the two remainder stages must be determined so that there exist σ_1, $\sigma_2 \in (0,1)$ and b^1, $b^2 \in \mathbb{R}^9$ such that

$$\begin{aligned} y_{n+\sigma_1} &= y_n + h_n \sum_{j=1}^{9} b_j^1 f_{n,j}, \\ y_{n+\sigma_2} &= y_n + h_n \sum_{j=1}^{9} b_j^2 f_{n,j}. \end{aligned} \tag{2.10}$$

are approximations of order five at the points $t_{n+\sigma_1} = t_n + \sigma_1 h_n$ and $t_{n+\sigma_2} = t_n + \sigma_2 h_n$. Then, with the data : y_n and y'_n at t_n, $y_{n+\sigma_1}$ at $t_{n+\sigma_1}$, $y_{n+\sigma_2}$ at $t_{n+\sigma_2}$ and y_{n+1}, y'_{n+1} at t_{n+1} we may construct the interpolatory polynomial which provides a continuous solution of uniform order five and requires eight effective function evaluations per step. In this approach, we are led to a new method whose Butcher tableau of coefficients has the form

$$\begin{array}{c|c} c & A \\ 1 & b^T \\ c_8 & a_8^T \\ c_9 & a_9^T \\ \hline & b(\theta)^T \end{array} \quad = \quad \begin{array}{c|c} \hat{c} & \hat{A} \\ \hline & b(\theta)^T \end{array}, \tag{2.11}$$

where $a_8^T = (a_{81}, \ldots, a_{87})$ and $a_9^T = (a_{91}, \ldots, a_{98})$. Observe that $b^1 = b(\sigma_1)$ and $b^2 = b(\sigma_2)$.

Finally, note that we have at our disposal 35 free parameters $(\sigma_1, \sigma_2, b^1, b^2, a_8, a_9)$ with 15 + 15 order conditions for the solutions (2.10). Since these order equations are non linear it is difficult, in general, to give a simple set of conditions which imply the existence of solution, however in most practical cases there are simplifying assumptions in the underlying solution of order five which make it possible to get in a simple way families of interpolants with several free parameters.

3. Interpolants for DOPRI5(4).

Let

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}, \tag{3.1}$$

be the Butcher tableau of coefficients of the fifth order solution of the pair DOPRI5(4). Note that this pair has seven stages but the solution of order five can be computed using exclusively the first six stages of the method, therefore the matrix A in (3.1) corresponds to an explicit RK formula with six stages. The Butcher tableau of our continuous extension of (3.1) with order five will have the form (2.11).

We assume that the matrix $\hat{A}$ satisfies the conditions

$$\begin{aligned} \hat{A}\hat{c} &= \frac{1}{2}(\hat{c}^2 - c_2^2\, e_2), \\ \hat{A}\hat{c}^2 &= \frac{1}{3}(\hat{c}^3 - c_2^3\, e_2), \end{aligned} \tag{3.3}$$

where $\hat{c}^q = (c_1^q, \ldots, c_9^q)^T$ and e_j is the j–unit canonical vector of components $(e_j)_i = \delta_{ij}$. These conditions hold for the matrix A and vector c and are usually referred to as the simplifying assumptions. Hence (3.3) are equivalent to the scalar equations

$$a_8^T \hat{c} = c_8^2/2 \quad a_8^T \hat{c}^2 = c_8^3/3, \tag{3.4$_1$}$$

$$a_9^T \hat{c} = c_9^2/2 \quad a_9^T \hat{c}^2 = c_9^3/3. \tag{3.4$_2$}$$

Now, we proceed to construct the desired interpolant for (3.1). Firstly, choosing $\sigma_1 \in (0,1)$, we set the conditions to be satisfied by $a_8 \in \mathbb{R}^8$ and $b(\sigma_1) \in \mathbb{R}^8$ so that

$$y_{n+\sigma_1} = y_n + h_n \sum_{i=1}^{8} b_i(\sigma_1) f_{n,i},$$

is a solution of order five. Taking into account the simplifying assumptions these conditions can be written in the form

$$
\begin{aligned}
&b_2(\sigma_1) = 0\\
&b(\sigma_1)^T c'^j = \sigma_1^{j+1}/(j+1) \qquad j = 0,\dots,4,\\
&b(\sigma_1)^T (c' \bullet A' e_2) = 0,\\
&b(\sigma_1)^T A' c'^3 = \sigma_1^5/20,\\
&b(\sigma_1)^T A' e_2 = 0,\\
&b(\sigma_1)^T A'^2 e_2 = 0.
\end{aligned}
\tag{3.5}
$$

where $A' \in \mathbb{R}^{8\times 8}$ is the submatrix of $\hat{A}$ obtained taking the first eight rows and columns of $\hat{A}$ and $c' \in \mathbb{R}^8$ has the first eight components of $\hat{c}$.Here, $u \bullet v$ denotes the vector of components $(u_1v_1, u_2v_2, \dots)$ where $u = (u_1, u_2, \dots)^T, v = (v_1, v_2, \dots)^T$.

To study the equations (3.4_1) and (3.5) we follow the ideas given in [1]. We introduce the new scalar variables $\gamma = b_8(\sigma_1)a_8^T c'^3$, $\gamma_1 = -b_8(\sigma_1)a_8^T A' e_2$ and $\gamma_2 = -b_8(\sigma_1)a_8^T e_2$. Then, since $c_7 = 1$ and denoting by $(x)_{i-j}$ the row vector $(x_i, \dots, x_j)$, equations (3.5) and (3.4_1) can be rewritten equivalently in the form

$$
\begin{pmatrix}
c_{3-7} & c_8 & 0 & 0\\
\vdots & \vdots & \vdots & \vdots\\
c_{3-7}^4 & c_8^4 & 0 & 0\\
(Ac^3)_{3-7} & 0 & 0 & 0\\
(A^2e_2)_{3-7} & 0 & -1 & 0\\
(Ae_2 \bullet c)_{3-7} & 0 & 0 & -c_8\\
(Ae_2)_{3-7} & 0 & 0 & -1
\end{pmatrix}
\begin{pmatrix}
b_3(\sigma_1)\\ \vdots\\ b_6(\sigma_1)\\ b_7(\sigma_1)\\ b_8(\sigma_1)\\ \gamma_1\\ \gamma_2
\end{pmatrix}
=
\begin{pmatrix}
\sigma_1^2/2\\ \vdots\\ \sigma_1^5/5\\ \sigma_1^5/20 - \gamma\\ 0\\ 0\\ 0
\end{pmatrix},
\tag{3.6}
$$

$$
\begin{pmatrix}
c_2 & c_3 & c_4 & c_5 & 1\\
c_2^2 & c_3^2 & c_4^2 & c_5^2 & 1\\
c_2^3 & c_3^3 & c_4^3 & c_5^3 & 1\\
0 & a_{32} & a_{42} & a_{52} & 0\\
1 & 0 & 0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
a_{82}\\ a_{83}\\ a_{84}\\ a_{85}\\ a_{87}
\end{pmatrix}
=
\begin{pmatrix}
c_8^2/2 - a_{86}c_6\\
c_8^3/3 - a_{86}c_6^2\\
\gamma/b_8(\sigma_1) - a_{86}c_6^3\\
-\gamma_1/b_8(\sigma_1) - a_{86}a_{62}\\
-\gamma_2/b_8(\sigma_1)
\end{pmatrix},
\tag{3.7}
$$

$$
b_1(\sigma_1) = \sigma_1 - \sum_{j=3}^{8} b_j(\sigma_1) \qquad b_2(\sigma_1) = 0 \qquad a_{81} = c_8 - \sum_{j=2}^{7} a_{8j}. \tag{3.8}
$$

It can be shown that the matrix of coefficients in (3.6) is singular only for $c_8 = 0$, $c_8 = 1$ and $c_8 = c_8^*$, where $c_8^* = 0.91661200243235649\dots$ is the real root of the polynomial $P(x) = -49950x^3 + 70945x^2 - 26322x + 2988$. Furthermore, the constant matrix in (3.7) is non-singular. In conclusion, we have obtained $y_{n+\sigma_1}$ with four free parameters $\sigma_1 \in (0,1)$, $c_8 (\neq 0, 1, c_8^*)$, γ and a_{86}.

Secondly, we choose $\sigma_2 \in (0,1)$, $\sigma_2 \neq \sigma_1$, and we set the conditions to be satisfied by $a_9 \in \mathbb{R}^9$ and $b(\sigma_2) \in \mathbb{R}^9$ so that

$$
y_{n+\sigma_2} = y_n + h_n \sum_{i=1}^{9} b_i(\sigma_2) f_{n,i},
$$

is a solution of order five. Proceeding as above, such a set of order conditions can be written again in the form (3.5) with $\hat{A}$, $\hat{c}$, σ_2 instead of A', c', σ_1 respectively. Then, putting $\alpha = b_9(\sigma_2)a_9^T \hat{c}^3$, $\alpha_1 = -b_9(\sigma_2)a_9^T \hat{A} e_2$, $\alpha_2 = -b_9(\sigma_2)a_9^T e_2$, the order conditions together with (3.4_2) become

$$\begin{pmatrix} c_{3-8} & 0 & 0 \\ \vdots & \vdots & \vdots \\ c_{3-8}^4 & 0 & 0 \\ (A'c^3)_{3-8} & 0 & 0 \\ (A'^2e_2)_{3-8} & -1 & 0 \\ (A'e_2 \bullet c)_{3-8} & 0 & -c_9 \\ (A'e_2)_{3-8} & 0 & -1 \end{pmatrix} \begin{pmatrix} b_3(\sigma_2) \\ \vdots \\ b_6(\sigma_2) \\ b_7(\sigma_2) \\ b_8(\sigma_2) \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \sigma_2^2/2 - b_9(\sigma_2)c_9 \\ \vdots \\ \sigma_2^5/5 - b_9(\sigma_2)c_9^4 \\ \sigma_2^5/20 - \alpha \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{3.9}$$

$$\begin{pmatrix} c_2 & c_3 & c_4 & c_5 & 1 \\ c_2^2 & c_3^2 & c_4^2 & c_5^2 & 1 \\ c_2^3 & c_3^3 & c_4^3 & c_5^3 & 1 \\ 0 & a_{32} & a_{42} & a_{52} & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{92} \\ a_{93} \\ a_{94} \\ a_{95} \\ a_{97} \end{pmatrix} = \begin{pmatrix} c_9^2/2 - a_{96}c_6 - a_{98}c_8 \\ c_9^3/3 - a_{96}c_6^2 - a_{98}c_8^2 \\ \alpha/b_9(\sigma_2) - a_{96}c_6^3 - a_{98}c_8^3 \\ -\alpha_1/b_9(\sigma_2) - a_{96}a_{62} - a_{98}a_{82} \\ -\alpha_2/b_9(\sigma_2) \end{pmatrix}, \tag{3.10}$$

$$b_1(\sigma_2) = \sigma_2 - \sum_{j=3}^{9} b_j(\sigma_2) \qquad b_2(\sigma_2) = 0 \qquad a_{91} = c_9 - \sum_{j=2}^{8} a_{9j}. \tag{3.11}$$

Here, the constant matrix in (3.10) is nonsingular again and the matrix in (3.9) depends on the parameters a_{8j} and c_9. then, the parameter c_9 has to be chosen so that this matrix be non singular.

Clearly in this process we may obtain $y_{n+\sigma_2}$ with six free parameters : $\sigma_2 (\in (0,1), (\sigma_1 \neq \sigma_2))$, $c_9 (\neq 0, 1)$, α, a_{96}, a_{98} and $b_9(\sigma_2)$.

In this way we have a family of fifth order methods with ten degrees of freedom ($\sigma_1, \sigma_2, c_8, c_9, \gamma, \alpha, a_{86}, a_{96}, a_{98}$ and $b_9(\sigma_2)$), and for each set $\mu = (\sigma_1, \sigma_2, c_8, c_9, \gamma, \alpha, a_{86}, a_{96}, a_{98}, b_9(\sigma_2))$ we may define a continuous solution $y_\mu(t_n + \theta h), \theta \in [0,1]$ whose local error will be given by

$$u_n(t_n + \theta h) - y_\mu(t_n + \theta h) = h^6 \sum_{\rho(\tau)=6} C_{\theta,\mu}(\tau) F(\tau)(y_n) + \mathcal{O}(h^7),$$

where the weights $C_{\theta,\mu}(\tau)$ depend polynomically on θ and the free parameters. Now we define as a measure of the local error of this formula the quantity

$$g_\mu^* = \int_0^1 g_\mu(\theta) d\theta, \tag{3.12}$$

where

$$g_\mu(\theta) = \sqrt{\frac{\sum_{\rho(\tau)=6} \{C_{\theta,\mu}(\tau)\}^2}{\sum_{\rho(\tau)=6} \{C_{1,\mu}(\tau)\}^2}}. \tag{3.13}$$

Because at the end point of the interval we have for all μ the fifth order solution of DOPRI5(4), the coefficients $C_{1,\mu}(\tau)$ are independent of μ and the denominator in (3.13) is the constant 3.99×10^{-4} calculated by Dormand and Prince [2].

Next we consider how to choose the free parameters in order to get a continuous method with minimal local error in the sense of (3.12). This minimization process has been carried out numerically in the following way : First a coarse grid was established in the space of parameters and g_μ^* was computed on the points of this grid. As a consequence of this search some grid points close to the minimum were located. Taking

them as starting values, better values were found by a descent method. Finally we took simple rational values close to the computed optimal solution because they are more convenient from a computational point of view. In this way we have found the following set of optimal parameters

$$c_8 = \frac{1277}{6000}, \quad \sigma_1 = \frac{109}{450}, \quad \gamma = \frac{1}{3000}, \quad a_{86} = -\frac{23}{1250}, \quad c_9 = \frac{643}{1500}, \quad \sigma_2 = \frac{441}{500},$$

$$\alpha = \frac{157}{90000}, \quad a_{96} = \frac{1}{60}, \quad a_{98} = \frac{61}{300}, \quad b_9(\sigma_2) = \frac{8673}{50000}.$$

From (3.6) to (3.11) the coefficients of the additional stages and the new fifth order solutions, in decimal form, are

$a_{81} = 9.883325946430815073\ E-2$
$a_{82} = 6.820075031771575901\ E-2$
$a_{83} = 6.068825543094543486\ E-2$
$a_{84} = -4.711877279672667038\ E-2$
$a_{85} = 3.359191935282494552\ E-2$
$a_{86} = -1.840000000000000000\ E-2$
$a_{87} = 1.703792156426571357\ E-2$

$b_1(\sigma_1) = 6.980896127696492989\ E-2$
$b_2(\sigma_1) = 0$
$b_3(\sigma_1) = -1.11239613509505654\ E-1$
$b_4(\sigma_1) = 1.467355407207903519\ E-2$
$b_5(\sigma_1) = -1.050001669088839517\ E-2$
$b_6(\sigma_1) = 5.664429110099507573\ E-3$
$b_7(\sigma_1) = -4.623134348070028627\ E-3$
$b_8(\sigma_1) = 2.78438042311542827\ E-1$

$a_{91} = 3.224588250952790918\ E-2$
$a_{92} = 2.15458255334826252\ E-1$
$a_{93} = -5.655554913580628390\ E-2$
$a_{94} = 8.148046202216559079\ E-2$
$a_{95} = -4.129025308081867483\ E-2$
$a_{96} = 1.666666666666666667\ E-2$
$a_{97} = -2.267213098322812698\ E-2$
$a_{98} = 2.03333333333333333\ E-1$

$b_1(\sigma_2) = 8.452120256722231973\ E-2$
$b_2(\sigma_2) = 0$
$b_3(\sigma_2) = 1.88621758699678963\ E-1$
$b_4(\sigma_2) = 5.00874032429383933\ E-1$
$b_5(\sigma_2) = -2.45108174447082267\ E-1$
$b_6(\sigma_2) = 9.864265620489277901\ E-2$
$b_7(\sigma_2) = -5.168670915601016669\ E-2$
$b_8(\sigma_2) = 1.32675233701914438\ E-1$
$b_9(\sigma_2) = 1.73460000000000000\ E-1.$

In Fig.3.1 we have plotted the graphs of the functions $g_\mu(\theta)$ (denoted by error) as functions of θ for the DPS method of Shampine (in the figure, -*-) as well as our method given in [1] (in the figure, -o-) and the new method (in the figure, -•-). Clearly the behavior of $g_\mu(\theta)$ for the new method is better than for the others.

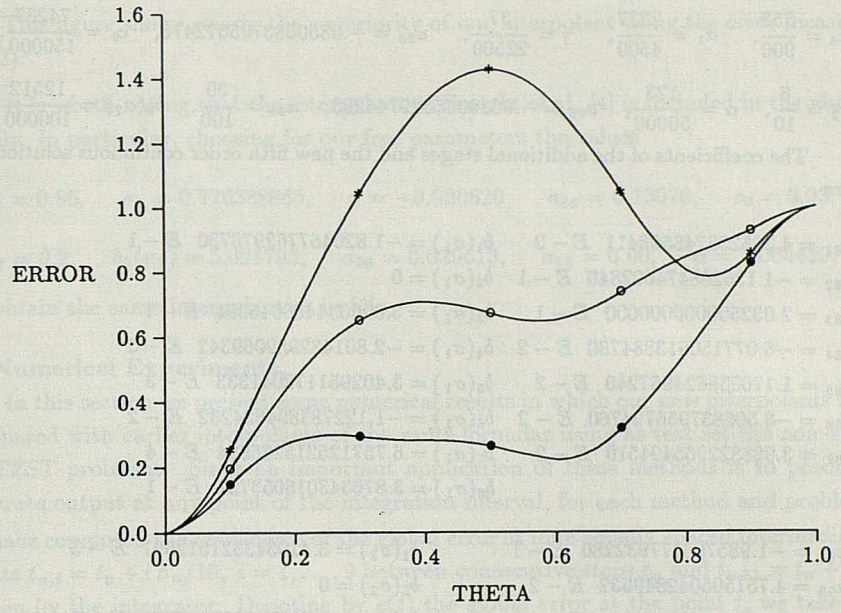


Fig 3.1

Remarks

1) It is worth noting that the interpolants of Shampine and Calvo et al. are included in the above family defined by (3.6)-(3.11). In particular, choosing for our free parameters the values

$$c_8 = \frac{2}{5}, \quad \sigma_1 = \frac{2}{5}, \quad \gamma = \frac{-1}{250}, \quad a_{86} = \frac{1}{20}, \quad c_9 = \frac{2}{5},$$

$$b_9(\sigma_2) = \frac{125}{36}\sigma_2^2(\sigma_2-1)^2(5\sigma_2-2), \quad a_{96} = \frac{13486}{402675}, \quad a_{98} = -\frac{1737}{3068}, \quad \alpha = \frac{1}{4}c_9^4 b_9(\sigma_2),$$

with arbitrary σ_2, we obtain the same interpolant as in [1]. (Note that with these parameters $\alpha_1 = \alpha_2 = a_{92} = 0$)

2) Although our continuous family has nine stages, we have determined the approximation $y_{n+\sigma_1}$ using only the first 8 stages of the RK method. It would be possible to consider a more general case computing this approximation with the full set of stages.

4. Interpolants for RKF4(5).

Due to the fact that the fifth order formula of Fehlberg pair satisfies the same simplifying assumptions as the fifth order formula of DOPRI5(4) and has also six stages, the determination of families of interpolants can be done in the same way. We merely outline this process. We have again a family of interpolants depending on the parameters $\mu = (\sigma_1, \sigma_2, c_8, c_9, \gamma, \alpha, a_{86}, a_{96}, a_{98}, b_9(\sigma_2))$ that can determined by the linear equations (3.6)–(3.11), where as the coefficients are the corresponding to Fehlberg formula. Next, following a minimization process as in the above section we get the optimal set of

$$c_8 = \frac{353}{900}, \quad \sigma_1 = \frac{2227}{4500}, \quad \gamma = \frac{37}{22500}, \quad a_{86} = -.0850683795672476, \quad c_9 = \frac{74237}{150000},$$

$$\sigma_2 = \frac{8}{10}, \quad \alpha = \frac{123}{50000}, \quad a_{96} = -.0610065529744293, \quad a_{98} = \frac{39}{100}, \quad b_9(\sigma_2) = \frac{12512}{100000}.$$

The coefficients of the additional stages and the new fifth order continuous solution are

$a_{81} = 4.385233748598411\ E-2$ $\quad b_1(\sigma_1) = -1.820467762970730\ E-1$

$a_{82} = -1.136288474622840\ E-1$ $\quad b_2(\sigma_1) = 0$

$a_{83} = 2.032500000000000\ E-1$ $\quad b_3(\sigma_1) = 3.006034468645994\ E-1$

$a_{84} = -6.077150613884780\ E-2$ $\quad b_4(\sigma_1) = -2.801482989069342\ E-3$

$a_{85} = 1.170038824967240\ E-2$ $\quad b_5(\sigma_1) = 3.402951172041333\ E-3$

$a_{86} = -8.506837956724760\ E-2$ $\quad b_6(\sigma_1) = -1.122783894634232\ E-2$

$a_{87} = 3.988822965494510\ E-2$ $\quad b_7(\sigma_1) = 6.757125215285924\ E-4$

$\quad b_8(\sigma_1) = 3.876343016053724\ E-1$

$a_{91} = -1.985706177932280\ E-1$ $\quad b_1(\sigma_2) = 3.419343321512391\ E-3$

$a_{92} = 4.751505042849532\ E-2$ $\quad b_2(\sigma_2) = 0$

$a_{93} = 3.000000000000000\ E-1$ $\quad b_3(\sigma_2) = 3.899120293887170\ E-1$

$a_{94} = 4.781403514569980\ E-2$ $\quad b_4(\sigma_2) = 2.594864526440950\ E-1$

$a_{95} = -3.132057379825690\ E-2$ $\quad b_5(\sigma_2) = -9.140970551625090\ E-2$

$a_{96} = -6.100655297442930\ E-2$ $\quad b_6(\sigma_2) = 1.030072673555981\ E-2$

$a_{97} = 4.819923250521310\ E-4$ $\quad b_7(\sigma_2) = -4.730936162733424\ E-2$

$a_{98} = 3.900000000000000\ E-1$ $\quad b_8(\sigma_2) = 1.504805150537014\ E-1$

$\quad b_9(\sigma_2) = 1.251200000000000\ E-1.$

Finally, in Fig.4.1 we have plotted the graphs of the functions $g_\mu(\theta)$ for the new interpolant (in the figure, -•-) and the one given by Enright et al [3] (in the figure,

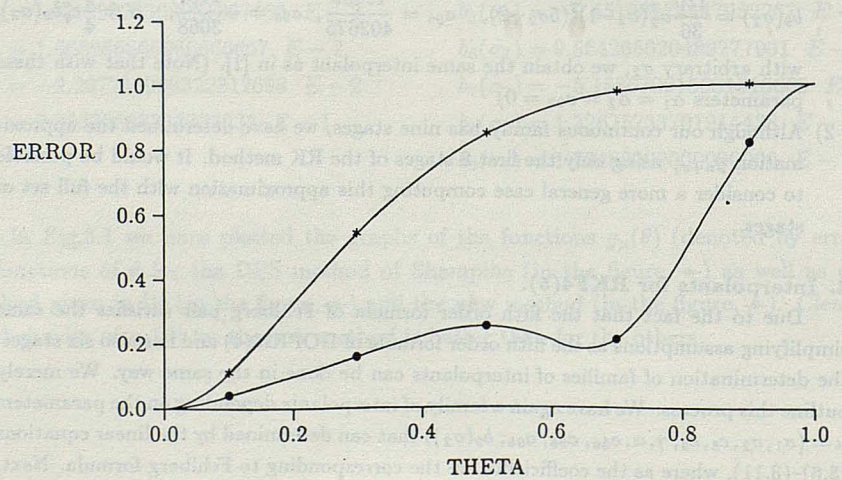


Fig 4.1

-+-). This figure shows clearly the superiority of our interpolant using the error measure (3.12)

It is worth noting that the interpolant of Enright el al. [4] is included in the above family. In particular, choosing for our free parameters the values

$$c_8 = 0.86, \quad \sigma_1 = 0.776388865, \quad \gamma = -0.030620, \quad a_{86} = 0.13076, \quad c_9 = 0.93,$$

$$\sigma_2 = 0.2, \quad b_9(\sigma_2) = 5.691793, \quad a_{96} = 0.029613, \quad a_{98} = 0.00, \quad \alpha = 1.064439,$$

we obtain the same interpolant as in [4].

5. Numerical Experiments.

In this section we present some numerical results in which our new interpolants are compared with earlier interpolants of the same formulas using as test set the non–stiff DETEST problems. Since an important application of these methods is to produce accurate output at any point of the integration interval, for each method and problem we have computed the max–norm of the global error at nine equally spaced intermediate points $t_{n,i} = t_n + i\,h_n/10$, $i = 1, \ldots, 9$ between consecutive steps t_n and $t_{n+1} = t_n + h_n$ chosen by the integrator. Denoting by $e(t)$ the global error at the point t, we take as measure of the error for a method and a given scalar problem the quantity

$$R = \max_{n \geq 0} \frac{\{\max\{e(t_{n,i}) | i = 1, \ldots, 10\}}{\max\{e(t_n), e(t_{n+1})\}\}}. \tag{5.1}$$

This means that we compute for each interval $[t_n, t_{n+1}]$ the ratio of the errors at eleven equally spaced intermediate points divided by the greatest of the errors at the two ends of the interval and then we take the maximum over all integration intervals.

The computations of the quantity (5.1) for the interpolants under consideration were carried out on a VAX–8300 computer in double precision with tolerances 10^{-i}, $i = 3, \ldots, 9$.

For brevity, we show here the results with some typical DETEST problems whose exact solution can be easily computed (A1,A2,A3,A4) and the simple non–linear problem

$$\begin{aligned} y' &= (y + \sqrt{t^2 + y^2})/t \quad t \in [1, 20] \\ y(1) &= 0 \end{aligned} \tag{5.2}$$

whose exact solution is $y(t) = (t^2 - 1)/2$.

In table (5.1) we present the results for the interplants of DOPRI5(4). For each problem the first, second and third row give the values of (5.1) corresponding to our method, the interpolant [1] and the DPS interpolant respectively.

In table (5.2) we show the results for two interpolants of Fehlberg's pair using the same set of test problems. For each problem, the first row gives the values of the factor (5.1) for our interpolant and the second row the corresponding values for Enright's interpolant.

TABLE 5.2

Problem	1D-3	1D-4	1.D-5	1.D-6	1.D-7	1.D-8	1.D-9
A1	1.261	1.015	1.000	1.000	1.000	1.000	1.000
A1	10.784	5.860	4.003	1.531	1.541	1.538	1.480
A2	1.000	1.000	1.000	1.225	1.000	1.149	1.000
A2	1.183	1.155	1.114	1.077	1.054	1.043	1.044
A3	1.245	1.658	1.190	1.211	1.187	1.150	1.323
A3	6.278	2.667	4.428	3.949	2.431	1.385	1.285
A4	1.588	1.161	1.123	1.094	1.088	1.098	1.108
A4	1.081	1.054	1.050	1.036	1.033	1.039	1.043
(5.2)	1.000	1.000	1.011	1.000	1.000	1.000	1.000
(5.2)	1.000	1.006	1.126	1.000	1.021	1.011	1.000

TABLE 5.1

Problem	1D-3	1D-4	1.D-5	1.D-6	1.D-7	1.D-8	1.D-9
A1	1.000	1.000	1.000	1.000	1.000	1.000	1.000
A1	1.123	1.125	1.077	1.099	1.013	1.000	1.000
A1	1.039	1.047	1.019	1.002	1.000	1.000	1.000
A2	1.000	1.000	1.000	1.000	1.000	1.000	1.000
A2	1.000	1.000	1.000	1.000	1.000	1.000	1.009
A2	1.000	1.000	1.000	1.000	1.000	1.000	1.000
A3	1.070	1.360	1.585	1.803	1.145	1.111	1.523
A3	1.438	1.631	1.507	1.822	1.633	3.916	1.795
A3	1.267	2.823	1.940	2.059	2.044	5.303	1.977
A4	1.096	1.291	1.202	1.167	1.137	1.111	1.146
A4	1.131	1.523	1.502	1.381	1.425	1.259	1.393
A4	1.132	1.664	1.467	1.229	1.428	1.460	2.587
(5.2)	1.000	1.000	1.515	1.177	1.075	1.000	1.000
(5.2)	1.000	1.000	2.262	1.449	1.441	1.723	1.000
(5.2)	1.108	1.380	6.232	6.319	11.62	3.590	1.220

From these tables it may be concluded that the errors of the interpolated values are of the same order as the error of the neighbouring grid points. Furthermore in our experiments we have seen that our new continuous extensions are in general more reliable than the others.

REFERENCES

[1] Calvo M., Montijano J.I., Rández L.; A fifth order interpolant for the Dormand and Prince Runge–Kutta method. J. Comput. Appl. Math., Vol. 29, (1990), pp. 91–100.

[2] Dormand J.R. and Prince P.J; A family of embedded Runge–Kutta formulae. J. Comput. Appl. Math 6, (1980).

[3] Enright W.H., Jackson K.R., Norsett S.P., Thomsen P.G.; Interpolants for Runge–Kutta formulas. ACM transactions on Mathematical Software, 12, N° 3, (1986), pp. 193–218

[4] Enright W.H., Jackson K.R., Norsett S.P.,Thomsen P.G.; Effective Solution of Discontinuous IVPs using a Runge–Kutta formula pair with interpolants.Applied Math. and Comp. 27 (1988), pp. 313–335.

[5] Fehlberg, E. ; Klassische Runge–Kutta–Formeln vierter und niedrigerer Ordnung mit Schrittweiten–Kontrolle und ihre Anwendung auf Warmeleitungs probleme. Computing 6, 1–2 (1970), pp. 61–71.

[6] Gladwell I., Shampine L. F., Baca L.S., Brankin R. W.; Practical aspects of interpolation in Runge–Kutta codes. SIAM J. Sci. Stat. Comput. 8, N° 3 (1987) pp. 322–341.

[7] Horn M.K.; Fourth and fifth–order, scaled Runge–Kutta algorithms for treating dense output. SIAM J. Numer. Anal. 20, (1983), pp. 558–568.

[8] Owren B. and Zennaro M. Order barriers for continuous explicit Runge–Kutta methods. Tech. Rep. University of Trondheim (1989).

[9] Shampine L. F.; Interpolation for Runge–Kutta methods. SIAM J. Numer. Anal. N° 22, 5, (1985), pp. 1014–1027.

[10] Shampine L. F.; Some Practical Runge–Kutta Formulas. Math. of Comp. N° 173 (1986), pp. 135–150.

Rev. Acad. Ciencias. Zaragoza 45, (1990)

THE ANALITICAL THEORY OF THE EARTH'S ROTATION USING A SYMMETRICAL GYROSTAT AS A MODEL.

R. Cid
Dpto. de Física Teórica. Universidad de Zaragoza. España.

A. Vigueras
Dpto. de Matemática Aplicada y Estadística. Universidad de Murcia. España.

Abstract.- In this work, the problem of the Earth's rotation when it is attracted by the Sun and the Moon is studied, using as a model a stationary symmetric girostat. We suppose that the two first components of the gyrostatic moment are null and we choose the third component as a constant in such a way that the free polar motion has a period of 430 days (Chandler's period).

The problem is formulated by means of dimensionless canonical variables referred to the mean ecliptic of an adopted epoch, assuming that the Sun moves in an elliptic orbit with zero inclination, and the Moon in an elliptic orbit where nodal and inclination arguments are constant. From these hypothesis and for the purpose of practical calculation, the periodic terms are eliminated by Deprit"s method and the secular perturbations to the second order are obtained.

1. INTRODUCTION.

The theory of rotation of the Earth about its center of mass constructed by Woolard (1953) and adopted by the IAU as the international standard, consider that, dynamically, the Earth is a symmetrical rigid body.

By using the Serret-Andoyer canonical variables λ, μ, ν, p_λ, p_μ, p_ν, (Andoyer, 1923), and a moving plane of reference (the mean acliptic of the epoch), Kinoshita (1977) developed a theory about the Earth's rotation. He adopted a triaxial rigid Earth model and the Hori's perturbation method. This theory has two fun-

damental advantages:

1) It treats separately the motions of the rotation axis and of the angular momentum axis. 2) It utilises the mean ecliptic of the epoch as plane of reference. In this way in the perturbation function doesn't appear mixed secular terms; but the corrections due to the not symmetry are little significant in the usual approximations and the same Kinoshita remove them in his development.

E. Cid (1982) deals with such a problem for a symmetric rigid body Earth, he utilised the canonical variables π, ζ, ν, p_π, p_ζ, p_ν, introduced by R. Cid and J.M. Correas (1973) also referred to the mean ecliptic of the epoch, and the Deprit's perturbation method (1969).

In both works, the free solution obtained when the external forces are cancelled gives us the Eulerian component of the polar motion. In the present paper, we use as a model of the Earth a symmetrical gyrostat wich has the two first components of the gyrostatic moment identically zero and the third one is constant, in such a way that in absence of external forces the free solution describes the Chandler's component of the polar motion.

The problem has been formulated in terms of a set of dimensionless variables π, ζ, ν, p_π, p_ζ, p_ν, referred to the mean ecliptic of the epoch, wich is determined by the planetary presence. In addition, we suppose that the Sun moves in a Keplerian orbit with null inclination and the Moon moves in a Keplerian orbit whose nodal and Inclination arguments are constant.

The Hamiltonian of the problem, stated in this way, has the form

$$H = H_0 + \varepsilon H_1 + \frac{1}{2}\varepsilon^2 H_2 + \ldots$$

and we can calculate the secular motions up to a second order of approximation after separating the periodical motions.

2. STATEMENT OF THE PROBLEM OF THE ROTATION OF THE EARTH WHEN IT IS CONSIDERED TO BE A GYROSTAT.

Let us suppose the Earth to be a gyrostat with constant gyrostatic momentum and wich turns around an axis passing through its center of mass O, with instantaneous angular velocity $\vec{\omega}$.

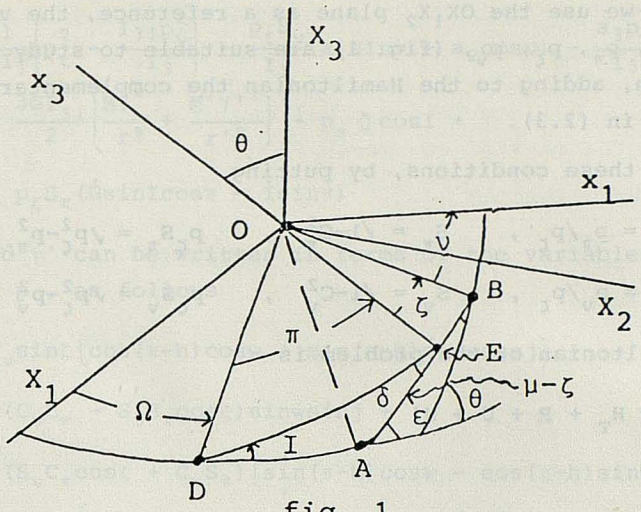


fig. 1

Therefore, we can consider a system of de principal axes of inertia $Ox_1x_2x_3$. which are rigidily attached to the rigid part of Earth, whose principal moments of inertia are denoted by I_1, I_2, I_3.

Let us consider a system $OX_1X_2X_3$, whose axis are palallel to those of other inertial one, and whose OX_1X_2 plane corresponds to the mean acliptic of the initial epoch $t_0 = 0$. The mean ecliptic for the epoch t, given by the $OX_1^*X_2^*$ plane of a new system $OX_1^*X_2^*X_3^*$, is referred to the $OX_1X_2X_3$ system through the functions $\Omega(t)$, and $I(t)$ (nodal and inclination angles), which are given by the following expressions (Newcomb, 1906)

$$\begin{aligned} \sin I \sin\Omega &= F_1 t + F_2 t^2 + F_3 t^3 + \ldots. \\ \sin I \cos\Omega &= G_i t + G_2 t^2 + G_3 t^3 + \ldots. \end{aligned} \qquad (2.1)$$

where the coefficients F_i, G_i, are constant.

In addition, we suppose that the Earth is attracted according to the Newton's law by two material points (the Moon and the Sun) of masses M, M', respectively, and that such masses describe eliptical orbits around O, in such a way that the Sun moves into the $OX_1^*X_2^*$ plane and the Moon in a plane, whose nodal and inclination angles with respect to the $OX_1^*X_2^*$ plane are respectively h and j.

If we use the OX_1X_2 plane as a reference, the variables π, ζ, ν, p_π, p_ζ, p_ν, (fig. 1) are suitable to study the Earth rotation, adding to the Hamiltonian the complementary function R given in (2.3).

In these conditions, by putting

$$C_\pi = p_\pi/p_\zeta \ , \qquad S_\pi = \sqrt{1-C_\pi^2} \ , \qquad p_\zeta S_\pi = \sqrt{p_\zeta^2 - p_\pi^2}$$
$$C_\nu = p_\nu/p_\zeta \ , \qquad S_\nu = \sqrt{1-C_\nu^2} \ , \qquad p_\zeta S_\nu = \sqrt{p_\zeta^2 - p_\nu^2} \qquad (2.2)$$

the Hamiltonian of the problem is

$$H = H_T + R + U + U'$$

where

$$H_T = \frac{1}{2}\left[p_\zeta^2 S_\nu^2 \left\{\frac{\sin^2\nu}{I_1} + \frac{\cos^2\nu}{I_2}\right\} + \frac{a_3 p_\nu^2}{I_3}\right] - \frac{a_3}{I_3} p_\nu -$$
$$- p_\nu S_\nu \left[\frac{a_1}{I_1}\sin\nu + \frac{a_2}{I_2}\cos\nu\right]$$

$$R = -p_\pi \frac{d\Omega}{dt}\cos I + p_\zeta S_\pi \left[\frac{d\Omega}{dt}\sin I\cos\pi - \frac{dI}{dt}\sin\pi\right] \qquad (2.3)$$

$$U = \frac{3GM}{2r^3}\left[(I_2-I_1)\beta^2 + (I_3-I_1)\gamma^2\right]$$

$$U' = \frac{3GM'}{2r'^3}\left[(I_2-I_1)\beta'^2 + (I_3-I_1)\gamma'^2\right]$$

being (β,γ), (β',γ') the second and third direction cosines of the position vectors Earth-Moon and Earth-Sun, respectively, with respect to the system of the principal axes of inertia $Ox_1x_2x_3$. In this statement the terms of the lunisolar potential of power superior to $1/r^3$ and $1/r'^3$, are omitted, by assuming that these terms are sufficiently small.

The problem is simplified by using a model of the Earth with dynamical symmetry ($I_1 = I_2$). In this case, by putting

$$I_{31} = I_3 - I_1 \ , \qquad I_o = (I_3-I_1)/I_1$$

the Hamiltonian becomes

$$H = \frac{1}{2I_1}\left[p_\zeta^2 - \frac{I_{31}p_\nu^2}{I_3}\right] - \frac{p_\zeta S_\nu}{I_1}(a_1\sin\nu + a_2\cos\nu) - \frac{a_3 p_\nu}{I_3} +$$

$$+ \frac{3GI_{31}}{2}\left[\frac{M\gamma^2}{r^3} + \frac{M'\gamma'^2}{r'^3}\right] - p_\pi \dot{\Omega}\cos I + \qquad (2.4)$$

$$+ p_\zeta S_\pi(\dot{\Omega}\sin I\cos\pi - \dot{I}\sin\pi)$$

where γ and γ' can be written in terms of the variables π, ζ, ν, p_π, p_ζ, p_ν, as follows

$$\gamma = S_\nu\sin\zeta\left[\cos(\pi-h)\cos w + \sin(\pi-h)\sin w\cos j\right] +$$

$$+ (C_\nu C_\pi - S_\nu S_\pi\cos\zeta)\sin w\sin j + \qquad (2.5)$$

$$+ (S_\nu C_\pi\cos\zeta + C_\nu S_\pi)\left[\sin(\pi-h)\cos w - \cos(\pi-h)\sin w\sin j\right]$$

$$\gamma = -(C_\nu S_\pi + S_\nu S_\pi\cos\zeta)\sin(w'-\pi) + S_\nu\sin\zeta\cos(w'-\pi)$$

The variables $w = w_o+f$, $w' = w'_o+f'$, define the Moon and Sun positions by means of the true anomalies f, f', and the perigee's arguments w_o, w'_o.

Let us carry out the transformation ($p'_i = I_1\omega_o p_i$, $t' = \omega_o t$) to other dimensionless variables, being $\omega_o \neq 0$, the initial angular velocity of the Earth, then we obtain for the Hamiltonian the expression

$$H = \frac{1}{2}\left[p_\zeta^2 - \frac{I_{31}p_\nu^2}{I_3}\right] - \frac{p_\zeta S_\nu}{I_1\omega_o}(a_1\sin\nu + a_2\cos\nu) - \frac{a_3 p_\nu}{I_3\omega_o} -$$

$$- \frac{p_\pi}{\omega_o}\dot{\Omega}\cos I + \frac{p_\zeta S_\pi}{\omega_o}(\dot{\Omega}\sin I\cos\pi - \dot{I}\sin\pi) + \qquad (2.6)$$

$$+ \frac{3GI_{31}}{2I_1\omega_o^2}\left[\frac{M\gamma^2}{r^3} + \frac{M'\gamma'^2}{r'^3}\right]$$

In the formula (2.6) we have maintained the notation p_i for de moment p'_i.

Now, consider the equalities

$$\frac{3GMI_o}{2r^3\omega_o^2} = \frac{3}{2}\left[\frac{n}{\omega_o}\right]^2\frac{MI_o}{M+M_o}\left[\frac{a}{r}\right]^3 = \frac{1}{2}\,e_o\left[\frac{a}{r}\right]^3$$

$$\frac{3GM'I_o}{2r'^3\omega_o^2} = \frac{3}{2}\left[\frac{n'}{\omega_o}\right]^2\frac{M'I_o}{M'+M_o}\left[\frac{a'}{r'}\right]^3 = \frac{1}{2}\,e'_o\left[\frac{a'}{r'}\right]^3$$

with

$$e_o = 3\left(\frac{n}{\omega_o}\right)^2 \frac{MI_o}{M+M_o} \quad , \quad e'_o = 3\left(\frac{n'}{\omega_o}\right)^2 \frac{M'I_o}{M'+M_o} \tag{2.7}$$

and where n, n', are the mean motions of the Moon and the Sun, and M_o is the mass of the Earth.

By the relations

$$\begin{aligned} &M = M_o/81.3 \quad , \qquad M' = 332.958\ M_o \\ &n = 2\pi/27.396 \text{ rad/sid.day} \ , \\ &n' = 2\pi/366.25 \text{ rad/sid. day} \end{aligned} \tag{2.8}$$

we obtain

$$\begin{aligned} &\omega_o = 2\pi \text{ rad/sid. day} \qquad I_o = 1/305 \\ &e_o = 1.592382\cdot 10^{-7} \qquad e'_o = 0.733272\cdot 10^{-7} \\ &U + U' = \frac{1}{2} e'_o \left\{ \delta \left(\frac{a}{r}\right)^3 \gamma^2 + \left(\frac{a'}{r'}\right)^3 \gamma'^2 \right\} \end{aligned} \tag{2.9}$$

where we define δ in the form

$$\delta = 1.592382/0.733272 = 2.171612$$

Now, make us an estimation of the terms of the complementary function R, by virtue of the relations

$$\begin{aligned} &(\dot{\Omega}/\omega_o)\cos I = -0.23602\cdot 10^{-7} - 0.58027\cdot 10^{-9} t \\ &(\dot{\Omega}/\omega_o)\sin I = -0.53943\cdot 10^{-11} t \\ &\dot{I}/\omega_o = 0.99319\cdot 10^{-9} - 0.22646\cdot 10^{-10} t \end{aligned} \tag{2.10}$$

where t is expressed in tropic centuries measured from 1850. The terms of the complementary function R can be written in the form

$$\begin{aligned} R = p_\pi &(0.23602\cdot 10^{-7} + 0.58027\cdot 10^{-9} t) - \left[0.53943\cdot 10^{-11} t \cos\pi \right. \\ &\left. + (0.99319\cdot 10^{-9} - 0.22646\cdot 10^{-10} t)\sin\pi \right] p_\zeta S_\pi \end{aligned}$$

and for the considered approximation it is sufficient to take

$$R = e'_o \alpha p_\pi \quad , \quad \alpha = \frac{0.23602}{0.733272} = 0.321872 \tag{2.11}$$

If we suppose that the two first components of the gyrostatic moment are null, $a_1 = a_2 = 0$, the third constant component

can be chosen in such a way that the free polar motion has a period of 430 days (Chandler's period). So, the equations of motion for a symmetric gyrostat, when the external forces are null and are given by

$$\omega_1 + I_o\omega_2\omega_3 + \omega_2(a_3/I_1) = \omega_1 + \rho\omega_2 = 0$$
$$\omega_2 - I_o\omega_1\omega_3 - \omega_1(a_3/I_1) = \omega_2 - \rho\omega_1 = 0 \qquad (2.12)$$

being

$$\omega_3 = \omega_o \quad (= 2\pi \text{ rad./sid. day})$$

$$\rho = I_o\omega_o + (a_3/I_1)$$

The solution of the above system is

$$\omega_1 = A\cos(\rho t+\beta) \quad , \quad \omega_2 = A\sin(\rho t+\beta)$$

where A and β are constant. This solution corresponds to a circular motion of the Pole with a period $T = 2\pi/\rho$, and making

$$T = 2\pi/\rho = 430$$

we obtain

$$b_3 = \frac{a_3}{I_1\omega_o} = \frac{1}{430} - I_o = -0.953107\cdot10^{-3} \qquad (2.13)$$

Then, taking the above hypothesis into account and introducing the constant $b_o = I_{31}/I_3$, the Hamiltonian (2.6) can be written in the form

$$H = \frac{1}{2}(p_\zeta^2 - b_o p_\nu^2) - b_3 p_\nu + e'_o\alpha p_\pi +$$
$$+ \frac{1}{2}e'_o\left[\delta(a/r)^3\gamma^2 + (a'/r')^3\gamma'^2\right] \qquad (2.14)$$

The functions γ^2 and γ'^2 can be expressed by means of the equalities

$$\gamma^2 = \sum_{i,s,q} \gamma_{i,s,q}\cos\{i\zeta + s(\pi-h) + qw\}$$
$$\gamma'^2 = \sum_{i,q} C_{i,q}\cos\{i\zeta + q(w'-\pi)\} \qquad (2.15)$$

where $i \in \{0,1,2\}$, $s \in \{-2,-1,0,1,2\}$ and $q \in \{-2,0,2\}$.

The coefficients $\gamma_{i,s,q}$, $C_{i,q}$ are functions of the moments p_π, p_ζ, p_ν, being

$$\gamma_{0,0,0} = (\text{sen}^2 j)G_1 + (1+\cos^2 j)G_2$$

$$\gamma_{0,0,2} = \gamma_{0,0,-2} = \frac{1}{2}(G_2 - G_1)\text{sen}^2 j$$

$$C_{0,0} = \frac{1}{4}(S_\nu^2 + 2C_\nu^2 S_\pi^2 + C_\pi^2 S_\nu^2) \qquad (2.15')$$

$$G_1 = \frac{1}{2}(C_\nu^2 C_\pi^2 + -S_\nu^2 S_\pi^2)$$

$$G_2 = \frac{1}{4}\{C_\nu^2 S_\pi^2 + -S_\nu^2(1+C_\pi^2)\}$$

Now, if we define the small parameter ε in the form

$$e_o' = \varepsilon^2 \quad , \quad \varepsilon = 2.7153121\cdot 10^{-4} \qquad (2.16)$$

we have

$$b_3 = - c_o\varepsilon \qquad c_o = 3.5101196 \qquad (2.17)$$

Therefore, the Hamiltonian (2.4) can be written as follows

$$H = H_0 + \varepsilon H_1 + \frac{1}{2}\varepsilon^2 H_2 \qquad (2.18)$$

being

$$H_0 = \frac{1}{2}(p_\zeta^2 - b_o p_\nu^2) = H_0(p_\zeta, p_\nu)$$

$$H_1 = c_o p_\nu = H_1(p_\nu) \qquad (2.19)$$

$$H_2 = 2\alpha p_\pi + \delta\left(\frac{a}{r}\right)^3\gamma^2 + \left(\frac{a'}{r'}\right)^3\gamma'^2 = H(\pi,\zeta,p_\pi,p_\zeta,p_\nu)$$

3. HAMILTONIAN OF THE SECULAR MOTION.

Let us start this paragraph by eliminating the variable of the Hamiltonian by the Deprit's method. Then the new Hamiltonian function is

$$H' = H_0' + \varepsilon H_1' + \frac{1}{2}\varepsilon^2 H_2' \qquad (3.1)$$

and the generating function

$$W = W_1 + \varepsilon W_2 \qquad (3.1')$$

where

$$H_0' = H_0 \quad , \quad H_1' = H_1 \quad , \quad W_1 = 0$$

$$H_2' = 2\alpha p_\pi + \delta\left(\frac{a}{r}\right)^3 \sum_{s,q} \gamma_{0,s,q} \cos\{s(\pi-h) + qw\} + \left(\frac{a'}{r'}\right)^3 \sum_{q} C_{0,q} \cos q(w'-\pi)$$

$$W_2 = \sum_{-\infty}^{\infty} \sum_{\substack{i,q \\ i\neq 0}} \left\{ C^k_{i,q} X^q_k \operatorname{sen}\{i\zeta + q(w_o'-\pi) + \delta_q k l'\} + \delta \sum \gamma^k_{i,s,q} X^q_k \operatorname{sen}\{i\zeta + s(\pi-h) + qw_o + \delta_q k l\} \right\} \tag{3.2}$$

being l and l' the mean anomalies, and where the coefficients $C^k_{i,q}$, $\gamma^k_{i,s,q}$, are given in (Vigueras, A., 1983).

In the precedent formulary we have suppressed the symbol (') in the new variables.

In these conditions, the Hamiltonian takes the form

$$H' = H_0'(p_\zeta', p_\nu') + \varepsilon H_1'(p_\nu') + \frac{1}{2}\varepsilon^2 H_2'(\pi', p_\pi', p_\zeta', p_\nu', t)$$

wich still depends of the angular variable π'.

A new application of the Deprit's method for eliminating the variable π, by means of a generating function

$$W' = W_1' + W_2' \tag{3.3}$$

allow us to obtain the Hamiltonian

$$H'' = H_0'' + \varepsilon H_1'' + \frac{1}{2}\varepsilon^2 H_2'' \tag{3.4}$$

where

$$H_0'' = H_0' \quad , \quad H_1'' = H_1' \quad , \quad W_1' = 0$$

$$H_2'' = 2\alpha p_\pi + \delta\left(\frac{a}{r}\right)^3 \sum_q \gamma_{0,0,q} \cos qw + \left(\frac{a'}{r'}\right)^3 C_{0,0} \tag{3.5}$$

$$W' = \frac{C_{0,2} S(\pi, f')}{n'(1-e'^2)^{3/2}} + \frac{\delta}{n(1-e^2)^{3/2}} \left(\sum_{\substack{s\neq 0 \\ q\neq 0}} \gamma_{0,s,q} E_{s,q} + \left\{ \sum_{s\neq 0} \gamma_{0,s,0} \operatorname{coss}(\pi-h) \right\}(f + \operatorname{esen} f) \right)$$

where f, f', denote the true anomalies of the Moon and the Sun,

and $S(\pi,f')$, $E_{s,q}$ the functions

$$S(\pi,f') = \sin 2(f'+w_0'-\pi) + \frac{e'}{3}\sin\{3f'+2(w_0'-\pi)\} + \\ + e'\sin\{f'+2(w_0'-\pi)\} \tag{3.6}$$

$$E_{s,q} = \frac{1}{q}\sin\{s(\pi-h)+qw_0+qf\} + \frac{e}{2(q+1)}\sin\{s(\pi-h)+qw_0 + \\ +(q+1)f\} + \frac{e}{2(q-1)}\sin\{s(\pi-h)+qw_0+(q-1)f\}$$

Finally, after this second elimination, we can write the Hamiltonian of secular motion in the form

$$H'' = H_0''(p_\zeta'',p_\nu'') + \varepsilon H_1''(p_\nu'') + \frac{1}{2}\varepsilon^2 H_2''(p_\pi'',p_\zeta'',p_\nu'',t) \tag{3.7}$$

with

$$H_0'' = -(p_\zeta''^2 - b_o p_\nu'')$$

$$H_1'' = c_o p_\nu'' \tag{3.8}$$

$$H_2'' = 2\alpha p_\pi'' + \delta\left(\frac{a}{r}\right)^3(\gamma_{0,0,0} + 2\gamma_{0,0,2}\cos 2w) + \left(\frac{a'}{r'}\right)^3 C_{0,0}$$

since $\gamma_{0,0,2} = \gamma_{0,0,-2}$. Then, by integrating the equations of motion whose Hamiltonian function is (3.7), we obtain the secular perturbations

$$\begin{aligned} p_\pi'' &= p_\pi^\circ \quad , & \pi'' &= \varepsilon^2(\alpha t + Q_1^\circ) + \pi_o'' \\ p_\zeta'' &= p_\zeta^\circ \quad , & \zeta'' &= p_\zeta^\circ t + \varepsilon^2 Q_2^\circ + \zeta_o'' \\ p_\nu'' &= p_\nu^\circ \quad , & \nu'' &= -(b_o p_\nu^\circ + c_o\varepsilon)t + \varepsilon^2 Q_3^\circ + \nu_o'' \end{aligned} \tag{3.9}$$

where we make use of the notation

$$Q_i^\circ = P_i^\circ F' + \delta\left[(1-\frac{3}{2}\sin^2 j)F + \frac{3}{2}E_{0,2}\sin^2 j\right]P_i^\circ \quad (i = 1,2,3)$$

$$F = f + e\sin f \qquad F' = f' + e'\sin f'$$

$$\mu_o = 1/\{4n_o(1-e^2)^{3/2}(p_\zeta^\circ)^5\} \qquad \mu_o' = 1/\{4n_o'(1-e'^2)^{3/2}(p_\zeta^\circ)^5\}$$

$$R_\pi = (p_\zeta^\circ)^2 - 3(p_\pi^\circ)^2 \qquad R_\nu = (p_\zeta^\circ)^2 - 3(p_\nu^\circ)^2$$

$$P_1^\circ = \mu_o p_\zeta^\circ p_\pi^\circ R_\nu \qquad P_2 = -\mu_o\left[(p_-)^2 R_\nu + (p_\nu)^2 R_\pi\right] \qquad P_3^\circ = \mu_o p_\zeta^\circ p_\nu^\circ R_\pi$$

The formulas for the coefficients $P_i^{\circ\prime}$ are obtained from the above P_i° on replacing μ_o by μ_o'.

In the preceding solution, we observe that the perturbations due to the Sun, the Moon and those due to the motion of the reference plane (the mean ecliptic of the epoch) appear separate.

REFERENCES

Andoyer, H. (1923): Cours de Mécanique Celeste. Gauthier-Villars. Paris.

Camarena, V., Ribera, J. and Pétriz, F. (1976): Urania, 184-200. Tarragona.

Cid, E. (1982): Sobre el movimiento de sólidos en torno a sus centros de masas con aplicación al estudio de la precesión y nutación terrestres. Tesis. Universidad de Zaragoza.

Cid, R. and Correas, J.M. (1973): Act. I Jorn. Matem. Hispano Lusas, 439-452.

Cid, R. and Vigueras, A. (1985): Cel. Mech. 36, 155. USA.

Deprit, A. (1972): Cel. Mech. 6(2), 127-150. USA.

Kinoshita, H. (1977): Cel. Mech. 15(3), 277-326. USA.

Leimanis, E. (1965): The General Problem of the Motion of Coupled Rigid Bodies about a Fixed Point. Springer-Verlag, Berlin.

San Saturio, M.E. and Vigueras, A. (1988): Cel. Mech. 41, 297.

Tsopa, M.P. (1981): P.M.M. 44, 285-287.

Vigueras, A. (1983): Movimiento rotatorio de giróstatos y aplicaciones. Tesis. Universidad de Zaragoza.

Woolard, E.W. and Clemence, G.M. (1966): Spherical Astronomy. Academic Press. New York.

Rev. Acad. Ciencias. Zaragoza 45, (1990)

CORRECCION DE ORBITAS DE ESTRELLAS DOBLES VISUALES

Carlos Osácar Soriano
Rafael Cid Palacios
Departamento de Física Teórica (Astronomía)
Universidad de Zaragoza, 50009 Zaragoza

Resumen

Se comparan los resultados obtenidos en el ajuste de órbitas de estrellas dobles visuales usando el Criterio de Mínimos Cuadrados y el Criterio de Mínimos Valores Absolutos. Dichos criterios se aplican a observaciones simuladas y a estrellas reales. En las observaciones simuladas se compara la fiabilidad de ambos criterios para recuperar la órbita inicial.

1 Introducción

Si las medidas efectuadas sobre una estrella doble careciesen de errores, no existiría problema alguno para calcular la órbita relativa de la misma a partir de las observaciones, pues bastaría coger un número suficiente de datos de observación para determinar los siete elementos que definen una órbita elíptica.

Los diferentes métodos de cálculo de órbitas proporcionan distintos resultados, según los lugares normales ó puntos escogidos. Por esto, se hace necesario determinar una órbita que represente, lo mejor posible, el conjunto de las observaciones.

Para esto se utilizan los métodos de corrección de órbitas. La cuestión de definir cuando una órbita es mejor que otra da lugar a que, según el criterio y los observables empleados, se obtengan distintos métodos para la corrección o mejora de órbitas.

Desde un punto de vista matemático, el problema de ajuste de funciones se puede expresar así:

"Consideremos una función diferenciable $f = f(t, \boldsymbol{a})$, que depende del tiempo t y de un conjunto de parámetros a_j $(j = 1, 2, \ldots, m)$, y supongamos que de dicha función se han hecho un cierto número n de medidas f_i^o $(i = 1, 2, \ldots, n > m)$ en tiempos t_i. El

problema de ajuste de funciones consiste en encontrar los valores de $\boldsymbol{a}$ que hacen mínima la suma de distancias entre las cantidades medidas y las calculadas $f_i^c = f(t_i, \boldsymbol{a})$". La definición de distancia dará lugar a los distintos criterios de ajuste.

Tradicionalmente, se ha utilizado el Criterio de Mínimos Cuadrados de forma sistemática para la reducción de observaciones. Sin embargo, no parece existir razón alguna *a priori* para que sea preferido a otros Criterios como puede ser el de Mínimos Valores Absolutos. Además, algunas experiencias hechas en otros campos de la Astronomía parecen indicar que el Criterio de Mínimos Cuadrados no es siempre la mejor elección para recobrar los valores de parámetros de interés astronómico a partir de las medidas realizadas (Branham, 1982).

2 Criterio de Mínimos Cuadrados

Este criterio considera como distancia el cuadrado de la diferencia de las ordenadas $f_i^o - f_i^c$ y por tanto el proceso consiste en minimizar el sumatorio

$$S = \sum_{i=1}^{n} (f_i^o - f_i^c)^2 = \sum_{i=1}^{n} [f_i^o - f(t_i, \boldsymbol{a})]^2. \qquad (1)$$

Su aplicación supone una distribución normal de los errores de observación. Los valores obtenidos con él tienen la varianza mínima para las diferencias observación - cálculo. Además presenta la ventaja de reducir las ecuaciones a un número manejable (igual al número de parámetros) de fácil tratamiento. Otra circunstancia que ha favorecido su utilización ha sido la existencia de un gran número de tests estadísticos para la media.

El problema se reduce a encontrar el mínimo de (1) en función de las variables a_j. Para resolverlo se puede utilizar cualquier algoritmo de minimización de funciones.

Dichos algoritmos pueden agruparse en tres categorías:

- Métodos que sólo utilizan los valores de la función a minimizar, como el algoritmo del simplex;
- métodos en los que además se emplean las derivadas parciales primeras, como en el algoritmo del gradiente;
- métodos en los que intervienen las derivadas parciales segundas, como en el algoritmo de Newton.

El método de Newton converge rápidamente, pero falla en ocasiones si la estimación inicial está muy alejada de la solución; los algoritmos como el del gradiente tienen un radio de convergencia mayor, aunque convergen más lentamente. Para evitar estos inconvenientes han surgido otros algoritmos como el de Levenberg-Marquardt (Marquardt, 1963) que combina los algoritmos del gradiente y de Newton; o el de Fletcher-Powell (Fletcher y Powell, 1963) que obtiene la dirección del gradiente y luego estima cuánto se debe mover en esa dirección para alcanzar el mínimo, usando las derivadas segundas.

En los métodos anteriores se supone que los parámetros $\boldsymbol{a}$ son independientes, pero no existe ninguna dificultad en considerar relaciones entre los distintos componentes de $\boldsymbol{a}$, que se introducen mediante sus correspondientes multiplicadores de Lagrange.

Es posible también usar métodos en los que intervengan las derivadas segundas, como el descrito por Eichorn y Clary (1974).

Cuando las relaciones que ligan los observables con los parámetros están dadas de forma implícita, el criterio de Mínimos Cuadrados se puede aplicar en la forma dada por Jefferys (1980,1981).

3 Criterio de Mínimos Valores Absolutos.

Las observaciones de estrellas dobles no siempre cumplen la hipótesis de distribución normal de errores subyacente bajo el Criterio de Mínimos Cuadrados, lo que puede invalidar los resultados del ajuste. Esto obliga a realizar un filtrado previo de los datos rechazando valores con un valor de corte arbitrario. Sería deseable un método de ajuste que incluyera todas las observaciones sin estar demasiado influenciado por observaciones claramente discordantes para poder determinar este valor de corte.

De los resultados obtenidos por Branham (1982) se deduce que otro criterio utilizable es considerar mejor aquella órbita que haga mínima la suma de los valores absolutos de las diferencias observación – cálculo. Este criterio es mucho menos usado. Laplace lo propuso en su "Mécanique Céleste" añadiéndole además la condición de que la suma de las desviaciones fuera nula.

Desde un punto de vista estadístico, este criterio considera que los errores de las medidas siguen una distribución de la forma

$$y = \frac{h}{2}\exp(-h|z|)$$

siendo h el módulo de precisión y z el error cometido. Esto equivale a decir que se toma como valor más probable de una medida la mediana de las medidas u observaciones. Puede demostrarse que la solución obtenida mediante este criterio satisface exactamente tantas ecuaciones de condición como parámetros de ajuste, usándose el resto de las medidas para determinar qué conjunto de ecuaciones ha de satisfacerse. Por otra parte, el cambio de algunas observaciones que no sean satisfechas no cambia los valores estimados de los parámetros. Esta propiedad es la que hace que los resultados obtenidos sean mucho menos sensibles a la aparición de medidas con errores grandes, o con una distribución de errores que no sea gaussiana. La insensibilidad al cambio en una medida es reflejo de las propiedades del estimador utilizado, esto es, la mediana.

El Criterio de Mínimos Valores Absolutos presenta el inconveniente práctico de que es necesario trabajar con todas las ecuaciones de condición, no pudiéndo reducirse a un número pequeño de ecuaciones. Por todo esto, ha sido poco utilizado hasta la aparición de ordenadores rápidos con los que esta última objeción pierde importancia, siendo primordial la calidad de la órbita obtenida. Además los algoritmos usados son más complejos que el de Mínimos Cuadrados. Asimismo, no existen prácticamente tests estadísticos para la mediana, lo que hace díficil el control de los resultados.

Este criterio considera como distancia el valor absoluto de la diferencia de las ordenadas $f_i^o - f_i^c$ y por tanto el proceso consiste en minimizar el sumatorio

$$\mathrm{L1} = \sum_{i=1}^{n} |f_i^o - f_i^c|. \qquad (2)$$

A diferencia del caso de mínimos cuadrados, en (2) no se pueden igualar a 0 las derivadas parciales de *L1*, ya que la función Valor Absoluto no tiene derivada en el punto 0. Por tanto, de las tres clases de algoritmos de minimización descritos anteriormente, sólo podemos usar los del primer tipo, que no recurren a los valores de las derivadas, sino que usan únicamente los valores de la función, aunque resulten más lentos. Si f es una función lineal, se han desarrollado algoritmos específicos para este problema basados en los métodos de programación lineal como el descrito en Barrodale and Roberts (1973). Cuando f es una función no lineal y conocemos unos valores iniciales de $\boldsymbol{a}$, podemos mejorarlos haciendo un desarrollo de la función f en serie de Taylor en torno de dichos valores previos, obteniendo

$$\mathrm{L1} = \sum_{i=1}^{n} |f_i^o - f(t_i, \boldsymbol{a}) - \sum_{j=1}^{m} \frac{\partial f(t_i, \boldsymbol{a})}{\partial a_j} \Delta a_j| = 0,$$

con lo que llegamos a un sistema lineal.

En el caso de estrellas dobles, la fórmula (2) se transforma en la siguiente

$$\mathrm{L1} = \sum_i |\boldsymbol{\theta}_i^o - \boldsymbol{\theta}_i^c(t_i, \boldsymbol{a})|, \quad i = 1, \ldots, n. \tag{3}$$

Suponiendo conocida una órbita previa podemos calcular las correcciones mediante un desarrollo en serie del tipo

$$\mathrm{L1} = \sum_i |\theta_i^o - \theta_i^c(t_i, \boldsymbol{a}_0) - \sum_{j=1}^{m} \frac{\partial \theta(t_i, \boldsymbol{a}_0)}{\partial a_j} \Delta a_j|, \tag{4}$$

con lo que llegamos a un sistema lineal, cuya aplicación reiterada nos conduce a la solución de (3). En el caso general, cuando las relaciones que ligan las variables medidas con los parámetros vengan dadas por ecuaciones implícitas, se aplican métodos como el descrito por Späth y Watson (1987).

4 Experimentos numéricos

Con el fin de comparar los diversos métodos de corrección de órbitas se ha considerado la minimización, por el criterio de Mínimos Cuadrados, de las funciones siguientes:

$$\mathrm{S1} = \sum_i p_i \left(\theta_i^o - \theta_i^c(t_i, \boldsymbol{a})\right)^2 + q_i \left(\rho_i^o - \rho_i^c(t_i, \boldsymbol{a})\right)^2 \tag{5}$$

$$\mathrm{S2} = \sum_i \left[\rho_i^{o2} + \rho_i^{c2}(t_i, \boldsymbol{a}) - 2\rho_i^o \rho_i^c(t_i, \boldsymbol{a}) \cos(\theta_i^o - \theta_i^c(t_i, \boldsymbol{a}))\right] \tag{6}$$

Con *S1* se minimizan conjuntamente los ángulos y las distancias. Las cantidades p_i y q_i son los pesos relativos de cada medida y sirven también para homogeneizar ambas sumas de cuadrados. Se tomó $p_i = 1$ y $q_i = 1$.

S2 es la suma de cuadrados de las distancias entre los puntos calculados y observados.

Para describir la órbita se usaron los elementos orbitales $a, e, P, T, \omega, \Omega, i$. Con la función *S2* se hicieron dos ajustes, uno usando estos elementos y otro usando las variables de Thiele-Innes A, B, F, G, e, P y T, con el fin de compararlos con el método de ajuste del mismo nombre.

El criterio de Mínimos Valores Absolutos se aplicó a la función

$$\mathrm{L1} = \sum_i P_i \, |\theta_i^o - \theta_i^c(t_i, \boldsymbol{a})| + Q_i \, |\rho_i^o - \rho_i^c(t_i, \boldsymbol{a})| \, . \tag{7}$$

Esta función es análoga a *S1*, pero usando valores absolutos en vez de la suma de cuadrados.

Supondremos que todas las medidas tienen el mismo peso estadístico. La introducción de pesos en *S1* y *L1*, para cada observación, no presenta ningún problema, puesto que basta cambiar los valores de p_i y q_i por su peso correspondiente, sin tener que modificar los algoritmos utilizados.

En lugar de la excentricidad e se ha usado una variable auxiliar s definida como $e = s^2/(1+s^2)$ (Soulié, 1986) que asegura que la excentricidad e se mantendrá en el intervalo $[0,1)$ para cualquier valor de s. La sustitución clásica $e = \operatorname{sen}\phi$ puede proporcionar valores negativos de la excentricidad si $\phi < 0$.

El criterio de mínimos cuadrados se ha implementado usando el método de Levenberg-Marquardt con la rutina ZXSSQ del paquete IMSL (IMSL, 1985) que halla el mínimo de una suma de cuadrados. Las derivadas de la función, usadas en el método, se calculan numéricamente como diferencia entre dos valores próximos de las variables dentro de la propia rutina.

La corrección usando el criterio de Mínimos Valores Absolutos se realiza usando la fórmula (3). Esta función se minimiza usando métodos de programación lineal mediante el algoritmo de Robers y Robers (1971) (CACM, algoritmo 458) obteniendo las correcciones de los parámetros orbitales. Con la nueva órbita se vuelve a minimizar *L1* para obtener otras nuevas correcciones; este proceso se repite hasta que los incrementos obtenidos sean tan pequeños como se quiera. En casi todos los casos ensayados la convergencia de este proceso ha sido muy rápida. En los casos en que las correcciones oscilaban fue suficiente con tomar una fracción (entre 0.9 y 0.7) de dichas correcciones. Con esto se disminuye la velocidad de convergencia pero se llega a una solución.

4.1 Casos simulados

Con estos métodos de ajuste estamos interesados en obtener la órbita verdadera. Para poder comprobar la concordancia de los valores recobrados con la órbita original, se aplicaron los métodos descritos anteriormente a un conjunto de casos simulados en los que se conoce la órbita verdadera.

Para ello se eligieron unos valores de los elementos orbitales y se calcularon las efemérides correspondientes a una serie de épocas. A cada valor obtenido para el ángulo y la distancia se le sumó un número aleatorio con una distribución gausiana de media 0 y varianza conocida con objeto de simular errores en las medidas. Este conjunto se identificó con un conjunto de "medidas". Como valores iniciales para los elementos orbitales se tomaron unos valores más ó menos parecidos a los originales, aunque en algunos casos la desviación típica de las diferencias observación - cálculo era completamente inaceptable (30°).

En todos los casos se usaron 139 puntos que cubrían aproximadamente período y medio de la órbita elegida y correspondían a las épocas de medida de una estrella real (ADS 11077) para que la distribución en el tiempo de éstas fuera realista. Los resultados correspondientes a cuatro experimentos numéricos se detallan a continuación.

Explicación de las tablas

En las tablas, *Original* designa la órbita elegida, *Inicial* el conjunto de constantes orbitales usadas como valores de partida para los cálculos de corrección. *S1* y *S2* son los resultados obtenidos al minimizar las funciones (5) y (6) respectivamente. *Thiele* representa los elementos resultantes de minimizar (6) en función de las variables de Thiele-Innes y *L1* el resultado de minimizar (3).

Con el fin de comprobar el ajuste de las órbitas recobradas por cada método, se calcularon para cada una de las órbitas descritas descritas en el párrafo anterior las sumas de cuadrados de las diferencias observación - cálculo para ángulos ($\sum(\Delta\theta)^2$) y distancias ($\sum(\Delta\rho)^2$) así como la media y desviación típica de estas diferencias. En el caso de la órbita *Original* las medias y desviaciones típicas corresponden a los datos usados como "medidas". Estas cantidades nos informan del ajuste con relación a los datos medidos, que en un caso ideal deberían ser iguales a los obtenidos para la órbita original, pero no de su aproximación a los valores teóricos. Para medir esta aproximación se calcularon las desviaciones típicas de las diferencias entre las efemérides de cada órbita y la órbita original.

Primer caso $\sigma_\theta = 1.0, \quad \sigma_\rho = 0.1$

Tabla I. Orbitas

	a	e	P	T	Ω	ω	i
Original	1.030	0.760	56.40	1885.60	55.80	105.00	32.00
Inicial	1.080	0.700	50.40	1890.60	59.80	120.00	39.00
S1	1.024	0.758	56.44	1885.74	61.93	100.12	29.61
S2	1.005	0.748	56.49	1885.69	55.41	106.34	26.58
Thiele	1.024	0.758	56.44	1885.85	63.87	98.64	29.48
L1	1.033	0.758	56.41	1885.68	58.38	103.13	30.61

Tabla II. Diferencias con las observaciones

	$\sum(\Delta\theta)^2$	$\sum(\Delta\rho)^2$	$\overline{\Delta\theta}$	$s_{\Delta\theta}$	$\overline{\Delta\rho}$	$s_{\Delta\rho}$
Original	$1.3054E{+}2$	$1.3200E{+}0$	$1.3378E{-}1$	$9.6328E{-}1$	$1.2072E{-}2$	$9.7048E{-}2$
Inicial	$1.4586E{+}5$	$7.0902E{+}0$	$-1.406E{+}1$	$2.9288E{+}1$	$7.1597E{-}2$	$2.1498E{-}2$
S1	$1.5036E{+}2$	$1.2719E{+}0$	$-5.8442E{-}3$	$1.0438E{+}0$	$3.5642E{-}4$	$9.6003E{-}2$
S2	$2.3462E{+}2$	$1.2765E{+}0$	$1.0650E{-}1$	$1.2995E{+}0$	$8.6696E{-}4$	$9.6172E{-}2$
Thiele	$1.9273E{+}2$	$1.2673E{+}0$	$6.3595E{-}2$	$1.1800E{+}0$	$3.6860E{-}4$	$9.5829E{-}2$
L1	$1.3444E{+}2$	$1.2825E{+}0$	$-1.4570E{-}2$	$9.8690E{-}1$	$-2.0699E{-}3$	$9.6382E{-}2$

Tabla III. Diferencias con la órbita Original

	$\sigma(\Delta\theta)$	$\sigma(\Delta\rho)$
S1	$3.9252E{-}1$	$1.0092E{-}2$
S2	$8.6004E{-}1$	$1.4632E{-}2$
Thiele	$6.7730E{-}1$	$1.2194E{-}2$
L1	$2.1165E{-}1$	$6.4389E{-}3$

Cambiando las condiciones iniciales tenemos

Tabla I-bis. Orbitas

	a	e	P	T	Ω	ω	i
Original	1.030	0.760	56.40	1885.60	55.80	105.00	32.00
Inicial	1.080	0.720	53.40	1887.60	57.80	103.00	30.00
S1	1.024	0.758	56.44	1885.74	61.89	100.15	29.63
S2	1.021	0.755	56.46	1885.77	60.95	101.19	28.99
Thiele	1.020	0.756	56.47	1885.80	62.82	99.57	28.88
L1	1.033	0.758	56.41	1885.68	58.38	103.13	30.61

Tabla II-bis. Diferencias con las observaciones

	$\sum(\Delta\theta)^2$	$\sum(\Delta\rho)^2$	$\overline{\Delta\theta}$	$s_{\Delta\theta}$	$\overline{\Delta\rho}$	$s_{\Delta\rho}$
Original	$1.3054E{+}2$	$1.3200E{+}0$	$1.3378E{-}1$	$9.6328E{-}1$	$1.2072E{-}2$	$9.7048E{-}2$
Inicial	$1.4434E{+}4$	$2.6108E{+}0$	$1.0382E{-}1$	$1.0226E{+}1$	$-3.1418E{-}2$	$1.3388E{-}1$
S1	$1.5032E{+}2$	$1.2686E{+}0$	$2.0422E{-}3$	$1.0437E{+}0$	$5.3808E{-}4$	$9.6003E{-}2$
S2	$1.8194E{+}2$	$1.2662E{+}0$	$6.9912E{-}2$	$1.1460E{+}0$	$3.1371E{-}4$	$9.5879E{-}2$
Thiele	$1.9136E{+}2$	$1.2673E{+}0$	$5.3320E{-}2$	$1.1764E{+}0$	$4.3435E{-}4$	$9.5786E{-}2$
L1	$1.3444E{+}2$	$1.2825E{+}0$	$-1.4570E{-}2$	$9.8690E{-}1$	$-2.0699E{-}3$	$9.6382E{-}2$

Tabla III-bis. Diferencias con la órbita Original

	$\sigma(\Delta\theta)$	$\sigma(\Delta\rho)$
S1	$3.9269E{-}1$	$1.0005E{-}2$
S2	$6.2112E{-}1$	$1.0806E{-}2$
Thiele	$6.7191E{-}1$	$1.2150E{-}2$
L1	$2.1165E{-}1$	$6.4389E{-}3$

La órbita elegida tiene una excentricidad moderadamente alta (0.76) y una inclinación media (32°). Las "medidas" de ángulos y distancias se prepararon con unas desviaciones típicas de 1° y de 0.1 segundos de arco respectivamente. Esto significa que el 95% de las medidas tienen un error menor de 3° y 0.3 segundos de arco. Puede considerarse, por tanto, una órbita muy bien medida.

La órbita *Inicial* es una órbita francamente mala, con un error medio de 14° y una desviación típica de 29°. La suma de cuadrados de diferencias angulares es 1000 veces la inicial.

Partiendo de estos datos se obtuvieron las órbitas mostradas en la Tabla I. Las diferencias observación - cálculo correspondientes se muestran en la Tabla II.

En todos los casos, las sumas de cuadrados son del mismo orden e inferiores a las iniciales, y en el caso de las distancias (ρ), son inferiores a la original.

Con el fin de investigar la posible influencia de los valores iniciales en los resultados obtenidos se probó una segunda órbita *Inicial* .Las soluciones obtenidas con estas nuevas condiciones iniciales se muestran en la Tabla I-bis y las diferencias observación - cálculo en la Tabla II-bis. La órbita *Inicial*-bis es bastante mejor, con una desviación típica de 10° frente a los 29 del caso anterior. Las sumas de cuadrados de diferencias observación - cálculo de ángulos y distancias son 1/10 y 1/3, respectivamente, de las diferencias de la órbita anterior.

Los resultados obtenidos son sensiblemente idénticos en ambos casos. Coinciden hasta las centésimas en el caso *L1* y difieren en unas centésimas de grado para *S1*. La órbita

Thiele difiere aproximadamente en un grado. Las mayores diferencias se encuentran en la orientación de la órbita de *S2*. Las diferencias observación - cálculo son, sin embargo, bastante similares.

En cuanto a los parámetros recobrados, el semieje mayor se ha obtenido con una diferencia de unas pocas milésimas. La excentricidad con una diferencia máxima de 0.12 y es menor que la original en todos los casos. Los errores en el período oscilan entre 0.01 y 0.09 años y los valores obtenidos son superiores al original en todos los casos. El periastro se ha recuperado con una precisión similar a la del período y también es mayor que en la órbita original.

Las mayores variaciones entre los distintos métodos se producen en la determinación de la orientación de la órbita. El argumento de periastro ω y la inclinación i obtenidos, son menores que los originales, aunque los valores iniciales son mayores en un caso y menores en el otro. El ángulo del nodo Ω obtenido es siempre mayor.

En conjunto, y a la vista de las diferencias observación - cálculo, la suma de cuadrados de diferencias de distancias es aproximadamente igual para los cuatro métodos. *L1* proporciona el mejor ajuste de ángulos. Los elementos dinámicos a, e, P y T se recobran bien en todos los casos. El valor de a obtenido por *S2* es bastante más pequeño (1.005) que el original (1.030), sin embargo, las diferencias observación - cálculo obtenidas son comparables con el resto de los resultados. La precisión en la recuperación de los ángulos es bastante menor.

En la tabla III aparecen las desviaciones típicas de las diferencias entre cada órbita obtenida y la *Original*. Podemos ver que la órbita *L1* es la que presenta menores desviaciones. Esto concuerda con que sus parámetros sean los más cercanos a la órbita *Original* y produzca la menor suma de cuadrados en los ángulos. En las distancias, en cambio, la órbita *S1* tiene una suma de cuadrados levemente menor que *L1*, pero la desviación con respecto a la *Original* es casi doble.

Segundo caso $\sigma_\theta = 1.0, \quad \sigma_\rho = 0.1$

Tabla IV. Orbitas

	a	e	P	T	Ω	ω	i
Original	0.120	0.500	50.00	1925.00	95.00	60.00	60.00
Inicial	0.130	0.450	47.00	1922.00	84.00	65.00	55.00
S1	0.134	0.497	49.99	1924.98	95.41	59.59	59.93
S2	0.123	0.489	49.51	1924.60	95.44	55.38	53.46
Thiele	0.124	0.487	49.59	1924.47	94.25	55.13	52.50
L1	0.124	0.498	49.99	1925.00	95.97	59.60	59.92

Tabla V. Diferencias con las observaciones

	$\sum(\Delta\theta)^2$	$\sum(\Delta\rho)^2$	$\overline{\Delta\theta}$	$s_{\Delta\theta}$	$\overline{\Delta\rho}$	$s_{\Delta\rho}$
Original	1.3056E+2	1.3200E+0	1.3416E−1	9.6330E−1	1.2073E−2	9.7048E−2
Inicial	1.9256E+5	1.3326E+0	−2.145E+1	3.0527E+1	1.6425E−3	9.8254E−2
S1	1.2511E+2	1.2983E+0	−1.9765E−3	9.5196E−1	−1.5084E−4	9.6996E−2
S2	2.9843E+3	1.2863E+0	−3.9064E−1	4.6338E+0	8.6843E−4	9.6540E−2
Thiele	3.6428E+3	1.2859E+0	−9.3281E−1	5.0518E+0	−1.0125E−3	9.6526E−2
L1	1.5515E+2	1.3085E+0	−4.5510E−1	9.5691E−1	8.3900E−3	9.7011E−2

Tabla VI. Diferencias con la órbita Original

	$\sigma(\Delta\theta)$	$\sigma(\Delta\rho)$
S1	$1.9451E{-}1$	$3.6374E{-}3$
S2	$4.5989E{-}1$	$6.1476E{-}3$
Thiele	$5.0217E{+}0$	$6.5649E{-}3$
L1	$1.7336E{-}1$	$1.1790E{-}3$

Esta órbita es un poco más inclinada que la del caso anterior (60°) y tiene un semieje 10 veces más pequeño. Los errores en las medidas son los mismos que en el caso anterior, lo que significa que los errores en la medida de distancias son del mismo orden que las medidas. La órbita está bien medida en ángulos y mal en distancias. Las diferencias observación - cálculo de la órbita *Inicial* son del mismo orden que en el caso anterior. Asimismo, los valores de las diferencias obtenidas después de la corrección son también similares. Al igual que en el primer caso, las diferencias para las distancias son menores que las de la órbita *Original*.

Por elementos, el semieje mayor obtenido es bastante próximo al original, excepto en *S1*, y mayor que él en todos los casos. La excentricidad obtenida es muy cercana a la original, aunque menor. La máxima diferencia es 0.013 para *Thiele*. El período está perfectamente recobrado en *S1* y *L1* y es levemente inferior en los otros casos, al igual que ocurre con la época de paso por el periastro.

En cuanto a la orientación de la órbita, los ángulos obtenidos por *S1* y *L1* son bastante parecidos entre sí y próximos a los originales. *S2* y *Thiele* recobran ω e i bastante peor.

En conjunto, la mejor órbita en cuanto a diferencias de cuadrados de ángulos es *S1*, aunque las diferencias de distancias no son tan buenas debido al valor obtenido para el semieje.

Sin embargo, comparando las desviaciones típicas de las diferencias en la tabla VI, se puede ver que la órbita *Thiele* está bastante alejada de la *Original*, con un error 10 veces mayor que *S2*, cosa que no se deduce de las diferencias con respecto a las medidas. Igualmente, *S1*, con unas sumas de cuadrados menores que *L1*, se desvía más que ésta de la *Original*.

Tercer caso $\sigma_\theta = 5.0, \quad \sigma_\rho = 0.05$

Tabla VII. Orbitas

	a	e	P	T	Ω	ω	i
Original	0.120	0.500	50.00	1925.00	95.00	60.00	60.00
Inicial	0.130	0.450	47.00	1922.00	84.00	65.00	55.00
S1	0.126	0.489	49.95	1924.94	97.02	58.31	60.19
S2	0.121	0.495	49.70	1924.72	95.17	57.69	56.34
Thiele	0.116	0.459	49.94	1923.93	97.79	49.51	53.41
L1	0.122	0.491	49.98	1924.98	97.92	58.01	59.62

Tabla VIII. Diferencias con las observaciones

	$\sum(\Delta\theta)^2$	$\sum(\Delta\rho)^2$	$\overline{\Delta\theta}$	$s_{\Delta\theta}$	$\overline{\Delta\rho}$	$s_{\Delta\rho}$
Original	3.2645E+3	3.2995E−1	6.6970E−1	4.8171E+0	6.0370E−3	4.8521E−2
Inicial	1.8796E+5	3.6637E−1	−2.0915E+1	3.0354E+1	−4.3953E−3	5.1337E−2
S1	3.1099E+3	3.2499E−1	−1.0763E−2	4.7471E+0	6.0224E−4	4.8524E−2
S2	4.1113E+3	3.2122E−1	−2.2927E−1	5.4534E+0	4.2184E−4	4.8244E−2
Thiele	6.3013E+3	3.2778E−1	6.0272E−1	6.7302E+0	−7.0485E−3	4.8220E−2
L1	3.1864E+3	3.2572E−1	−4.5214E−1	4.7837E+0	3.1290E−3	4.8481E−2

Tabla IX. Diferencias con la órbita Original

	$\sigma(\Delta\theta)$	$\sigma(\Delta\rho)$
S1	8.0689E−1	2.3518E−3
S2	2.8490E+0	3.4156E−3
Thiele	6.3097E+0	5.1494E−3
L1	7.5390E−1	1.7559E−3

Este caso tiene los mismos elementos orbitales *Original* e *Inicial* que el caso anterior pero los errores en las medidas de los ángulos son ahora 5 veces mayores y los errores en la medida de distancias, la mitad.

Los semiejes obtenidos son más próximos al original que en el caso anterior, sobre todo en el valor proporcionado por *S1*. En cambio la excentricidad está peor recobrada en todos los casos excepto en *S2*. Los períodos son similares al caso anterior, así como las épocas de paso por el periastro. De nuevo, los valores proporcionados por *S1* y *L1* son los mejores.

Los ángulos Ω y ω están mucho más alejados de los valores originales que en el caso 2, sin embargo las inclinaciones obtenidas son muy similares.

En comparación con el caso anterior, la suma de diferencias de cuadrados de distancias es 4 veces menor en correspondencia con la disminución del error inicial y las diferencias de ángulos han aumentado con el error de los ángulos. Esta variación es menos perceptible para *S2* y *Thiele*, cuyas sumas de errores se han duplicado, mientras que las de *S1* y *L1* se han multiplicado por 20.

La comparación con la órbita *Original* (Tabla IX) vuelve a mostrar como más cercana a la original la órbita *L1*. Podemos notar, igualmente, que a pesar de tener sumas de diferencias de cuadrados muy similares, sus diferencias con respecto a la *Original* son bien distintas.

Cuarto caso $\sigma_\theta = 5.0, \quad \sigma_\rho = 0.05$

Tabla X. Orbitas

	a	e	P	T	Ω	ω	i
Original	0.120	0.100	50.00	1925.00	95.00	60.00	60.00
Inicial	0.130	0.150	47.00	1922.00	84.00	65.00	55.00
S1	0.128	0.093	50.03	1924.25	91.47	54.04	60.16
S2	0.125	0.084	49.68	1922.98	90.67	45.61	58.38
Thiele	0.125	0.085	49.72	1923.05	90.29	46.36	58.34
L1	0.126	0.101	50.02	1924.84	91.24	58.51	60.67

Tabla XI. Diferencias con las observaciones

	$\sum(\Delta\theta)^2$	$\sum(\Delta\rho)^2$	$\overline{\Delta\theta}$	$s_{\Delta\theta}$	$\overline{\Delta\rho}$	$s_{\Delta\rho}$
Original	$3.2645E{+}3$	$3.3003E{-}1$	$6.6993E{-}1$	$4.8170E{+}0$	$6.0367E{-}3$	$4.8527E{-}2$
Inicial	$1.3286E{+}5$	$3.9057E{-}1$	$-2.3857E{+}1$	$1.9735E{+}1$	$-8.4389E{-}3$	$5.2521E{-}2$
S1	$3.1175E{+}3$	$3.2534E{-}1$	$-1.9491E{-}2$	$4.7529E{+}0$	$-5.3039E{-}5$	$4.8386E{-}2$
S2	$4.7693E{+}3$	$3.2311E{-}1$	$-2.1274E{-}1$	$5.2219E{+}0$	$3.7732E{-}4$	$4.8386E{-}2$
Thiele	$3.6786E{+}3$	$3.2329E{-}1$	$-4.0640E{-}2$	$5.1628E{+}0$	$3.8655E{-}4$	$4.8399E{-}2$
L1	$3.1721E{+}3$	$3.2581E{-}1$	$-1.2777E{-}1$	$4.7927E{+}0$	$1.9542E{-}3$	$4.8550E{-}2$

Tabla XII. Diferencias con la órbita Original

	$\sigma(\Delta\theta)$	$\sigma(\Delta\rho)$
S1	$7.7806E{-}1$	$1.8962E{-}3$
S2	$2.5043E{+}0$	$1.9218E{-}3$
Thiele	$2.3537E{+}0$	$1.8421E{-}3$
L1	$5.3686E{-}1$	$1.5679E{-}3$

En este caso, los errores de "medida" son los mismos que en el caso anterior, pero la excentricidad de la órbita elegida es menor (0.1). En una órbita menos excéntrica, el periastro y los ángulos relacionados con él estarán mal determinados, admitiendo unos errores mayores, aunque la suma de cuadrados de diferencias sea del mismo orden.

Partiendo de unos valores iniciales con diferencias observación - cálculo similares al caso 3, se obtuvieron unos semiejes algo más alejados del original que el caso anterior y también mayores. La excentricidad está mejor recobrada, aunque proporcionalmente estaba más alejada de la verdadera. Como en los casos anteriores, los mejores valores fueron los obtenidos usando *S1* y *L1*, y son menores que en la *Original*, salvo el valor marginal de *L1* (1.001). Los períodos recobrados por S1 y *L1* son prácticamente idénticos al original y un poco inferiores los obtenidos por *S2* y *Thiele*. Los valores de la época de paso por el periastro recobrados por *S1* y *L1* son comparables, aunque algo peores que los del caso 3. La determinación de este elemento realizada por *S2* y *Thiele* es francamente pobre.

Como era de esperar, la recuperación de ω en este caso es peor, excepto para *L1*, que proporciona un valor similar al caso anterior. Las inclinaciones muestran poca influencia de la excentricidad en *S1* y *L1* y dando mejores valores en este caso que en el anterior. En cambio, el ángulo del nodo recobrado en este caso está más alejado del original con un error del orden de 5° frente a 2° en el caso 3.

Sin embargo, a pesar de esta peor determinación de los parámetros orbitales, las sumas de cuadrados de diferencias son similares al caso anterior. Al igual que en el resto de los casos anteriores, la suma de los cuadrados de las diferencias de distancias es menor que la original.

Resumen de todos los casos

En todos los casos el método de ajuste tendió a dar una media de diferencias cercana a 0. Este efecto es más perceptible en las distancias que en los ángulos. A pesar de las

discrepancias en los valores de las sumas de diferencias obtenidas por los distintos métodos de ajuste, las desviaciones típicas de éstas se mantienen relativamente constantes y son del orden de la desviación típica de la órbita *Original*, que nos da la precisión de las medidas. Así pues, estas desviaciones típicas pueden servir de índice de la precisión global de la órbita.

Como era de esperar, *S1* proporciona siempre una suma de cuadrados menor que la de los otros métodos, ya que trata de minimizar dicha función. No obstante, *L1* proporciona unos resultados comparables en todos los casos. Este método es el que ha mostrado un comportamiento más regular y es menos sensible a los cambios en las condiciones iniciales.

Todos los métodos usados han proporcionado valores de la excentricidad ligeramente inferiores a la original, si bien *S1* y *L1* han dado siempre unos valores muy próximos. El período se ha determinado bien en todos los casos, ya que las medidas abarcaban más de una revolución. La época de paso por el periastro y el argumento de periastro se determinan tanto mejor cuanto mayor es la excentricidad.

En conjunto, los métodos *S1* y *L1* recuperan mejor los parámetros orbitales que los métodos de Thiele-Innes. Esto puede ser debido a que las magnitudes medidas (ρ, θ), son aquellas sobre las que se minimiza, mientras que el método de Thiele-Innes, al minimizar las distancias, mezcla observaciones de ángulos y distancias que no están correlacionadas entre sí. Este método parece preferible cuando las medidas a ajustar provengan de placas fotográficas en las que se miden las distancias x e y en vez de θ y ρ.

Cuando se comparan las órbitas recuperadas con la *Original*, el método *L1* ha proporcionado en todos los casos estudiados una órbita más próxima que los restantes.

Desde un punto de vista más práctico, la implementación efectuada de *S1* es mucho más rápida que *L1* (1 minuto frente a 10 minutos). Esta diferencia es atribuíble al método de iteración tan simple usado en *L1*, frente a la eficiencia del método de Levenberg-Marquardt. Otro punto a tener en cuenta es el aumento de tiempo de cálculo con el número de puntos. El tiempo usado por el algoritmo de minimización usado en *L1*, derivado del "simplex", crece mucho más rápidamente que el algoritmo de Levenberg-Marquardt, llegando en ocasiones a tardar varias horas en casos con más de 150 puntos.

4.2 Casos reales

Una vez visto el comportamiento de los distintos método de ajuste frente a casos simulados, vamos aplicarlos a estrellas reales para comparar las correcciones obtenidas con cada uno de ellos. Se aplicaron los cuatro métodos a pesar de que *S2* y *Thiele* proporcionan resultados menos fiables.

En las tablas que siguen, aparecen los parámetros orbitales de partida (*Original*) y los obtenidos con cada uno de los métodos con la misma notación que en el apartado anterior. En todos los casos se trata de órbitas que cubren casi completamente una revolución y tienen un número suficiente de puntos (> 50). A todas las medidas se les asignó el mismo peso.

ADS 1538

	a	e	P	T	Ω	ω	i
Original	1.050	0.708	170.30	1893.35	40.41	220.72	73.59
S1	1.049	0.740	160.76	1892.29	38.16	225.11	74.18
S2	1.079	0.764	163.96	1892.72	37.23	228.46	74.41
Thiele	1.101	0.772	165.05	1893.18	36.38	230.20	74.86
L1	1.015	0.714	162.26	1892.64	39.47	221.67	73.08

Diferencias ADS 1538

	$\sum(\Delta\theta)^2$	$\sum(\Delta\rho)^2$	$\overline{\Delta\theta}$	$s_{\Delta\theta}$	$\overline{\Delta\rho}$	$s_{\Delta\rho}$
Original	1.3953 E+3	9.7226 E+0	5.6240 E-1	2.4370 E+0	-3.2046 E-2	2.0632 E-1
S1	1.2821 E+3	8.5123 E+0	4.6917 E-3	2.3978 E+0	1.7822 E-3	1.9537 E-1
S2	1.4954 E+3	8.4982 E+0	-1.2250 E-1	2.5867 E+0	1.8120 E-3	1.9521 E-1
Thiele	1.7067 E+3	8.4990 E+0	4.4623 E-1	2.7301 E+0	1.2964 E-3	1.9522 E-1
L1	1.2507 E+3	8.7365 E+0	1.3577 E-2	2.3682 E+0	1.3588 E-2	1.9746 E-1

Esta órbita tiene muchas medidas (224) a lo largo de casi una revolución. Es bastante excéntrica y bastante inclinada, lo que produce una elipse aparente muy estrecha, que dificulta su cálculo. Por otra parte, la periodicidad de signos en las diferencias parece sugerir la existencia de una estrella triple.

La variabilidad de los elementos obtenidos es grande, tanto en los elementos dinámicos como en la orientación aunque las sumas de cuadrados son similares. En conjunto, las medidas son buenas, ya que la desviación típica de las diferencias angulares es pequeña (2°). En cambio, las diferencias en las distancias son proporcionalmente mayores (casi 2 décimas de segundo de arco con un semieje de 1).

Atendiendo a las sumas de cuadrados, las órbitas *S1* y *L1* son más parecidas entre sí que a *S2* o *Thiele*.

ADS 1709

	a	e	P	T	Ω	ω	i
Original	0.908	0.253	143.60	1898.80	99.20	321.60	63.00
S1	0.918	0.249	140.38	1897.71	98.93	317.72	63.12
S2	0.919	0.249	144.02	1898.81	98.81	321.07	62.15
Thiele	0.922	0.252	140.54	1897.48	98.82	312.18	63.66
L1	0.907	0.252	142.11	1898.77	98.42	321.24	63.09

Diferencias ADS 1709

	$\sum(\Delta\theta)^2$	$\sum(\Delta\rho)^2$	$\overline{\Delta\theta}$	$s_{\Delta\theta}$	$\overline{\Delta\rho}$	$s_{\Delta\rho}$
Original	$1.6574E+3$	$9.2018E-1$	$-4.4821E-1$	$3.1445E+0$	$1.9527E-2$	$6.9819E-2$
S1	$1.6861E+3$	$8.3979E-1$	$-7.8861E-3$	$3.1040E+0$	$4.5707E-3$	$6.9122E-2$
S2	$2.2843E+3$	$8.2211E-1$	$1.3975E+0$	$3.3300E+0$	$2.9587E-3$	$6.8476E-2$
Thiele	$2.0650E+3$	$8.1311E-1$	$5.1633E-1$	$3.3958E+0$	$4.5800E-3$	$6.8009E-2$
L1	$1.6438E+3$	$8.9788E-1$	$-1.4706E-1$	$3.0613E+0$	$1.3771E-2$	$7.0285E-2$

Se trata de una órbita con bastantes observaciones (> 179) a lo largo de algo más de un período. En conjunto, dichas observaciones son buenas con una desviación típica en ángulos del orden de 3°

Los ajustes proporcionados por los métodos *S1* y *L1* son bastante similares y cercanos a los de partida. Las sumas de cuadrados de diferencias de ángulos son prácticamente iguales para ambas órbitas. En la suma correspondiente a distancias hay algo más de diferencia.

Analizando las diferencias observación - cálculo vemos que existen 3 puntos cuyas diferencias angulares son mayores de 10°. Estas medidas tienen muy probablemente un error no estadístico, ya que al exceder 3 veces la desviación típica, tienen menos del 1% de probabilidad de pertenecer a una distribución de tipo normal

Eliminando estas tres medidas, y repitiendo la corrección se obtienen los parámetros orbitales y diferencias que aparecen a continuación.

ADS 1709

	a	e	P	T	Ω	ω	i
Original	0.908	0.253	143.60	1898.80	99.20	321.60	63.00
S1	0.917	0.249	140.24	1897.71	99.14	317.67	63.03
S2	0.921	0.254	140.02	1896.38	99.48	314.16	63.27
Thiele	0.917	0.249	141.08	1897.42	99.49	317.27	62.77
L1	0.906	0.252	142.09	1898.80	98.45	321.28	63.05

Diferencias 1709

	$\sum(\Delta\theta)^2$	$\sum(\Delta\rho)^2$	$\overline{\Delta\theta}$	$s_{\Delta\theta}$	$\overline{\Delta\rho}$	$s_{\Delta\rho}$
Original	9.3041 E+2	8.9418 E-1	-1.9384 E-1	2.3177 E+0	1.9879 E-2	6.9291 E-2
S1	9.5726 E+2	8.1270 E-1	1.8106 E-2	2.3811 E+0	4.6531 E-3	6.8580 E-2
S2	1.1922 E+3	7.8998 E-1	3.8349 E-2	2.6324 E+0	3.5340 E-3	6.7678 E-2
Thiele	1.0947 E+3	7.9611 E-1	1.9369 E-2	2.5227 E+0	4.0137 E-3	6.7914 E-2
L1	9.3073 E+2	8.7034 E-1	1.0374 E-1	2.3239 E+0	1.4504 E-2	6.9632 E-2

Lo primero que se puede ver en las tablas es que estos 3 puntos son reponsables de casi el 50% de las sumas de cuadrados de la órbita *Original* y reducen la desviación típica de 3.1445 a 2.3177. Examinando los resultados del ajuste se puede constatar que la disminución de la suma de cuadrados de diferencias de ángulos se ha reducido casi a la mitad, mientras que las desviaciones típicas se reducen en un 25%. Como se ve, un filtrado de datos tan simple como éste, mejora el conjunto del ajuste. Debemos notar que las variaciones de los parámetros orbitales obtenidas con el filtrado son muy similares a las obtenidas sin él, siendo *L2* el que tiene las variaciones mayores. Como era de esperar, *L1* presenta las diferencias menores. Como en los casos simulados, esta órbita obtuvo la mínima suma de cuadrados de ángulos.

ADS 11077

	a	e	P	T	Ω	ω	i
Original	1.123	0.798	56.04	1942.35	63.50	98.90	43.20
S1	1.095	0.759	56.03	1941.41	40.51	116.56	42.40
S2	1.072	0.766	55.91	1941.79	46.71	112.31	39.77
Thiele	1.089	0.766	55.89	1941.59	43.64	114.11	41.77
L1	1.067	0.753	56.16	1941.51	39.22	118.21	38.84

Diferencias ADS 11077

	$\sum(\Delta\theta)^2$	$\sum(\Delta\rho)^2$	$\overline{\Delta\theta}$	$s_{\Delta\theta}$	$\overline{\Delta\rho}$	$s_{\Delta\rho}$
Original	4.3954 E+3	2.9087 E+0	2.1910 E-1	5.6393 E+0	5.8908 E-3	1.4506 E-1
S1	2.5748 E+3	2.7858 E+0	1.4480 E-2	4.3194 E+0	4.9392 E-3	1.4199 E-1
S2	3.3455 E+3	2.6960 E+0	3.8352 E-1	4.9086 E+0	5.5163 E-3	1.3966 E-1
Thiele	3.0938 E+3	2.7038 E+0	4.2869 E-1	4.7152 E+0	-2.7483 E-3	1.3995 E-1
L1	2.4762 E+3	2.9329 E+0	4.1173 E-1	4.2158 E+0	-1.1287 E-2	1.4534 E-1

Las medidas de este par cubren 120 años, lo que viene a ser más de dos revoluciones completas. Esto hace que el período se pueda determinar con bastante precisión. Las épocas de medida de esta estrella se usaron para los casos simulados.

Los cuatro métodos proporcionan un ajuste mejor que la órbita *Original*, con unos parámetros dinámicos muy similares entre sí y una dispersión mayor de la orientación de la órbita.

ADS 13169

	a	e	P	T	Ω	ω	i
Original	0.380	0.260	126.49	1905.45	131.80	333.30	50.40
S1	0.336	0.228	126.48	1905.30	129.25	337.15	47.17
S2	0.367	0.226	136.69	1908.04	135.51	345.04	42.91
Thiele	0.367	0.226	134.25	1907.31	134.81	342.26	43.69
L1	0.371	0.217	136.75	1909.98	128.37	357.84	41.62

Diferencias ADS 13169

	$\sum(\Delta\theta)^2$	$\sum(\Delta\rho)^2$	$\overline{\Delta\theta}$	$s_{\Delta\theta}$	$\overline{\Delta\rho}$	$s_{\Delta\rho}$
Original	9.6292 E+2	1.0133 E-1	-7.9719 E-1	4.2700 E+0	-4.5369 E-3	4.4388 E-2
S1	9.0698 E+2	1.2956 E-1	-1.4318 E-2	4.2171 E+0	2.9521 E-2	4.0642 E-2
S2	1.0055 E+3	8.0944 E-2	-1.1340 E-1	4.4387 E+0	1.7607 E-3	3.9812 E-2
Thiele	9.9498 E+2	8.1051 E-2	2.6481 E-2	4.4169 E+0	1.2398 E-3	3.9846 E-2
L1	9.6014 E+2	8.9729 E-2	-7.9243 E-1	4.2645 E+0	-4.8631 E-3	4.1657 E-2

Las observaciones de esta órbita cubren 80 años, lo que viene a ser algo menos de una revolución.

En este caso podemos ver dos tipos de resultados. *Original* y *L1*, muy similares, con un período de 126.48 años y el resto con un período de 136 años. Las sumas de cuadrados $\sum(\Delta\theta)^2 + \sum(\Delta\rho)^2$ son similares en ambos casos.

Esta indeterminación del período puede deberse a que tenemos un arco abierto, con una determinación muy pobre del periastro.

Conclusiones

Las aplicaciones de los cuatro métodos a casos reales proporciona unos resultados comparables para todos ellos. Esto no se corresponde con los pobres resultados obtenidos por *S2* y *Thiele* para los casos simulados. El Criterio de Mínimos Valores Absolutos sigue siendo válido, con unos resultados comparables al de Mínimos Cuadrados, pero es mucho menos sensible a la aparición de medidas discordantes.

Agradecimientos

Agradecemos al Profesor J. A. Docobo el habernos facilitado las observaciones de estrellas dobles usadas.

BIBLIOGRAFIA

Barrodale,I. y Roberts, F. D. K.: 1973, An improved Algorithm for discrete L_1 linear Approximation, *SIAM J. Numer. Anal.*, **10**, 839–848

Branham, R. L.: 1982, Alternatives to Least Squares, *Astron. J.*, **87**, 928-937

Eichhorn, H. y Clary, W. G.: 1974, Least Squares Adjustement with Relatively Large Observation Errors, Inaccurate Initial Approximations, or both, *Monthly Notices Roy. Astron. Soc.*, **166** 425–432

Fletcher,W. y Powell, M.J.D. 1963, *Comput. J.*, **6**, 163

IMSL: 1985, IMSL User's Manual V 9.2

Jefferys, W. H.: 1980, On the Method of Least Squares, *Astron. J.*, **85**, 177–181

Jefferys, W. H.: 1981, On the Method of Least Squares II, *Astron. J.*, **86**, 149–155

Robers, P. D. y Robers, S. S.: 1971, *Collected Algorithms from ACM*, Algoritmo 458

Soulié, E. J.: 1986, L'amélioration de l'orbite d'une étoile double visuelle, *Astron. Astrophys.* **164**, **141**, 408–414

Späth, H. y Watson, G.A.:1987, On Orthogonal Linear l_1 Approximation, *Numerische Matematik*, **51**, 531–543

Rev. Acad. Ciencias. Zaragoza 45, (1990)

On the Polar Orbits for the Zeeman Effect in a Moderately Strong Magnetic Field

A. Deprit* and S. Ferrer
Departamento de Física Teórica (Astronomía)
Universidad de Zaragoza, 50009 Zaragoza, Spain

Abstract

In the ground state, the reduced equations for the averaged quadratic Zeeman effect in a small magnetic field are analogous to those of a rigid body in free rotation about a fixed point.

PACS numbers: 0320, 3260V, 3345B, 4170, 9510C

1. INTRODUCTION

Much has been said lately about the motion of an electron for an hydrogen-like atom in the field of a small uniform magnetic field, yet large enough that one should consider perturbations of the order of the square of the Larmor frequency. Apparently the interest sprung from Zimmerman *et al.* (1980) indicating that the system admits a symmetry, and from Solov'ev (1982) producing an approximate integral, at least to the first order, to explain the newly discovered symmetry.

The three-dimensional version of the quadratic Zeeman effect admits a one-parameter family of singular two-dimensional manifolds, each made of the orbits lying permanently in a fixed plane parallel to the magnetic field through the Coulomb center of attraction. In quantum mechanics, they correspond to the ground state. These manifolds have been analyzed extensively either by a so-called Gustavson normalization (Reinhardt and Farrelly, 1982; Robnik, 1984; Robnik and Schrüfer, 1985) or by Poincaré's cross-sections (Robnik, 1981; Harada and Hasegawa, 1983; Delande and Gay,

*Permanent address: National Institute of Standards and Technology, Gaithersburg, MD 20899, U. S. A.

1986; Saini and Farrelly, 1987). We propose to supplement these researches with a global portrait of the phase flow in the ground state from the standpoint of classical mechanics. In fact, we are completing on an essential point the classical theory developed by Richards (1983): the reader is probably aware that Richards' disc model does not accommodate the ground state.

Our model is built along the lines of a treatment given recently for the general three-dimensional problem (Coffey *et al.*, 1986). Normalization beyond the first order, i.e., past terms proportional to the second power of the Larmor frequency ω, is not needed; it suffices to average the diamagnetic perturbation without entering the complications of a full-fledged Lie transformation. Yet we need to know the infinitesimal contact transformation leading to such a reduction in order to establish that it has no effect on the linear orbits of the unreduced system. Therefore, in Section 3, we go beyond a straightforward application of Pauli's theory of secular perturbations in explaining how to obtain the generator of the averaging transformation.

Conventional Delaunay coordinates are used rather than the parabolic coordinates favored by authors interested in high order expansions. In the averaged problem, we introduce a new set of coordinates consisting of the norm of the angular momentum and of two orthogonal components of the Runge-Lenz vector. In these coordinates, the phase space in the domain of bound states appears as a two-dimensional sphere; even more significantly, the phase flow looks exactly like that of a rigid body in free rotation about a fixed point.

The analogy with the *rotator* will be examined in detail. In the spherical model, the circular orbits at ground state, one direct and the other retrograde, are mapped onto unstable equilibria at both ends of a diameter which we regard as the poles of the sphere. The great circle perpendicular to that axis, we call the equator. Each point on the equator represents a collinear orbit. Along the same equator we find four stable equilibria. Not unsurprisingly, these critical solutions are identical to special collinear solutions of the original system. The phase flow around these centers is separated by four homoclinic orbits emanating from one pole and ending at the other. That these asymptotic manifolds are indeed homoclinic is confirmed by the exact expressions we give for them in terms of the *lambda* function.

Poincaré surfaces of section drawn by Harada and Hasegawa (1983) or Delande and Gay (1986) seem to support our conclusions about the global phase flow in the ground state at first order. As one can see in their figures, chaos at low energy sets in near the unstable circular orbits where the breaking of the artificial symmetry imposed by the reduction causes the

asymptotic manifolds emananting from the unstable equilibria to loose their homoclinic character.

2. RECTILINEAR SOLUTIONS

In cylindrical coordinates, the quadratic Zeeman effect (QZE) is represented in the Gaussian electromagnetic units by the Hamiltonian

$$\mathcal{H} = \frac{1}{2}\left(P^2 + Z^2 + \frac{\Lambda^2}{\rho^2}\right) - \frac{\mu}{\sqrt{\rho^2 + z^2}} + \frac{\omega^2}{2}\rho^2 \qquad (1)$$

where $\mu = |qQ|/m\epsilon$ and $\omega = qB/2mc$ is the Larmor frequency. [For the meaning of the physical parameters m, q, Q, ϵ, B and c, the reader is referred to Coffey *et al.* 1986.] The coordinate z is the elevation above the plane perpendicular to the magnetic field through the center of the Coulomb field while ρ is the distance to the magnetic field line through the same center; P and Z are the momenta conjugate to ρ and z respectively. The system is referred to a coordinate frame precessing about the magnetic field at the rate ω. Since the longitude λ is ignorable, its conjugate momentum Λ, which is the projection of the angular momentum in the direction of the magnetic field, is an integral.

Here we are exclusively interested in the special manifold of orbits defined by the condition that Λ be zero. Orbits in that manifold are what we call the polar orbits, for they stay permanently in the meridian plane of their starting point. When confined to a meridian plane, the QZE is described by the Hamiltonian

$$\mathcal{H} = \frac{1}{2}(P^2 + Z^2) - \frac{\mu}{\sqrt{\rho^2 + z^2}} + \frac{\omega^2}{2}\rho^2. \qquad (2)$$

Among the polar orbits, we draw attention to two special solutions,

a) the linear solutions along the field line passing through the center of Coulombian attraction, thus those orbits for which the coordinate ρ and its conjugate momentum P are permanently zero;

b) the linear solutions perpendicular to the field lines, i.e., those solutions for which the coordinate z and its conjugate momentum Z are permanently zero. We call these the linear *equatorial* orbits.

These two classes of singular polar orbits, we shall show, play a critical role in determining the structure of the average phase space.

3. REDUCTION AT FIRST ORDER

From here on, we take ω small enough that we can view the QZE as a perturbed Keplerian system. Accordingly, the Hamiltonian in (2) is decomposed into the sum $\mathcal{H} = \mathcal{H}_0 + \omega^2\mathcal{H}_1$ of a Keplerian Hamiltonian

$$\mathcal{H}_0 = \tfrac{1}{2}(P^2 + Z^2) - \frac{\mu}{r}$$

and a perturbation function

$$\mathcal{H}_1 = \tfrac{1}{2}\rho^2.$$

Our purpose now is to build an infinitesimal contact transformation

$$\chi : (\rho, z, P, Z) \mapsto (\rho', z', P', Z')$$

to convert $\mathcal{H}$ into a power series

$$\mathcal{H}' = \sum_{n\geq 0} \frac{1}{n!}\omega^{2n}\mathcal{H}'_n$$

such that

$$\begin{aligned} &a) \quad \mathcal{H}'_0(\rho', z', P', Z') = \mathcal{H}_0(\rho, z, P, Z), \\ &b) \quad (\mathcal{H}'_0; \mathcal{H}'_1) = 0. \end{aligned}$$

As will be seen in Section 5, there is no need to carry out the normalization beyond the first order in ω^2 since the first order normalization determines the global behavior of the polar QZE in a definitive manner.

According to condition b), the principal part $\mathcal{H}'_0$ in the transformed Hamiltonian is an integral of the normalized system to the first order; incidentally, it can be made so to any order had we extended the normalization beyond the first order. This formal integral will be used to operate a reduction by one degree of freedom (Meyer, 1973; see also Marsden and Weinstein, 1974).

Being an infinitesimal contact transformation, χ is defined by the equations

$$\begin{aligned} \rho &= \rho' + \omega^2\frac{\partial \mathcal{W}}{\partial P'}, & P &= P' - \omega^2\frac{\partial \mathcal{W}}{\partial \rho'}, \\ z &= z' + \omega^2\frac{\partial \mathcal{W}}{\partial Z'}, & Z &= Z' - \omega^2\frac{\partial \mathcal{W}}{\partial z'} \end{aligned}$$

emanating from a generating function

$$\mathcal{W} \equiv \mathcal{W}(\rho', z', P', Z').$$

The term of first order in the transformed Hamiltonian and the generator must satisfy the identity

$$(\mathcal{W}, \mathcal{H}_0') + \mathcal{H}_1' = \mathcal{H}_1(\rho', z', P', Z'). \tag{3}$$

Resolving this identity is simple in the domain $\mathcal{B}$ of bound orbits. The reason is twofold:

a) All orbits of the Keplerian $\mathcal{H}_0$ are periodic wherever $\mathcal{H}_0 < 0$. Hence the vector space $\mathcal{C}^\infty(\mathcal{B})$ of smooth functions in $\mathcal{B}$ is the topological direct sum

$$\mathcal{C}^\infty(\mathcal{B}) = \ker\mathcal{L}_0|\mathcal{B} \oplus \mathrm{im}(\mathcal{L}_0|\mathcal{B})$$

of the kernel and the image of the Lie derivative

$$\mathcal{L}_0 \,:\, F \mapsto (F; \mathcal{H}_0)$$

restricted to $\mathcal{B}$ (Deprit, 1982; Cushman, 1984);

b) Anywhere in the phase space where $\mathcal{H}_0 < 0$, one can proceed in the Delaunay rather than in the cylindrical variables. With ℓ and L standing respectively for the mean anomaly and its conjugate momentum, the Keplerian Hamiltonian is expressed as the function

$$\mathcal{H}_0 = -\frac{\mu^2}{2L^2};$$

therefore its associated Lie derivative reduces to the single partial derivative

$$\mathcal{L}_0 = n\frac{\partial}{\partial \ell} \qquad \text{where} \qquad n = \mu^2/L^3,$$

the frequency n designating what astronomers call the *mean motion*. With $\mathcal{L}_0$ in this elementary form, it is readily seen that $\ker\mathcal{L}_0$ is made of the functions independent of ℓ. For instance, if F is periodic in ℓ, then the average $< F >_\ell$ of F over the mean anomaly belongs to $\ker\mathcal{L}_0$. Accordingly, in the domain of bound states, the identity (3) is resolved by adopting

$$\begin{aligned} \mathcal{H}_1' &= < \mathcal{H} >_\ell, \\ \mathcal{W} &= \int^\ell [\mathcal{H}_1 - < \mathcal{H} >_\ell]\, d\ell. \end{aligned}$$

Tradition in celestial mechanics would suggest at this point that the perturbation function $\mathcal{H}_1$ be expressed explicitly as a function of ℓ by developing it as a mixed series, in the powers of the eccentricity e on the one hand, and in cosines and sines of multiples of ℓ on the other hand. But such classical expansions do not make sense in atomic physics, and, fortunately, all sorts of techniques are currently in development so the operations of averaging and normalizing could be carried out uniformly for all eccentricities between 0 and 1. For instance, the polar QZE is most adequately handled with a technique devised sometime ago (Deprit, 1983) to deal with the Stark effect in two dimensions whereby the perturbation is expressed explicitly and *exactly* in terms of what astronomers know as the *eccentric anomaly*.

The Delaunay momentum G is the norm of the angular momentum; the coordinate canonically conjugate to G is the inclination g of the Runge-Lenz vector over the ρ-axis in the meridian plane, an angle that astronomers have dubbed the *argument of perigee*. To g is attached an angle f, often named the *true anomaly*, so as to have that

$$\rho = r\cos(f+g), \qquad z = r\sin(f+g).$$

The angle f itself is related to the mean anomaly ℓ through the *eccentric* anomaly E *via* the relations

$$r\cos f = a(cosE - e), \qquad r\sin f = a\eta\sin E.$$

The link between E and ℓ is provided by Kepler's equation

$$E - e\sin E = \ell. \tag{4}$$

In the definitions above, we have introduced a few abbreviations like a and η defined respectively by the relations $\mu = n^2a^3$ and $G = L\eta$, and also the eccentricity $e = (1-\eta^2)^{1/2}$.

Stopping short of forcing an inversion of the transcendental (4) to obtain E explicitly, although approximately, as a function of ℓ, we shall be satisfied with obtaining the perturbation as an exact Fourier series in E with coefficients in $\ker\mathcal{L}_0$:

$$\begin{aligned}\mathcal{H}_1 \;=\; a^2\Big\{&\frac{1}{4}+\frac{1}{8}e^2(1+3e^2\cos 2g)\\ &-\frac{1}{2}e(1+\cos 2g)\cos E+\frac{1}{2}e\eta\sin 2g\sin E\\ &+\frac{1}{8}[e^2+(2-e^2)\cos 2g]\cos 2E-\frac{1}{4}\eta\sin 2g\sin 2E\Big\}\end{aligned}$$

According to (4), $\partial E/\partial \ell = r/a$, therefore

$$\mathcal{H}_1' = \frac{1}{2\pi}\int_0^{2\pi} \mathcal{H}_1\, d\ell = \frac{1}{2\pi}\int_0^{2\pi} (1 - e\cos E)\mathcal{H}_1\, dE.$$

A few straightforward quadratures lead eventually to the crucial result

$$\mathcal{H}_1' = \frac{1}{4}a'^2\left[1 + \frac{1}{2}e'^2(3 + 5\cos 2g')\right]. \tag{5}$$

For the sake of comparison with some Poincaré sections available in the literature, it is worth observing that the normalizing transformation leaves unchanged, at least to the first order, a few remarkable curves in the domain of bound states. These are the curves ($G = 0$, L and g = constant) whose projection on the coordinate plane are lines through the origin either parallel ($g \equiv \pi/2 \bmod \pi$) or perpendicular ($g \equiv 0 \bmod \pi$) to the lines of the magnetic field. We shall now prove that χ maps these curves onto similar lines in the phase space (ℓ', g', L', G'). This we do by producing the generator of the normalizing transformation χ. According to (3),

$$n\frac{\partial \mathcal{W}}{\partial \ell} = \mathcal{H}_1 - < \mathcal{H}_1' >_\ell,$$

hence the generator stems from the quadrature

$$\begin{aligned} \mathcal{W} &= \frac{1}{n'}\int^{\ell} (\mathcal{H}_1 - < \mathcal{H}_1 >_\ell)\, d\ell \\ &= \frac{1}{n'}\int^{E} (1 - e\cos E)\,(\mathcal{H}_1 - < \mathcal{H}_1 >_\ell)\, dE. \end{aligned}$$

The result is the Fourier series in g'

$$\mathcal{W} = \frac{1}{2n'^2}(L'\mathcal{W}_0 + L'\mathcal{W}_1\cos 2g' + G'\mathcal{W}_2\sin 2g') \tag{6}$$

whose coefficients are the periodic functions in E'

$$\begin{aligned} \mathcal{W}_0 &= -\left(1 - \frac{3}{8}e'^2\right)e'\sin E' + \frac{3}{8}e'^2\sin 2E' - \frac{1}{24}e'^3\sin 3E', \\ \mathcal{W}_1 &= -\frac{5}{8}(2 - e'^2)e'\sin E' + \frac{1}{8}(2 + e'^2)\sin 2E' - \frac{1}{24}(2 - e'^2)e'\sin 3E', \\ \mathcal{W}_2 &= -\frac{5}{4}e'\cos E' + \frac{1}{4}(1 + e'^2)\cos 2E' - \frac{1}{12}e'\cos 3E'. \end{aligned}$$

By definition of an infinitesimal contact transformation, the Delaunay variables in the original problem are linked to the Delaunay variables in the averaged problem by the relations

$$\begin{aligned} \ell &= \ell' + \omega^2 \frac{\partial \mathcal{W}}{\partial L'}, & L &= L' - \omega^2 \frac{\partial \mathcal{W}}{\partial \ell'}, \\ g &= g' + \omega^2 \frac{\partial \mathcal{W}}{\partial G'}, & G &= G' - \omega^2 \frac{\partial \mathcal{W}}{\partial g'}. \end{aligned}$$

Thus, in the particular case of (6), we find that

$$\begin{aligned} g &= g' + \frac{1}{2}\frac{\omega^2}{n'^2}\mathcal{W}_2 \sin 2g' + \frac{\partial \mathcal{W}}{\partial e'}\frac{\partial e'}{\partial G'} + \frac{\partial \mathcal{W}}{\partial E'}\frac{\partial E'}{\partial G'}, \\ G &= G' + \frac{\omega^2}{n'^2}(L'\mathcal{W}_1 \sin 2g' + G'\mathcal{W}_2 \cos 2g'). \end{aligned} \tag{7}$$

The curves defined by the conditions ($G' = 0$, and $L', g' =$ constant) correspond to straight lines through the origin in the phase space (ℓ', g', L', G'). Along these straight lines, $f' = \pi$, hence

$$\frac{\partial e'}{\partial G'} = -\frac{G'}{e'L'^2} = 0 \quad \text{and} \quad \frac{\partial E'}{\partial G'} = -\frac{1}{e'L'} \sin f' = 0.$$

Under these conditions, according to (7),

$$\begin{aligned} g &= g' + \frac{1}{2}\frac{\omega^2}{n'^2}\mathcal{W}_2 \sin 2g', \\ G &= \frac{\omega^2}{n'^2} L' \mathcal{W}_1 \sin 2g'. \end{aligned}$$

There follows that the infinitesimal contact transformation χ maps the straight lines in anyone of the cardinal directions $g' \equiv 0 \bmod \pi/2$ onto themselves, at least to the first order.

4. GLOBAL REPRESENTATION

Since ℓ' is not contained in (5), it is an ignorable coordinate to the first order; the corresponding integral is the conjugate momentum L'. The equation $\dot{\ell}' = \partial \mathcal{H}'/\partial L'$ can be integrated by a quadrature when the rest of the equations have been solved; thus the equations for ℓ' and L' disappear from the reduced system.

Since, by assumption, $\omega \ll n'$, the time t will be replaced as the independent variable by a *slow* time τ such that $\omega\, dt = n'\, d\tau$, which means

essentially that the Larmor frequency will serve as the time scale for the reduced polar QZE. At the same time, terms independent of the variables g' and G' will be dropped from (5). We are thus left with a Hamiltonian

$$\mathcal{K}' = \frac{n'}{\omega}\left[\mathcal{H}' + \frac{1}{2}n'^2 a'^2 - \frac{1}{4}\omega^2 a'^2\right] = \frac{1}{8}\omega L' e'^2(3 + 5\cos 2g') \qquad (8)$$

which is none other than the integral produced by Solov'ev (1982). The equations of motion are

$$\begin{aligned} \frac{dg'}{d\tau} &= \frac{\partial \mathcal{K}'}{\partial G'} = -\frac{1}{4}\omega\eta'(3 + 5\cos 2g'), \\ \frac{dG'}{d\tau} &= -\frac{\partial \mathcal{K}'}{\partial g'} = \frac{5}{4}\omega L' e'^2 \sin 2g'. \end{aligned}$$

It must be observed at this point that the map (g', G') does not cover the entire phase space, for it excludes the points $e' = 0$ at which the argument of perigee g' is not defined. This pole-like singularity, inherent to the Delaunay map, does not belong to the system itself: it disappears when the system is handled in variables like

$$\zeta_1 = L' e' \cos g', \qquad \zeta_2 = L' e' \sin g', \qquad \zeta_3 = G'$$

where one recognizes the Cartesian components of the Runge-Lenz vector perpendicular and parallel to the magnetic field lines, and the norm of the angular momentum. In the global map $(\zeta_1, \zeta_2, \zeta_3)$, since

$$\zeta_1^2 + \zeta_2^2 + \zeta_3^2 = L'^2, \qquad (9)$$

the reduced phase space reveals itself as a bundle of two- dimensional spheres $S^2(L')$, one above each point of the axis of the formal integral L'. In geometric terms, the reduction performed in the preceding section has achieved a *fibration* of the phase space.

The points such that $\zeta_3 > 0$ which comprise what we shall call here the northern hemisphere of $S^2(L')$ stand for ellipses traveled in the direct sense while those for which $\zeta_3 < 0$ in the southern hemisphere represent ellipses traveled in the retrograde sense. Any point on the equatorial circle $\zeta_3 = 0$ corresponds to a segment of length $2a'$ along a straight line passing through the origin having precisely the origin as its midpoint. The north pole $(\zeta_1 = \zeta_2 = 0, \zeta_3 = L')$ is the circle of radius a' traveled in the direct sense, and the south pole $(\zeta_1 = \zeta_2 = 0, \zeta_3 = -L')$, the same circle but traveled in the retrograde sense.

In the global coordinates,

$$\mathcal{K}' = \frac{\omega}{4L'}(4\zeta_1^2 - \zeta_2^2);$$

therefore, on account of the Poisson brackets

$$(\zeta_1;\zeta_2) = \zeta_3, \qquad (\zeta_2;\zeta_3) = \zeta_1, \qquad (\zeta_3;\zeta_1) = \zeta_2,$$

the equations of motion on each sphere $S^2(L')$ are

$$\begin{aligned}
\frac{d\zeta_1}{d\tau} &= (\zeta_1;\mathcal{K}') = -\frac{\omega}{2L'}\zeta_2\zeta_3, \\
\frac{d\zeta_2}{d\tau} &= (\zeta_2;\mathcal{K}') = -\frac{2\omega}{L'}\zeta_3\zeta_1, \\
\frac{d\zeta_3}{d\tau} &= (\zeta_3;\mathcal{K}') = \frac{5\omega}{2L'}\zeta_1\zeta_2.
\end{aligned}$$

These are similar to the equations of motion of a rigid body in torque-free rotation about a fixed point:

$$A\dot{p} = (C-B)qr, \quad B\dot{q} = (A-C)rp, \quad C\dot{r} = (B-A)pq,$$

A, B, and C designating the principal moments of inertia while p, q and r stand for the components of the angular velocity in the principal frame of inertia. Our task now is to show that there is here more than an analogy of form. In any manifold G = constant, where G stands for the norm of the angular momentum, the motions of the Euler-Poinsot problem are represented by the level contours of the Hamiltonian $\frac{1}{2}(Ap^2+Bq^2+Cr^2)$ on the ellipsoid $G^2 = A^2p^2 + B^2q^2 + C^2r^2$; likewise, the orbits of the reduced polar QZE in the integral manifold L' = constant are the level contours of the Hamiltonian $\mathcal{K}'$ on the sphere $S^2(L')$. Furthermore, as we shall now see, the phase flow is topologically identical in both cases.

5. EQUILIBRIA

The equilibria of the reduced QZE are the extrema of $\mathcal{K}'$ on $S^2(L')$. Introducing the Lagrange multiplier β, we obtain them as the extrema of the function

$$F \equiv F(\zeta_1,\zeta_2,\zeta_3,\beta) = \mathcal{K}' + \beta(\zeta_1^2+\zeta_2^2+\zeta_3^2)$$

satisfying the constraint in (9). In other words, we solve the system made of the four equations

$$0 = \frac{\partial F}{\partial \zeta_1} = 2(4+\beta)\zeta_1, \qquad 0 = \frac{\partial F}{\partial \zeta_3} = 2\beta\zeta_3,$$

$$0 = \frac{\partial F}{\partial \zeta_2} = 2(\beta-1)\zeta_2, \qquad 0 = \zeta_1^2 + \zeta_2^2 + \zeta_3^2 - L'^2$$

in the four unknowns $\zeta_1, \zeta_2, \zeta_3$ and β. Visibly, there are six equilibria on any sphere $S^2(L')$ when L' is not zero:

	ζ_1	ζ_2	ζ_3	β	K'	g'	G'	
E_0	0	0	L'	0	0		L'	Saddle
E_1	L'	0	0	$\omega/4L'$	$\omega L'$	0	0	Maximum
E_2	0	L'	0	$-\omega/L'$	$-\frac{1}{4}\omega L'$	$\pi/2$	0	Minimum
E_3	$-L'$	0	0	$\omega/4L'$	$\omega L'$	π	0	Maximum
E_4	0	$-L'$	0	$-\omega/L'$	$-\frac{1}{4}\omega L'$	$3\pi/2$	0	Minimum
E_5	0	0	$-L'$	0	0		$-L'$	Saddle

The equilibria E_0 and E_5 respectively at the north and south poles stand for circular orbits in the magnetic meridian plane. The characteristic equation at these points being $\lambda^2 - \omega^2 = 0$, the equilibria are unstable. All other equilibria lie on the equatorial circle, hence they correspond to linear solutions, E_1 and E_3 to the linear solutions perpendicular to the magnetic field lines, E_2 and E_4 to the linear solutions along the magnetic field lines. We have shown in Section 2 that such linear solutions in the reduced system are also solutions of the original system. At E_1 and E_3, the characteristic equation is $\lambda^2 + 5\omega^2 = 0$, at E_2 and E_4, it is $\lambda^2 + 5\omega^2/4 = 0$; hence all four linear equilibria are stable.

6. FIRST ORDER SOLUTIONS

Having located the equilibria and assessed their stability, we now consider the phase flow globally. It is simple enough that it can be characterized in its entirety by producing the complete catalog of solutions outside the equilibria. From (9) and (8) taken as a system of two equations in the

unknowns ζ_1^2 and ζ_2^2, we deduce that

$$\begin{aligned} \zeta_1^2 &= \frac{1}{5}\left(L'^2 + 4\frac{L'}{\omega}\mathcal{K}' - \zeta_3^2\right), \\ \zeta_2^2 &= \frac{4}{5}\left(L'^2 + \frac{L'}{\omega}\mathcal{K}' - \zeta_3^2\right). \end{aligned} \tag{10}$$

But, from the analysis of the equilibria, there appears that, on any sphere $S^2(L')$,

$$-\frac{1}{4}\omega L' \leq \mathcal{K}' \leq \omega L';$$

therefore

$$\begin{aligned} 0 &\leq L'^2 + 4\frac{L'}{\omega}\mathcal{K}' \leq 5L'^2, \\ 0 &\leq L'^2 + \frac{L'}{\omega}\mathcal{K}' \leq \frac{5}{4}L'^2. \end{aligned} \tag{11}$$

In consequence, we introduce the dimensionless quantities $\kappa_1 \geq 0$ and $\kappa_2 \geq 0$ such that

$$\kappa_1^2 = 1 + 4\frac{\mathcal{K}'}{\omega L'}, \qquad \kappa_2^2 = 1 - \frac{\mathcal{K}'}{\omega L'}.$$

According to the inequalities (11),

$$0 \leq \kappa_1^2 \leq 5 \qquad \text{and} \qquad 0 \leq \kappa_2^2 \leq 5/4.$$

In those notations, the solutions in (10) take on the form

$$\zeta_1 = \pm\frac{L'}{\sqrt{5}}\sqrt{\kappa_1^2 - \frac{\zeta_3^2}{L'^2}}, \qquad \zeta_2 = \pm\frac{2L'}{\sqrt{5}}\sqrt{\kappa_2^2 - \frac{\zeta_3^2}{L'^2}}.$$

Hence the resolution of the equations of motions in the global spherical coordinates hinges on the elliptic quadrature

$$\frac{d\zeta_3}{d\tau} = \pm\omega L'\sqrt{\left(\kappa_1^2 - \frac{\zeta_3^2}{L'^2}\right)\left(\kappa_2^2 - \frac{\zeta_3^2}{L'^2}\right)}. \tag{12}$$

Three cases are to be considered:

a) In the interval $0 < \mathcal{K}' < \omega L'$, where $0 < \kappa_2^2 < 1 < \kappa_1^2 < 5$, we introduce the modulus $k = \kappa_2/\kappa_1$ which, by assumption, is < 1, and we set $\zeta_3 = \kappa_2 L' \sin\psi$. In the Jacobian notations for elliptic functions, there follows from (12) that

$$\psi = \mathrm{am}(\upsilon, k) \quad \text{where} \quad \upsilon = \kappa_2\omega(\tau - \tau_0),$$

so that

$$\zeta_1 = \pm\frac{\kappa_1}{\sqrt{5}}\,L'\mathrm{dn}(\upsilon,k),\quad \zeta_2 = \frac{2\kappa_2}{\sqrt{5}}\,L'\mathrm{cn}(\upsilon,k),\quad \zeta_3 = \kappa_2 L'\mathrm{sn}(\upsilon,k). \qquad (13)$$

In this class of solutions, the phase flow induces a conterclockwise circulation about either E_1 or E_3. Geometrically speaking, the perigee librates about either the equilibrium E_1 or E_3 depending on whether we take ζ_1 with the upper or lower sign. Astronomers view these orbits as resulting from a continuous deformation of what they call an osculating ellipse. At $\upsilon = 0$, this curve is a linear orbit making an angle $g' = \arctan 2k$ with the magnetic field line. As υ increases, the linear orbit opens into an ellipse whose perigee moves away from the direction of the magnetic field lines; when υ passes through the quarter-period $K(k)$, the perigee is perpendicular to the magnetic field, and the osculating ellipse reaches its maximum eccentricity $e' = \sqrt{1-\kappa_2}$. Past the quarter-period, the ellipse flattens, the perigee tends to re-align itself with the magnetic field. Eventually, at the half-period $\upsilon = 2K(k)$, the ellipse degenerates again into a linear orbit ($e' = 0$) with a perigee at minimum inclination over the magnetic field. Thereafter the evolution will repeat itself in the southern hemisphere on $S^2(L')$.

b) The solution in the interval $-\omega L'/4 < \mathcal{K}' < 0$, where $0 < \kappa_1^2 < 1 < \kappa_2^2 < 5/4$, derives from the preceding one by Jacobi's transformation by *reciprocal modulus*. By assumption, the inverse modulus $k_1 = 1/k = \kappa_1/\kappa_2$ is < 1, and the solution is given by the equations

$$\zeta_1 = \frac{\kappa_1}{\sqrt{5}}\,L'\mathrm{cn}(\upsilon',k_1),\quad \zeta_2 = \pm\frac{2\kappa_2}{\sqrt{5}}\,L'\mathrm{dn}(\upsilon',k_1),\quad \zeta_3 = \kappa_1 L'\mathrm{sn}(\upsilon',k_1), \qquad (14)$$

the amplitude being now the function $\upsilon' = \kappa_1\omega(\tau - \tau_0)$. In this case, the phase flow circulates clockwise about the equilibria E_2 and E_4 while the perigee librates about either E_2 or E_4 depending on whether the upper or lower sign is adopted for ζ_2.

c) When $\mathcal{K}' = 0$, then $\kappa_1 = \kappa_2 = 1$. Besides the equilibria at the north and south poles, the equations of motion admit admit four solutions asymptotic to these equilibria. To define them in a concise manner, we set $\upsilon = \omega(\tau - \tau_0)$ and we introduce the argument ψ such that

$$\tfrac{1}{2}(\psi + \pi/2) = \mathrm{lambda}(\upsilon) = \arctan e^{\upsilon}.$$

In those notations, the four asymptotic solutions are:

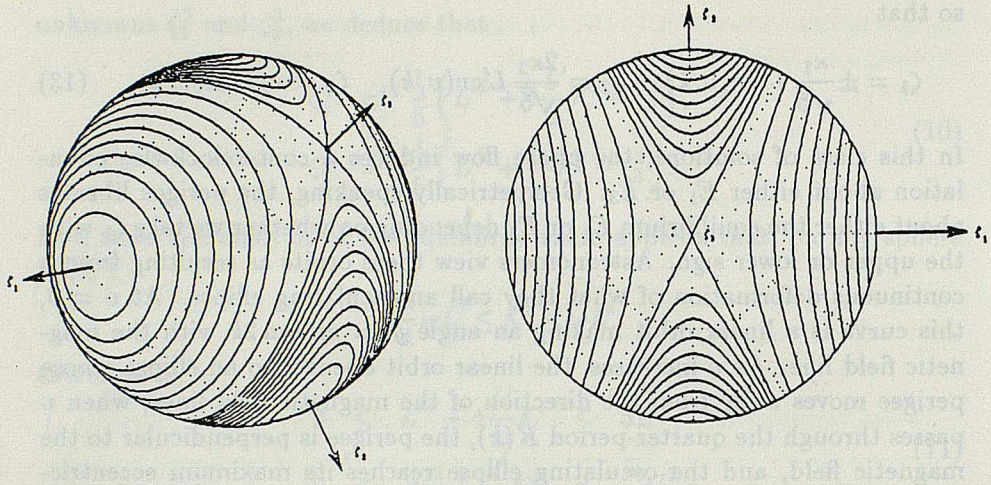


Figure 1: Two orthographic projections of the orbital sphere for the polar quadratic Zeeman effect after reduction

	ζ_1	ζ_2	ζ_3
I_1	$(L'/\sqrt{5})\cos\psi$	$(2L'/\sqrt{5})\cos\psi$	$L'\sin\psi$
I_2	$-(L'/\sqrt{5})\cos\psi$	$(2L'/\sqrt{5})\cos\psi$	$-L'\sin\psi$
I_3	$-(L'/\sqrt{5})\cos\psi$	$-(2L'/\sqrt{5})\cos\psi$	$L'\sin\psi$
I_4	$(L'/\sqrt{5})\cos\psi$	$-(2L'/\sqrt{5})\cos\psi$	$-L'\sin\psi$

The solutions I_1 and I_3 having detached themselves slowly from the equilibrium E_5 tend toward the northern equilibrium E_0 as τ tends to $+\infty$. The solutions I_2 and I_4 do the reverse, going asymptotically to E_0 and E_5 as τ tends to $-\infty$ and $+\infty$ respectively.

The figures below are orthographic projections of an orbital sphere $S^2(L')$ on which are plotted a few solutions as given either by (13) and (14).

CONCLUSIONS

The flow of a perturbed Keplerian system on the orbital sphere $S^2(L')$ after reduction is susceptible of many forms; it may present degeneracies in the form of manifolds of non isolated equilibria; it could transit through

parametric bifurcations from one regime to another. In the face of so wide a diversity of possible behaviors, we are glad to report that the QZE in its manifolds of polar orbits behaves simply like a rigid body in torque-free rotation about its center of mass. On the ellipsoid of given angular momentum in the space of the three components of the angular velocity, the level contours of the kinetic energy are analogous to those of the reduced QZE on its orbital sphere. Permanent rotations about the axis of smallest and largest inertia correspond to the linear equilibria of the reduced QZE, and the permanent rotations about the axis of intermediate inertia to the circular equilibria in the QZE. In both problems the unstable equilibria are located at the same level of energy, hence in the QZE as in the Euler-Poinsot problem, the homoclinic solutions joining asymptotically these singular solutions.

ACKNOWLEDGEMENTS

Partial support for this research came in the form of a Giuseppe Colombo fellowship which the European Space Agency granted S. Ferrer to spend the academic year 1986–1987 as a guest scientist in the Center for Applied Mathematics at the U. S. National Bureau of Standards. Partial support came also from Comisión Interministerial Científica y Técnica of Spain (PB87-0637).

REFERENCES

Coffey, S., Deprit, A., Miller, B. R. and Williams, C. A.: 1986, An. N. Y. Ac. Sc. **497**, 22–36.
Cushman, R.: 1984, in *Differential Methods in Mathematical Physics*, ed. S. Sternberg, D. Reidel Publishing Company, 125–144.
Delande, D. and Gay J. C.: 1986, Pys. Rev. Lett. **57**, 2006– 2009.
Deprit, A.: 1982, Celest. Mech. **26**, 9–21.
Deprit, A.: 1983, Celest. Mech. **29**, 229–248.
Harada, A. and Hasegawa, H.: 1983, J. Phys. A: Math. Gen. **16**, L259–L263.
Marsden, J. and Weinstein, A.: 1974, Rep. Math. Phys. **5**, 121– 130.
Meyer, K. R.: 1973, in *Dynamical Systems*, ed. M. M. Peixoto, Academic Press, New York, 259–272.
Reinhardt, W. P. and Farrelly, D.: 1982, J. Physique **43**, C29– C41.
Richards, D.: 1983, J. Phys. B: At. Mol. Phys. **16**, 749–765.
Robnik, M.: 1981, J. Phys. A: Math. Gen. **14**, 3195–3216.
Robnik, M.: 1984, J. Phys. A: Math. Gen. **17**, 109–130.
Robnik, M. and Schrüfer, E.: 1985, J. Phys. A: Math. Gen. **18**, L853–L859.
Saini, S. and Farrelly, D.: 1987, Phys. Rev. A **36**, 3556–3574.
Solov'ev, E. A.: 1982, Sov. Phys. JETP **55**, 1017–1022.
Zimmerman, M. L., Kash, M. M. and Kleppner, D.: 1980, Phys. Rev. Lett **45**, 1092–1094.

Rev. Acad. Ciencias. Zaragoza 45, (1990)

Inversion of seismic wave velocities by means of the stochastic inverse operator

J. Badal

Department of Theoretical Physics (Geophysics), University of Zaragoza, Plaza de San Francisco, 50009 Zaragoza, Spain

SUMMARY

We expose in detail the problem of inversion of seismic wave velocities which is related to the general problem of inversion of geophysical data. The problem is solved (based on the variational formulation for Love and Rayleigh waves) in the framework of the *generalized inverse theory*, using the *stochastic inverse operator*. Special attention is given to two aspects which are closely linked to the calculation of a particular solution of the problem: the error and the resolution achieved. In this manner, an analysis of variance as well as of the accuracy achieved relative to the solution, is carried out. After describing the calculation procedure, we apply the theory to invert dispersion data of Rayleigh waves, consisting of average interstation velocities. These velocities were previously determined using spectral methods and digital filtering techniques. The results obtained allow us to propose theoretical 2-D earth models for the structure of the lower crust and the upper mantle and also show some features of the Iberian lithosphere-asthenosphere system.

1. Formulation of the problem.

After having determined the dispersion characteristics of the long period seismic waves, one must determine a sufficiently realistic earth model capable of explaining the observations. In other words: from the velocity data obtained after application of adequate digital filtering techniques, we must *invert* these data or solve the corresponding *inverse problem*, i. e., determine a structural model through the calculation of certain parameters of the model like the mass density and the velocities of the P and S body waves, which describe the elastic structure of the medium. The problem has no unique solution, but it is possible to determine, with certain guarantees, a sufficiently acceptable model from a geophysical point of view (Backus and Gilbert, 1970).

The discrete model is defined by a finite number of plane-horizontal, homogeneous layers overlying a half-space. Each of these layers is specified by its density ρ and by the transmission velocities of the P and S waves, α and β respectively. The layers are also defined by their thickness. In principle, the problem consists in determining the relation between a set of observations or experimental data (phase or group velocities of surface waves) and a set of real values corresponding to the parameters of the model. Often the number of parameters in the model is reduced by making use of empirical relations, some of which are well known, as Rayleigh waves are in general more sensitive to changes in shear constants than to other structural parameters (Bloch, Hales and Landisman, 1969). Each layer of the earth model is specified here by its thickness, and by the velocity of the shear wave, β. In our inversion method, the thicknesses of the layers are fixed at the start, and the β values for the layers are estimated in an iterative manner, from the dispersion data.

Let **Y** be the vector of data or observations and **X** be the vector of the structural model; the problem in matrix form can be formulated as follows:

$$\mathbf{Y} = \mathbf{F}\,\mathbf{X} \qquad (1.1)$$

where **F** is a matrix which represents the operator which relates both vectors. The dimension of **Y** need not be equal to the dimension of **X**, and thus the matrix **F** need not be square. If dim **Y** > dim **X**, as occurs in our case, then the problem is *overdetermined*. We are interested in determining **X** from **Y**. If $\mathbf{F}^{-1}$ exists, then:

$$X = F^{-1} Y$$

This equation poses a typical data inversion problem (Wiggins, 197 The problem presented in this form is very important as it expres the possibility of investigating the interior of the Earth usi dispersion measurements of surface waves registered at seismologi stations.

In the case where the problem is linear, where the dimensions of and X are equal, then the problem is trivial and easily solved usi elementary matrix algebra. In the linear case, if (1.1) is not exac true, then this equation becomes

$$Y = A X + e$$

where $e = AX-Y$ is the residual vector. The solution obtained afi minimizing the sum of squares of the residuals, i. e., after applying t condition

$$e^T e \rightarrow \text{minimum}$$

where the symbol T indicates transpose, is (Jackson, 1972)

$$X = (A^T A)^{-1} A^T Y$$

But if the problem posed is not linear, then it might be impossible obtain a solution of the same type $X = F^{-1}Y$. In this case, we try linearize the problem making use of a Taylor expansion to first ord (the first derivative) for an initial or starting earth model. Let X_o I an initial model; then

$$Y = Y_o + A (X - X_o)$$

where

$$Y_o = F (X_o)$$

is the solution obtained solving the corresponding direct problem t any classical method, and

$$A = (A_{jk}) = (\partial F_j/\partial x_{ok})$$

is the Jacobian of partial derivatives. This equation can be written the form

$$\Delta Y = A \Delta X$$

where $Y-Y_o = \Delta Y$, $X-X_o = \Delta X$. The above thus represents a first order approximation to the problem posed in (1.1). If the problem is overdetermined, then the solution is

$$\Delta X = (A^T A)^{-1} A^T \Delta Y$$

As far as velocity data inversion is concerned, if we consider the phase velocity c, for example, then small perturbations in the velocities β_j, which characterize the initial model, correspond to fluctuations in the velocity of the fundamental mode at different frequencies or periods according to the approximately linear relation:

$$\delta c_j = \sum_{i=1}^{N} \frac{\partial c_j}{\partial \beta_i} \delta\beta_i = \frac{\partial c_j}{\partial \beta_1} \delta\beta_1 + \ldots + \frac{\partial c_j}{\partial \beta_N} \delta\beta_N \qquad (1.2)$$

In the above equation, N is the number of layers in the model, δc_j is the difference between the observed phase velocity and the theoretical phase velocity obtained from the model, corresponding to the period j, $\partial c_j/\partial\beta_i$ is the first term in the Taylor expansion of the velocity with regard to the proposed model, and $\delta\beta_i$ (which has to be determined) is the change in the parameter β_i, which is the β-velocity for the i^{th} layer of the model. If **d** is a vector in the data space with M components, composed of velocity differences, and **m** is a vector in the model space with N components, which contains the small variations in velocity (unknowns) defining the initial model, equation (1.2) can be written as

$$\mathbf{d} = \mathbf{G}\,\mathbf{m} \qquad (1.3)$$

where **G** is the M×N matrix of partial derivatives $G_{ji} = \partial c_j/\partial\beta_i$. The above equation is only valid as a first order approximation to our problem. Similarly, if the problem is overdetermined, then the solution is

$$\mathbf{m} = (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{d} \qquad (1.4)$$

Equations (1.3) and (1.4) permit an iterative computation process, namely: The theoretical velocities are calculated by solving the direct problem for the initial model. We can now construct the data vector **d** and also construct the matrix **G** of partial derivatives. Once the solution, i. e., the model vector **m**, is obtained, the starting earth model can be defined again, since thereafter the earth model can be redefined substituting β_i by $\beta_i + \delta\beta_i$, i = 1, 2, ..., N, for the S wave velocity in the i^{th} elastic layer. The procedure is restarted, but now

with this new model, and the solution can be obtained by sucesive approximations. The procedure is repeated until the differences between observed and theoretical velocities are really small, i. e., until the model converges to a solution which reduces the residuals to a satisfactory minimum. It should be noted that the above procedure requires that not only the data vector but also the matrix of partial derivatives be calculated for each iteration.

We need not limit ourselves to the inversion of the phase velocity; the group velocity can also be inverted. Moreover, phase and group velocities can be inverted jointly. Obviously, in joint inversion of dispersion data, the dimension of the problem increases, but the increase in information leads to a more realistic earth model. It should be noted that in the joint inversion case, the matrix of partial derivatives not only contains the derivatives with respect to the model parameters of the phase velocity, but also the derivatives of the group velocity.

The partial derivatives of the phase velocity can be calculated by the method proposed by Takeuchi, Dorman and Saito (1964), based on the variational formulation for Love and Rayleigh waves (Aki and Richards, 1980). The partial derivatives of the group velocity can be calculated from the phase velocity derivatives via a quick and precise method introduced by Rodi *et al.* (1975), based on the well known relationship between the group velocity and the phase velocity.

In the same manner, one may not only invert fundamental mode dispersion data, but also higher mode dispersion data, and Love or Rayleigh wave velocities. As would be expected, the difficulty in solving the problem increases notably when dealing with and having to make compatible a greater volume of information. Often, consistency and convergence problems arise in the search for a solution.

In our inversion scheme of Rayleigh waves, we only consider the model parameter β; other model parameters like ρ and α are not considered since Rayleigh waves are more sensitive to the shear wave velocity. So, the only significant variations occur in β. However, rigorously, equation (1.2) could be rewritten more completely by including terms or contributions due to similar small changes in ρ and α, such that the dimension of the model vector would be greater by a factor of 3.

2. Solution of the inverse problem.

The problem we have to solve is that described by equation (1.3)

$$\mathbf{d} = \mathbf{G}\,\mathbf{m}$$

in order to determine β-velocity as a function of depth which describes the elastic structure of the medium from surface wave observations, or more specifically, from the dispersion characteristics of the fundamental mode Rayleigh waves. We are dealing thus with the inversion of geophysical data, which is a problem that can be divided into two parts. The first part consists in determining a particular solution

$$\mathbf{m}_p = \mathbf{G}_p^{-1}\,\mathbf{d}$$

where the matrix $\mathbf{G}_p^{-1}$ operates on the data. The second part consists in finding the resolution and the error for this particular solution. The calculation of $\mathbf{G}_p^{-1}$ is what is known as *generalized inversion* of the available data, which leads to an earth model related always to known dispersion data.

Substituting equation (1.3), we obtain the following equation

$$\mathbf{m}_p = \mathbf{G}_p^{-1}\,\mathbf{G}\,\mathbf{m}$$

which expresses the particular solution as a weighted average of the true solution, with weights given by row vectors of the matrix $\mathbf{G}_p^{-1}\mathbf{G}$. This weighting matrix is known as the *resolution matrix*.. If $\mathbf{G}_p^{-1}\mathbf{G}$ is the identity matrix I, then the resolution is perfect and the particular solution coincides with the true solution. On the contrary, if the row vectors of $\mathbf{G}_p^{-1}\mathbf{G}$ contain nonzero elements about the principal diagonal, then the particular solution represents a weighted or smoothed solution.

The error $\Delta\mathbf{m}_p$ in the estimation of a particular solution, due to the error $\Delta\mathbf{d}$ in the data, is described by its *covariance matrix*

$$< \Delta\mathbf{m}_p\,\Delta\tilde{\mathbf{m}}_p > = \mathbf{G}_p^{-1} < \Delta\mathbf{d}\,\Delta\tilde{\mathbf{d}} > \tilde{\mathbf{G}}_p^{-1}$$

where ~ indicates conjugate transpose and < > indicates averaging. In this fashion, the covariance matrix of the error in the solution is obtained in terms of the covariance matrix of the error in the data. Once the operator $\mathbf{G}_p^{-1}$ is known for a particular solution, the resolution and the error in the solution can be obtained with relative

ease. However, for the overdetermined case of the problem, we can run into instability problems in the calculation of the *generalized inverse*, i. e., of the operator G_p^{-1}.

To better understand this, it is worth recalling the Lanczos (1961) decomposition, which permits the matrix G to be factorized in the form

$$G = U \Lambda \tilde{V} \tag{2.1}$$

where U is a matrix whose columns are orthogonal eigenvectors (after normalization $\tilde{U}\,U=I$) of $G\tilde{G}$ with nonzero real eigenvalues, V a matrix whose columns are orthogonal eigenvectors (after normalization $\tilde{V}\,V=I$) of $\tilde{G}G$ with nonzero real eigenvalues, and Λ a diagonal matrix whose elements are positive square roots of the nonzero eigenvalues of $\tilde{G}G$ or $G\tilde{G}$ (Aki and Richards, 1980).

Since

$$\tilde{G}G\,m = \tilde{G}d$$

the solution according as Lanczos (1961) theory is

$$m = (\tilde{G}G)^{-1}\tilde{G}d = V\,\Lambda^{-1}\tilde{U}\,d$$

the Lanczos *generalized inverse operator* being

$$G_g^{-1} = (\tilde{G}G)^{-1}\tilde{G} = V\,\Lambda^{-1}\tilde{U} \tag{2.2}$$

Now it is easy to see that some of the instability problems appear as a consequence of Λ^{-1}, when some of the eigenvalues approach zero. In this case, even though a mathematically valid solution can be obtained, it might not be physically admissible, and thus we would need to recur to methods which restrict the instability of the solution. On the other hand, from what has been said, we see that the calculation of the generalized inverse requires an eigenvalue analysis.

A method of avoiding this is due to Franklin (1970), who introduced the *stochastic inverse*, as an alternative form to the generalized inverse (Aki and Richards, 1980). Suppose that the data consist of signal and noise, then the problem to be solved is

$$d = G\,m + n$$

where both **m** and **n** are uncorrelated stochastic processes. If their means are zero and their covariance matrices are

$$< \mathbf{m}\tilde{\mathbf{m}} > = \sigma_m^2 \mathbf{I}$$
$$< \mathbf{n}\tilde{\mathbf{n}} > = \sigma_n^2 \mathbf{I}$$

a good approximation to the generalized inverse is given by the stochastic inverse operator

$$\mathbf{L}_o = \tilde{\mathbf{G}}(\mathbf{G}\tilde{\mathbf{G}} + \varepsilon^2\mathbf{I})^{-1}$$

where

$$\varepsilon^2 = \sigma_n^2/\sigma_m^2$$

In effect, according to the factorization (2.1), in terms of eigenvectors and eigenvalues, we can deduce that

$$\mathbf{L}_o = \mathbf{V}[\Lambda /(\Lambda^2 + \varepsilon^2\mathbf{I})]\,\tilde{\mathbf{U}} \qquad (2.3)$$

and comparing (2.3) with (2.2) it is obvious that $\mathbf{L}_o$ is a good approximation to $\mathbf{G}_g^{-1}$.

The operator $\mathbf{L}_o$ can also be expressed in the form

$$\mathbf{L}_o = (\tilde{\mathbf{G}}\mathbf{G} + \varepsilon^2\mathbf{I})^{-1}\tilde{\mathbf{G}} \qquad (2.4)$$

We will use this solution to invert surface wave velocities. The damping parameter ε^2 plays an important role in that it permits control of the inherent instability of the inversion process. This solution reduces to that due to Lanczos for $\varepsilon = 0$.

Solution (2.4) was already obtained by Lavenberg (1944) by minimizing the sum of squares of the residuals and of squares of the parameters of the model with weighting values which were inversely proportional to their variances, i. e., minimizing $\sigma_n^{-2}|\mathbf{d}-\mathbf{G}\mathbf{m}|^2 + \sigma_m^{-2}|\mathbf{m}|^2$, where $\sigma_n^2/\sigma_m^2 = \varepsilon^2$. Marquardt (1963) also suggested the use of solutions of this type. The same solution was later obtained by Twomey (1977), who proposed a technique based on minimizing a similar quantity $|\mathbf{G}\mathbf{m}-\mathbf{d}|^2 + \gamma|\mathbf{m}|^2$, where γ is a factor which can have an arbitrary value between zero and infinity. The result is the same (with $\gamma = \varepsilon^2$). This solution has been discussed in depth by Jackson (1979), who has studied the use of data or information *a priori* to solve the nonuniqueness of linear inversion.

The resolution matrix, according to (2.3), is

$$L_o\, G = V\, [\Lambda^2 /(\Lambda^2 + \varepsilon^2 I)]\, \tilde{V}$$

and the covariance matrix of the errors in the parameters of the model, as a function of the covariance matrix of the error in the data, is

$$\sigma_d^2\, L_o\, \tilde{L}_o = \sigma_d^2\, V[\Lambda^2 /(\Lambda^2 + \varepsilon^2 I)^2]\, \tilde{V}$$

where σ_d^2 is the variance of the error in the data Δd. In the stochastic model to which we are referring, Δd corresponds to $n = d\text{-}Gm$ and $\sigma_d^2 = \sigma_n^2$. We can now understand how the Lavenberg-Marquardt parameter, ε^2, (Lawson and Hanson, 1974) acts with respect to the solution of the problem, and the error in the estimation of the solution. In effect, ε^2 degrades the resolution, but reduces the error in the solution and stabilizes it, thus reducing the covariance.

In the stochastic inversion scheme, the best choice for the damping constant ε^2 is σ_n^2/σ_m^2, i. e., the ratio of the variance associated with noise to the variance associated with the model parameters. What is certain is that a small ε^2 value could cause the calculation to diverge and a large value cause the convergence to be too slow. The damping stochastic inverse requires an adequate value for the damping constant such that a rapid convergence is achieved in the calculation. We have generally considered $\varepsilon^2 = 30$, although experience has demonstrated the necessity of reducing this value to 10 or even 5, for problems involving joint inversion of phase and group velocities from quality data (Corchete *et al.*, 1990).

3. Error analysis and resolving kernels.

We will now refer to two basic questions: the variances or standard deviations associated with the solution vector and the resolving kernels. The variances are evaluated assuming that the data and its errors are independent random variables. Hence, given the relation

$$x = \sum_i a_i y_i$$

where y_i represents a set of independent random variables, a_i is a constant and x is a random variable, it has been shown (Johnston, 1963) that

$$var(x) = \sum_i a_i^2\, var(y_i)$$

where $var(y_i)$ is the variance of y_i and $var(x)$ the variance of x. Then, if the solution obtained by means of the stochastic inverse operator $\mathbf{L}_o$ is

$$\hat{\mathbf{m}} = \mathbf{L}_o \mathbf{d} \tag{3.1}$$

or

$$m_k = \sum_{l=1}^{M} L_{okl} d_l$$

$$\mathbf{L}_o = (L_{okl}) \quad , \quad k = 1, 2, ..., N$$

and the data vector $\mathbf{d}$ is supposed to be composed of independent random variables, we obtain the expression

$$var(m_k) = \sum_{l=1}^{M} (L_{okl})^2 var(d_l) \tag{3.2}$$

This equation shows the relation between the variance of an element of the solution vector and the respective variances of elements of the data vector, i. e., between the error in the estimation of a parameter of the model and the errors in the observed dispersion data. More precisely, the error in the estimation of a parameter in the model is given by the standard deviation

$$\sigma_k = [var\ (m_k)]^{1/2} \tag{3.3}$$

The matrix resolution or simply the resolution of the problem is nothing more than a measure of the degree of agreement between the true and computed solutions, as was mentioned in section 2. We have

$$\hat{\mathbf{m}} = \mathbf{L}_o \mathbf{G}\ \mathbf{m}$$

and clearly we see that the resolution matrix $\mathbf{L}_o\mathbf{G}$ connects the set of all possible solutions of the problem with the actual model. The equality $\mathbf{L}_o\mathbf{G}=\mathbf{I}$ is never verified, so the particular solution determined is never the true solution. What is obtained is a *smoothed* solution, in the sense that each parameter of the model obtained is a "weighted" mean of the parameters that define the real earth model. The row vectors of the resolution matrix form the so called *resolving kernels*, which are nothing more than the weights involved. The smoothing of the solution is due to the elements of each row vector which occur on either side of the element on the principal diagonal of the weighting matrix. The smaller these elements are, i. e., the values making up the

band on either side of the principal diagonal of the weighting matrix, the more correct the solution determined is, as the matrix approaches the identity matrix I. Similarly, the wider the band of significantly nonzero values, the greater the difference between the determined and true solutions, and consequently the problem will be more poorly resolved.

Usually the resolving kernels are graphically represented as continuous distributions, in our case as a function of depth; each one is represented with respect to a reference depth which corresponds to the model parameter previously fixed. A well resolved problem requires that in each distribution the absolute maximum appear clearly differentiated from other possible relative maxima and also with respect to the reference depth considered, as mentioned before. Really, the resolving kernels are a measure of the exactness achieved in calculating the solution with respect to depth. The lack of coincidence between the absolute maximum of a kernel and the respective reference depth implies poor agreement between the calculated solution and the true solution at that depth, which implies that the validity of the estimated solution is limited. This is obviously important when discussing the results obtained.

4. Computation procedure.

Bearing in mind that we are interested in inverting dispersion data related to the fundamental mode Rayleigh waves, starting from an initial theoretical earth model, presented in terms of a velocity distribution of the shear wave for homogeneous layers of given thickness, then the algorithm for these calculations can be summarized in the following steps:

4.1. Construction of the data vector **d**. To obtain this vector we must solve the motion differential equations numerically for different periods (Aki and Richards, 1980), starting from the initial earth model, and determine the displacement amplitudes and the phase and group velocities of the fundamental Rayleigh mode. The differences between the observed and theoretical dispersion data for the different periods considered, completely define the data vector.

4.2. Construction of the matrix **G**. This requires calculation of the partial derivatives of the phase velocity $\partial c/\partial \beta$ (Takeuchi, Dorman and Saito, 1964) as well as the partial derivatives of the group velocity

$\partial U/\partial\beta$ (Rodi *et al.*, 1975), for each period and each layer of the model, considering the information obtained in 4.1. Thereafter, the program used plots the dispersion curves for the phase and group velocities such that the theoretical and observed dispersion curves may be compared. The partial derivatives calculated in this step for each reference period are also plotted.

4.3. Construction of an *improved model.* This is the main step in the inversion process. Here we need to calculate a particular (smoothed) solution, $\hat{\mathbf{m}}$, to the problem, via equation (3.1); however, to do so, the stochastic inverse operator L_o must first be determined using equation (2.4), considering a reasonable value for the damping constant, which is chosen after considering the resultant model for each trial value. Then, from the initial earth model and the solution $\hat{\mathbf{m}}$, we can construct an improved model.

4.4. Analysis of results. Having reached this point, where we now have an improved model, with the variances, ec. (3.2), or standard deviations, ec. (3.3), of each and every parameter in the model, and having obtained some resolving kernels at various reference depths, we can opt to end the calculation process or continue the process, but starting now with the improved model as initial model.

Whether the calculation process should continue in an iterative manner is determined by the values (positive or negative) in the data vector, i. e., if the theoretical velocity dispersion curves differ substantially from the observed velocity dispersion curves, then obviously the calculation should continue; however, if they coincide sufficiently to within the standard deviation band, then obviously the calculation can be ended. Also, comparison of these curves with each iteration gives a good indication of the convergence of the process.

If finally we opt to end the calculation process, the program used allows us to obtain the resolving kernels for the reference depths of interest, as well as the final improved model. Thus, on the one hand, we obtain a final solution or structure of the medium, and on the other hand an accuracy indication or confidence limits of the solution found. Figure 1 shows a flow-chart of the velocity inversion process.

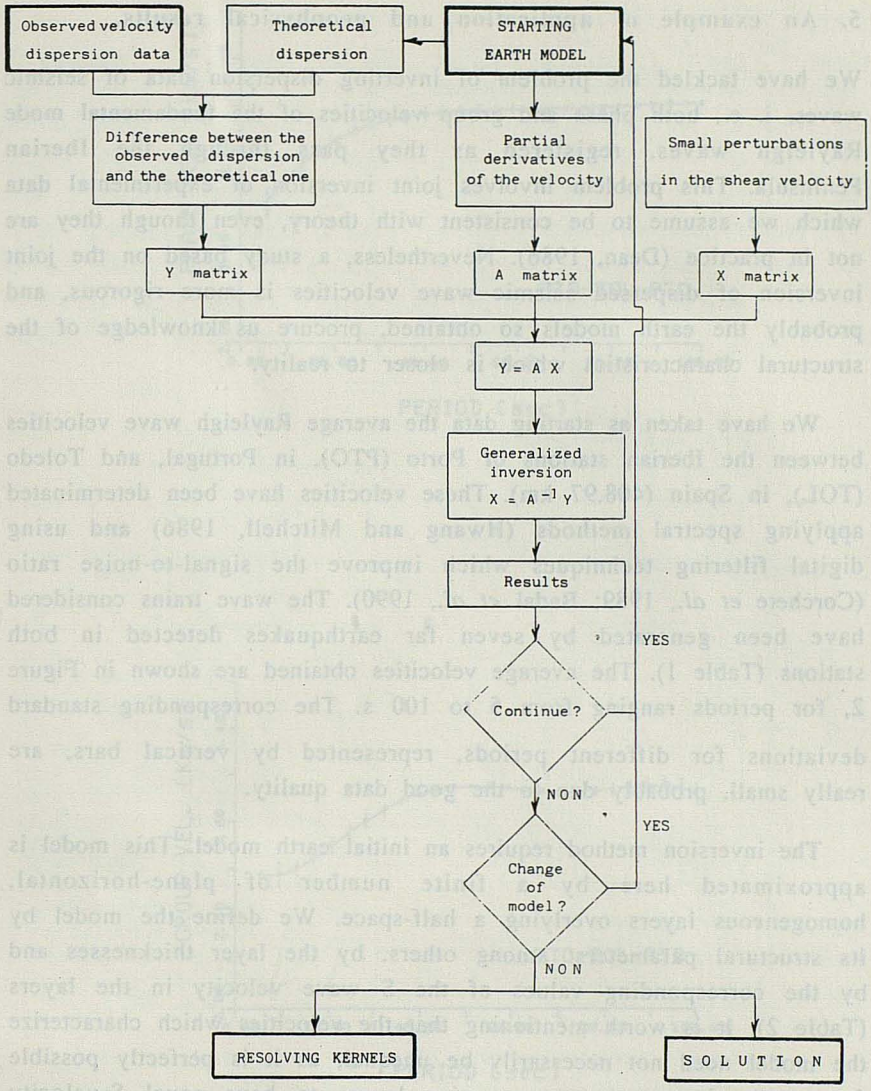


Fig. 1. Flow-chart of the inversion process.

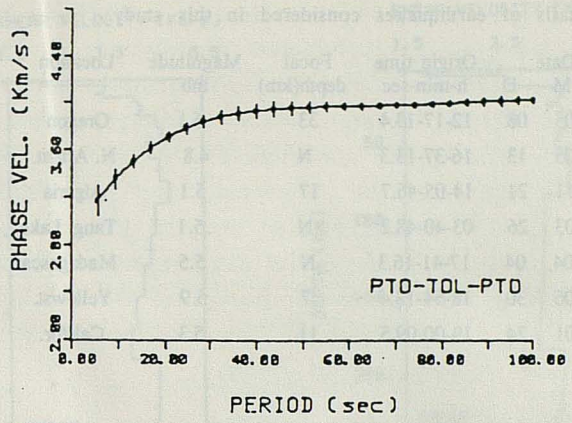
5. An example of application and geophysical results.

We have tackled the problem of inverting dispersion data of seismic waves, i. e., both phase and group velocities of the fundamental mode Rayleigh waves, registered as they pass through the Iberian Peninsula. This problem involves joint inversion of experimental data which we assume to be consistent with theory, even though they are not in practice (Dean, 1986). Nevertheless, a study based on the joint inversion of dispersed seismic wave velocities is more rigorous, and probably the earth models so obtained, procure us knowledge of the structural characteristics which is closer to reality.

We have taken as starting data the average Rayleigh wave velocities between the Iberian stations of Porto (PTO), in Portugal, and Toledo (TOL), in Spain (408.97 km). These velocities have been determinated applying spectral methods (Hwang and Mitchell, 1986) and using digital filtering techniques which improve the signal-to-noise ratio (Corchete *et al.*, 1989; Badal *et al.*, 1990). The wave trains considered have been generated by seven far earthquakes detected in both stations (Table 1). The average velocities obtained are shown in Figure 2, for periods ranging from 5 to 100 s. The corresponding standard deviations for different periods, represented by vertical bars, are really small, probably due to the good data quality.

The inversion method requires an initial earth model. This model is approximated here by a finite number of plane-horizontal, homogeneous layers overlying a half-space. We define the model by its structural parameters: among others, by the layer thicknesses and by the corresponding values of the S wave velocity in the layers (Table 2). It is worth mentioning that the velocities which characterize the model need not necessarily be unequal, as it is perfectly possible for consecutive or non consecutive layers to have equal S-velocity values. As far as the velocities are concerned, the inversion process, which is an iterative process, will gradually modify the model till it converges to a final structural model representative of the medium through which the waves have travelled.

The dimension or number of degrees of freedom of the problem is fixed *a priori* when the maximum number of layers constituing the model has been decided. This does not mean that the medium need necessarily have that number of layers, as the velocities determined via the generalized inversion for two consecutive layers can be almost



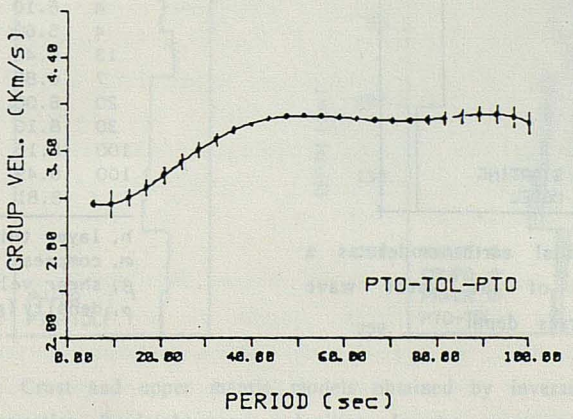


Fig. 2. Average Rayleigh wave velocities for the PTO-TOL seismic path. Vertical bars are standard deviations at various periods. Periods range from 5 to 100 seconds.

Table 1. Details of earthquakes considered in this study.

Event No.	Date Y	M	D	Origin time h-min-sec	Focal depth(km)	Magnitude mb	Location	Epicentre Lat.	Long.
1	1968	05	08	12-17-13.4	33	6.1	Oregon	43.6N	127.9W
2	1972	05	13	16-37-13.3	N	4.8	N. Atlant.	45.0N	28.1W
3	1973	11	24	14-05-46.7	17	5.1	Algeria	36.1N	4.4E
4	1975	03	26	03-40-48.2	N	5.1	Tang. Lake	5.4S	30.2E
5	1975	04	04	17-41-16.3	N	5.5	Madagascar	21.2S	45.1E
6	1975	06	30	18-54-13.4	7	5.9	Yellowst.	44.7N	110.6W
7	1980	01	24	19-00-09.5	11	5.3	Califor.	37.8N	121.8W

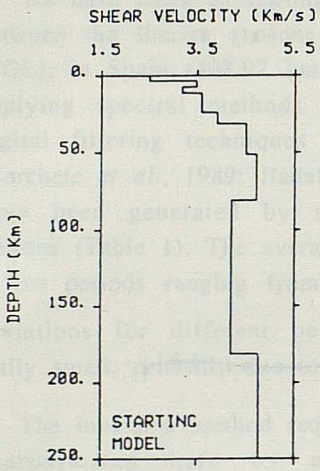


Fig. 3. Initial earth model as a distribution of the shear wave velocity versus depth.

Table 2

Parameters defining the initial earth model

h	α	β	ρ
3	3.30	2.50	2.28
4	6.10	3.48	2.79
4	5.60	3.18	2.74
13	6.40	3.58	2.80
7	6.85	3.90	3.05
20	8.00	4.50	3.20
30	8.10	4.70	3.35
100	8.15	4.20	3.40
100	8.49	4.77	3.53
	8.81	4.89	3.60

h, layer thickness (km)
α, compres. velocity (km/s)
β, shear velocity (km/s)
ρ, density (g/cc)

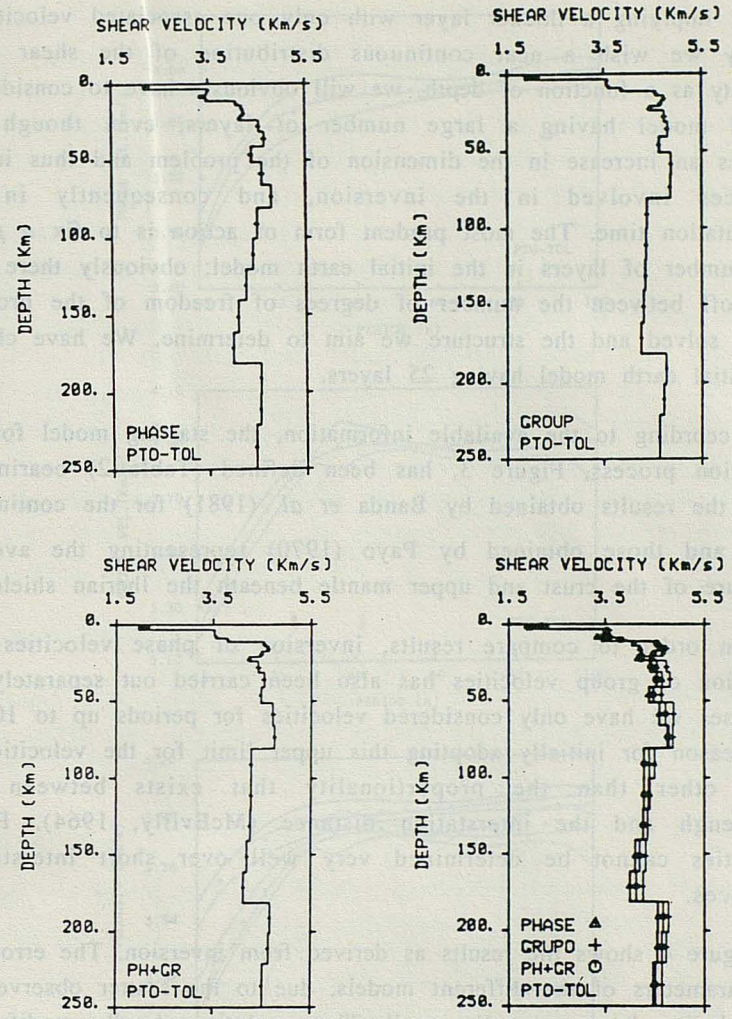


Fig. 4. Crust and upper mantle models obtained by inversion of interstation Rayleigh wave velocities. In the upper part, results obtained by inversion of phase velocities and group velocities. In the lower part, results obtained by joint inversion of phase and group velocities (PH+GR), along with a superposition of all those models. Errors in shear velocity are very small.

equal, implying a thicker layer with only one associated velocity. If finally we wish a near continuous distribution of the shear wave velocity as a function of depth, we will obviously have to consider an initial model having a large number of layers, even though this implies an increase in the dimension of the problem and thus in the matrices involved in the inversion, and consequently in the computation time. The most prudent form of action is to fix *a priori* the number of layers in the initial earth model; obviously there is a trade-off between the number of degrees of freedom of the problem to be solved and the structure we aim to determine. We have chosen an initial earth model having 25 layers.

According to the available information, the starting model for the inversion process, Figure 3, has been defined (Table 2) bearing in mind the results obtained by Banda *et al.* (1981) for the continental crust and those obtained by Payo (1970) representing the average structure of the crust and upper mantle beneath the Iberian shield.

In order to compare results, inversion of phase velocities and inversion of group velocities has also been carried out separately. In all cases we have only considered velocities for periods up to 100 s. The reason for initially adopting this upper limit for the velocities is none other than the proportionality that exists between the wavelengh and the interstation distance (McEvilly, 1964). Phase velocities cannot be determined very well over short interstation distances.

Figure 4 shows the results as derived from inversion. The errors in the parameters of the different models, due to the scatter observed in the velocity data, are really small. These solutions hardly modify the characteristics of the structure of the continental crust. Our results do not affect the top-most layers of the initial earth model, as we work with surface waves having periods greater than 10 s. This introduces an initial constraint obeying the wavelengh-depth relationship, affecting the resolution power of our operations, and in particular the precise determination of S-velocity at crustal depths. The essential conclusions reached refer to new features of the Iberian lithosphere-asthenosphere system: firstly, a thin low-velocity layer, 6 km thick, in the lower crust; secondly, another low-velocity layer, about 20 km thick, beneath the Moho; thirdly, a clear low-velocity zone in the peninsular asthenosphere, approximately 100 km thick, having a negative velocity gradient.

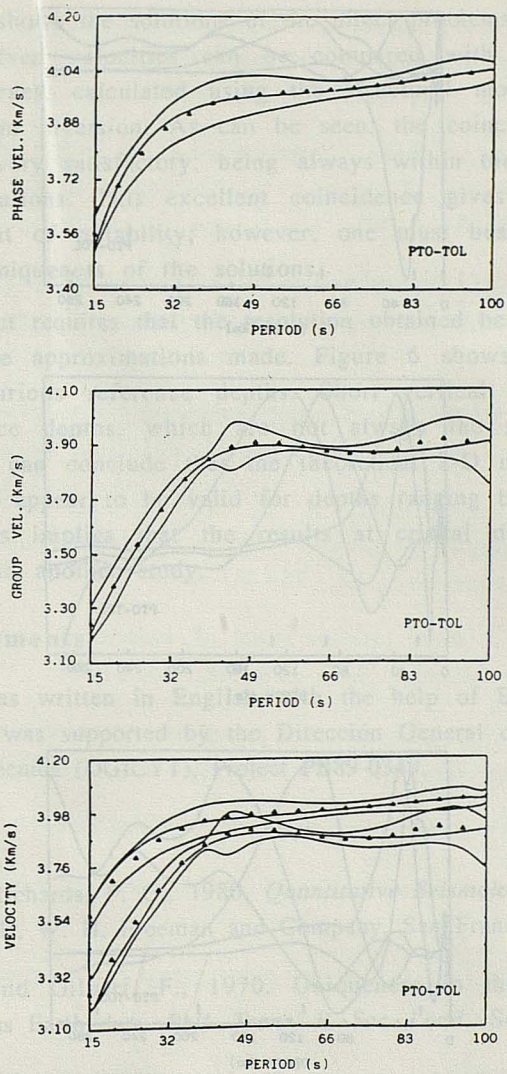


Fig. 5. A comparison between the respective average observed velocities (triangles) and the velocities calculated by solving the corresponding direct problem (central continuous line) from the adequate PH+GR earth model. In each case, upper and lower solid lines point out the limits of the respective standard deviations.

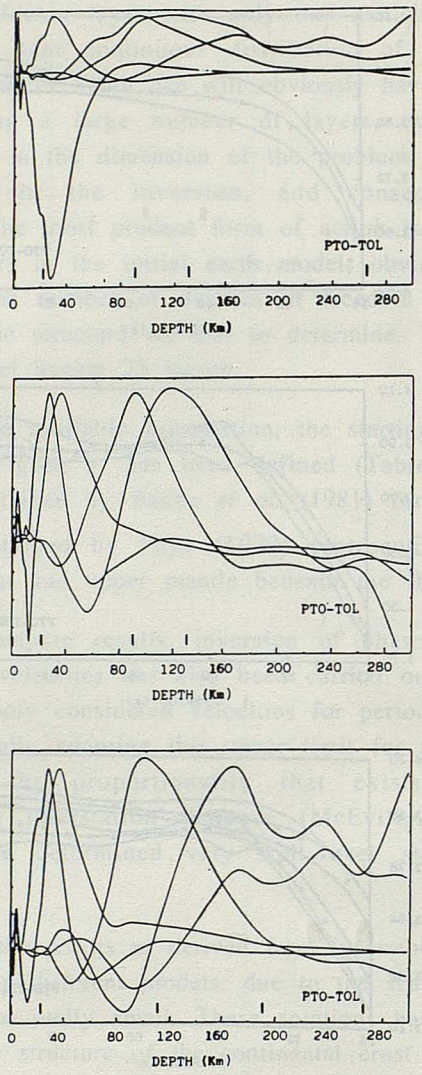


Fig. 6. Resolving kernels at various depths and solutions as derived by simple inversion of phase velocities (upper part), of group velocities (central part), and by inversion of both phase and group velocities (lower part). Relative maxima of these functions should occur at the reference depths indicated by short vertical marks in the lower part of each diagram.

Figure 5 shows the solutions of the direct problems, such that the average observed velocities can be compared with the respective dispersion curves, calculated using the structural model determined before by joint inversion. As can be seen, the coincidence between velocities is very satisfactory, being always within the limits of the standard deviations. This excellent coincidence gives the results a certain amount of reliability; however, one must bear in mind the intrinsic nonuniqueness of the solutions.

This aspect requires that the resolution obtained be examined with respect to the approximations made. Figure 6 shows the resolving kernels at various reference depths. Short vertical marks indicate these reference depths, which are not always the same for each problem. We can conclude that the theoretical 2-D models that we have obtained appear to be valid for depths ranging between 25 and 200 km. This implies that the results at crustal depths must be reconsidered in another study.

Acknowledgments

This paper was written in English with the help of Ezequiel Gurría. This research was supported by the Dirección General de Investigación Científica y Técnica (DGICYT), Project PB89-0349.

References

Aki, K. and Richards, P. G., 1980. *Quantitative Seismology. Theory and Methods*, vol. II, W. H. Freeman and Company, San Francisco.

Backus, G. and Gilbert, F., 1970. Uniqueness in the inversion of inaccurate gross Earth data. *Phil. Trans. R. Soc. Lond.*, Ser. A **266**, 123-192.

Badal, J., Corchete, V., Payo, G., Canas, J. A., Pujades, L. and Serón, F. J., 1990. Processing and inversion of long-period surface-wave data collected in the Iberian Peninsula. *Geophys. J. Int.*, **100**, 193-202.

Banda, E., Suriñach, E., Aparicio, A., Sierra, J. and Ruiz de la Parte, E., 1981. Crust and upper mantle structure of the central Iberian Meseta (Spain). *Geophys. J. R. astr. Soc.*, **67**, 779-789.

Bloch, S., Hales, A. L. and Landisman, M., 1969. Velocities in the crust and upper mantle of southern Africa from multi-mode surface-wave dispersion. *Bull. seism. Soc. Am.*, **59**, 1599-1629.

Corchete, V., Badal, J., Payo, G., Canas, J. A., Pujades, L. and Serón, F. J., 1990. An attempt of joint inversion of Rayleigh wave phase and group velocities in Iberia. *Rev. Geofísica*, **46**, 83-96.

Corchete, V., Badal, J., Payo, G. and Serón, F. J., 1989. Filtrado de ondas sísmicas dispersadas (Filtering of dispersed seismic waves). *Rev. Geofísica*, **45**, 39-58.

Dean, E. A., 1986. The simultaneous smoothing of phase and group velocities from multi-event surface wave data. *Bull. seism. Soc. Am.*, **76**, 1367-1383.

Franklin, J. N., 1970. Well-posed stochastic extensions of ill-posed linear problems. *J. Math. Anal. Appl.*, **31**, 681-716.

Hwang, H. J. and Mitchell, B. J., 1986. Interstation surface wave analysis by frequency-domain Wiener deconvolution and modal isolation. *Bull. seism. Soc. Am.*, **76**, 847-864.

Jackson, D. D., 1972. Interpretation of inaccurate, insufficient and inconsistent data. *Geophys. J. R. astr. Soc.*, **28**, 97-109.

Jackson, D. D., 1979. The use of *a priori* data to resolve non-uniqueness in linear inversion. *Geophys. J. R. astr. Soc.*, **57**, 137-157.

Johnston, J., 1963. Econometric Methods. McGraw-Hill. New York.

Lanczos, C., 1961. Linear Differential Operators. D. Van Nostrand Co. London.

Lawson, Ch. L. and Hanson, R. J., 1974. Solving Least Squares Problems. Prentice-Hall. New York.

Lavenberg, K., 1944. A method for the solution of certain nonlinear problems in least squares. *Quart. Appl. Math.*, **2**, 164-168.

Marquardt, D. W., 1963. An algorithm for least squares estimation of nonlinear parameters. *SIAM J.*, **11**, 431-441.

McEvilly, T. V., 1964. Central U. S. crust-upper mantle structure from Love and Rayleigh wave phase velocity inversion. *Bull. seism. Soc. Am.*, **54**, 1997-2015.

Payo, G., 1970. Structure of the crust and upper mantle in the Iberian Shield by means of a long period triangular array. *Geophys. J. R. astr. Soc.*, **20**, 493-508.

Rodi, W. L., Glover, P., Li, T. M. C. and Alexander, S. S., 1975. A fast, accurate method for computing group-velocity partial derivatives for Rayleigh and Love modes. *Bull. seism. Soc. Am.*, **65**, 1105-1114.

Takeuchi, H., Dorman, J. and Saito, M., 1964. Partial derivatives of surface wave phase velocity with respect to physical parameters changes within the Earth. *J. geophys. Res.*, **69**, 3429-3441.

Twomey, S., 1977. Introduction to the mathematics of inversion in remote sensing and indirect measurements. Elsevier Scientific Publishing Company. Amsterdam.

Wiggins, R. A., 1972. The general linear inverse problem: Implications of surface waves and free oscillations for earth structure. *Rev. Geophys. Space Phys.*, **10**, 251-285.

Rev. Acad. Ciencias. Zaragoza 45, (1990)

TELEDETECCION: FUNDAMENTOS Y APLICACIONES

M.A. SORIANO. Departamento de Geología. Universidad de Zaragoza. 50009 Zaragoza.

ABSTRACT.- Remote sensing techniques are becoming more and more important for the last years. Today multispectral sensors are employed and they can register a wide variety of wavelengths in the electromagnetic spectrum. The spectral signature of the different canopies is obtained. Remote sensing is an important tool for a great variety of sciences such as Topography, Archaeology, Oceanography, Geomorphology, Geology, Biology, Environmental studies,... In Geology it is possible determining the lithology and structure of vast regions, which allow the study of wide areas with the important saving of money in mineral purposes, for instance. The possibility of monitorizing the current most dynamic environments (glacial, fluvial, coastal,...) is also very important. Some works addressed to establish the relationships among vegetation, climate and productivity of the plants, carried out at the University of Maryland are shown.

1. INTRODUCCION

Bajo el término de Teledetección se consideran aquellos métodos que obtienen información de los objetos sin tener contacto físico con ellos. De modo más restringido solo se incluyen los que utilizan la energía electromagnética. Las primeras imágenes fotográficas de la superficie de la Tierra se obtuvieron desde un globo aerostático a mediados del siglo pasado. Pero este tipo de técnica no se utlilizó de forma más sistemática hasta la I y II guerras mundiales, durante las cuales se comenzó a usar películas de color y de infrarrojos con fines militares. Sin embargo, debido sobre todo a las imágenes sinópticas que se obtienen, el paso más importante se produjo con las primeras fotografías obtenidas de la Tierra desde el espacio en 1960.

Pese a que solo proporcionan información de la porción visible del espectro electromagnético, las fotografías aéreas son un instrumento muy útil en campos tan variados como la Geología, Biología, Geografía, Meteorología, etc.. Por ello se pensó en extender la observación a otras longitudes de onda del espectro electromagnético con objeto de obtener más información. Como consecuencia, comenzaron a usarse en los satélites, fundamentalmente, los sensores multiespectrales que registran datos en varios intervalos de longitudes de onda. Esto ocurrió a partir, sobre todo, del lanzamiento en 1972 del Landsat I. Al mismo tiempo que se registraban un gran número de datos, se intentaba que ocupasen el menor espacio posible, lo que provocó el desarrollo de sensores cada vez más complejos que digitalizan las imágenes. Este hecho trajo como cómo consecuencia el desarrollo paralelo del procesado de estos datos para transformarlos de nuevo en imágenes. Consecuentemente, tanto los programas como los ordenadores utilizados experimentaron un gran avance. Desde entonces se ha continuado con el

lanzamiento de nuevos satélites Landsat y recientemente (en 1986), el satélite francés SPOT. Por otra parte, el número de satélites meteorológicos, de comunicaciones, astronómicos, geodésicos, etc. ha aumentado de un modo espectacular en los últimos años. En un futuro (a mediados de esta década) está previsto que los nuevos satélites lleven sensores hiperespectrales (SCHOTT, 1989). Estos podrán registrar 100 o más bandas reflectivas, un número mucho mayor que los que se utilizan hoy en día (el satélite Landsat registra 6 bandas mediante su sistema TM). Debido a las enormes cantidades de datos que recogerán estos nuevos sensores se necesitan nuevas técnicas de análisis de imágenes y ordenadores para el procesado de éstas mucho más potentes que los actuales.

La importancia de la utilización de la teledetección por satélite reside en varios factores tales como que proporciona un recubrimiento periódico de la superficie de la Tierra, ofrece una visión sinóptica de la misma, su formato digital que permite el tratamiento de la imagen, se tiene homogeneidad en la toma de datos y se registra información en regiones de espectro electromagnético que no son la zona del visible (CHUVIECO, 1990).

En este trabajo se pretende exponer cuales son las características físicas más importantes en que se basan los estudios de teledetección, así como comentar algunas de las aplicaciones que se están efectuando desde hace ya largo tiempo, haciendo un especial hincapié en Geología y Geomorfología y en la experiencia adquirida por la autora en el Laboratory for Remote Sensing de la Universidad de Maryland.

2. CARACTERISTICAS FUNDAMENTALES.

Como es conocido, la energía electromagnética se mueve con la velocidad de la luz en un modelo de ondas armónicas. Sólo puede detectarse cuando interactúa con la materia ya que entonces se liberan fotones. Estos pueden ser registrados mediante aparatos especiales (sensores) que miden la cantidad de energía electromagnética para un determinado rango de frecuencia y que la transforman en impulsos eléctricos proporcionales al número de fotones recibidos. Hay gran variedad de tipos de sensores pero los más utilizados por los satélites son los scanner que emplean un detector con un campo de vista reducido que va barriendo el terreno para obtener finalmente una imagen. Hay varios clases de scanner pero sólo mencionaremos dos de ellos, el *cross-track* y el *along-track* que son los que portan los satélites Landsat y SPOT, respectivamente (tabla 1). El primero de ellos consta de un espejo que gira y va barriendo a través del terreno mediante líneas paralelas orientadas perpendicularmente a la dirección de vuelo. En el segundo hay una fila de detectores que recogen cada línea a la vez. Estos sistemas registran una imagen sencilla que representa una banda espectral. Los satélites portan sistemas multiespectrales, colocando varios detectores en los scanner, de tal manera que cada uno registra un intervalo de longitud de onda (tabla 1).

Sátelite	Sensor	Banda	λ (mm)	Resolución
Landsat(MSS)	Cross-track	4 o 1	0.5-0.6	76 m
	o	5 o 2	0.6-0.7	
	Whisk broom	6 o 3	0.7-0.8	
		7 o 4	0.8-1.1	
Landsat(TM)	Cross-track	1	0.45-0.52	30 m
	o	2	0.52-0.60	
	Whisk broom	3	0.63-0.69	
		4	0.76-0.90	
		5	1.55-1.75	
		7	2.08-2.35	
		6	10.4-12.5	120 m
SPOT	Along-track	1	0.50-0.59	20 m
	o	2	0.61-0.69	
	Push-broom	3	0.79-0.90	
		pancromática	0.50-0.90	10 m

Tabla 1.- Características fundamentales de los sistemas de imágenes de los satélites Landsat y SPOT.

El total de energía recibida por el sensor dependerá no sólo de la cantidad de energía incidente, sino también de la interacción que ésta ha sufrido con la atmósfera y con la superficie de la Tierra. De esta manera, la atmósfera puede transmitir, dispersar (reflejar en todas direcciones) y absorber todas o alguna de las longitudes de onda de la radiación incidente. En cuanto a la superficie de la Tierra, dependiendo de sus características, también producirá los efectos anteriores y, además de todo ello, hay que tener en cuenta la orientación que presente dicha superficie, la época del año y la hora del día en que se toma la imagen ya que el albedo será distinto. También dependen de estas circunstancias las sombras que aparecen en la imagen, ya que dependen de la posición relativa del Sol con respecto a la Tierra. Como ejemplo basta señalar que el rango de longitudes de onda que se utiliza en teledetección es limitado debido a la absorción atmosférica que ejercen varios gases (CO_2 , O_2 , O_3 ,...) sobre la radiación electromagnética. En teledetección se registra energía de las regiones de microondas, infrarrojo (con intervalos reducidos a causa de la absorción), visible y las longitudes de onda más largas de la región del ultravioleta.

La reflectancia espectral de los objetos que hay en la superficie de la Tierra se define como

$$R_\lambda = (E_r / E_i) \times 100$$

donde E_r es la Energía de longitud de onda λ reflejada y E_i es la energía incidente. En las imágenes viene dada por la intensidad del tono gris, de tal manera, que el tono estará más próximo al blanco cuanto mayor sea el valor de la reflectancia, y más próximo al negro cuanto menor. R_λ varía para un mismo objeto según la longitud de onda que se utilice (característica espectral). De hecho, en algunos casos estos cambios son tan típicos que permiten reconocer inmediatamente la cubierta que se está analizando tal como puede verse en la figura 1. En líneas generales, la disposición de estas curvas se mantiene. Sin embargo, tenemos que considerar que

la reflectancia de un objeto (por ejemplo una roca o suelo) no suele ser "pura", ya que en la mayor parte de las zonas están recubiertos por distintos tipos de vegetación, las rocas están alteradas, etc.. Todo esto produce una distorsión de la señal que recibe el sensor, llegando la influencia de la cobertera vegetal a ser tan importante que si un 50% de roca está recubierto por vegetales, la curva de reflectancia que se obtiene es tipo a la que corresponde con la vegetación y no con la roca en cuestión.

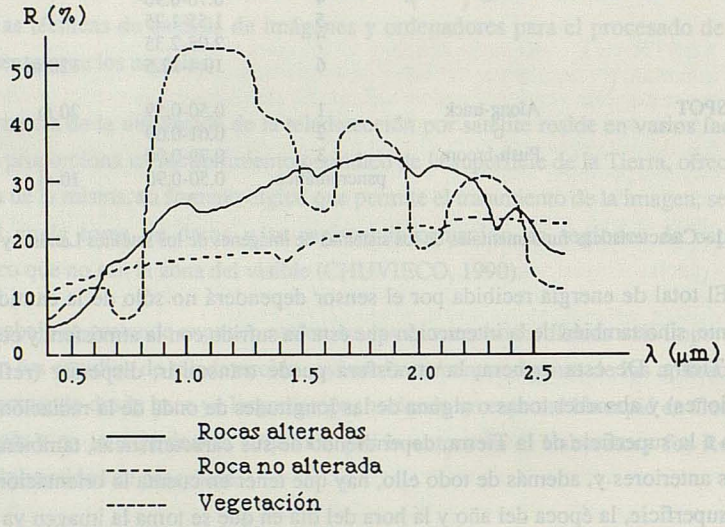


Figura 1.- Curvas de reflectancia espectral para rocas alteradas hidrotermalmente, rocas no alteradas y vegetación (según SABINS, 1986)

A partir de la información espectral, se realizan clasificaciones mediante técnicas de reconocimiento de modelos estadísticos (Figura 2). La brillantez de cada pixel en las bandas espectrales viene dada por un punto. Si se representan gráficamente dos o tres de esas bandas, se obtendrán nubes de puntos. Los que pertenezcan a una determinada clase o tipo de cubierta tenderán a agruparse formando elipses debido a que presentan características radiométricas similares. De esta manera pueden distinguirse diversos tipos de cubiertas de la Tierra.

Sin embargo, normalmente las imágenes originales que se obtienen deben ser mejoradas debido sobre todo a que las condiciones atmosféricas influyen en su nitidez. También hay que realzarlas por las propias condiciones de los sensores, pues éstos están preparados para registrar todos los niveles de gris y en una imagen es casi imposible que se encuentren desde el negro absoluto hasta el blanco más intenso. Además también hay que restaurarlas cuando se producen pequeños fallos mecánicos en los sensores. Todo ello es posible realizarlo mediante los distintos programas de procesado de imágenes que existen en el mercado.

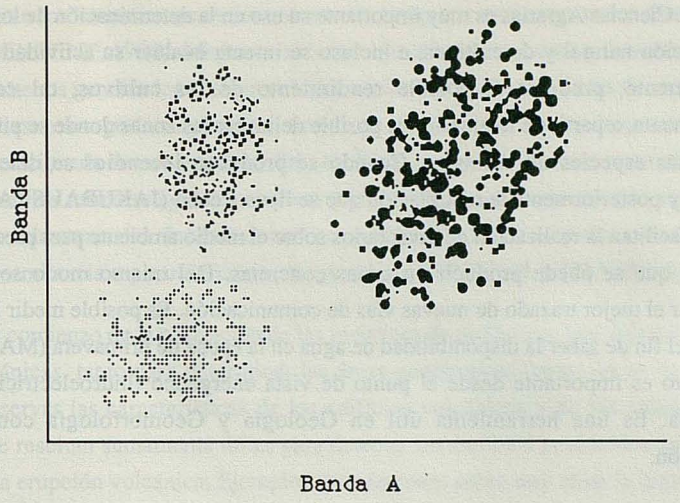


Figura 2: Nubes de puntos correspondientes a reflectancia de pixels en dos bandas distintas

3. APLICACION DE LAS IMAGENES DE SATELITES.

Debido a las características que tienen las imágenes obtenidas mediante satélite, y que ya se señalaron en la Introducción, su estudio se ha aplicado en campos muy diversos . Aunque no pretendemos dar una enumeración exhaustiva de todos ellos, sí nombraremos algunos que consideramos más interesantes.

Una de las principales aplicaciones de la teledetección es la Cartografía. Se está procediendo al levantamiento topográfico de nuevos mapas a distintas escalas especialmente a partir de las imágenes del satélite SPOT. La calidad de la cartografía obtenida a escala 1:50.000 es superior a la de los mapas topográficos ya existentes de la misma zona (GRABMAIER et al., 1988). En Arquelología se emplea alguno de los sistemas de Teledetección (radar) para encontrar posibles yacimientos en zonas con climatología adversa o donde hay mucha vegetación, ya que en estos casos la fotografía aérea tradicional no sirve. En Meteorología se usa para hacer prediciones del tiempo a un corto plazo, también para estudiar los tipos de nubes, variaciones de temperatura, etc.. De igual modo tiene una gran importancia en estudios de la Hidrosfera, en especial del mar, ya que se pueden determinar factores como su temperatura, turbidez, salinidad, contenido en materia orgánica, altura de las olas, determinación de batimetrías en zonas costeras o lacustres, direcciones de corrientes, determinación de movimientos de icebergs en sonas polares, etc. A gran escala es posible también su empleo para realizar modelos de distribución de poblaciones y de estimación del número de habitantes. En

Biología y Ciencias Agrarias es muy importante su uso en la determinación de los distintos tipos de vegetación natural y de cultivos, e incluso se intenta evaluar su actividad fotosintética e indirectamente, predecir el posible rendimiento de los cultivos, tal como veremos. Indirectamente, a partir de lo anterior es posible delimitar las zonas donde se pueden encontrar las distintas especies de animales. Cuando se producen incendios se determina bien su extensión y posteriormente la repoblación que se lleva a cabo (JAKUBAUSKAS et al., 1990). También facilitan la realización de inventarios sobre el medio ambiente para precisar el impacto ambiental que se puede producir en zonas concretas. Del mismo modo se emplean para seleccionar el mejor trazado de nuevas vías de comunicación. Es posible medir la acumulación nival con el fin de saber la disponibilidad de agua en la época de primavera (MACIAS y SOLE, 1988). Esto es importante desde el punto de vista enérgetico (hidroelectricidad) y para la agricultura. Es una herramienta útil en Geología y Geomorfología como veremos a continuación.

3.1. *Aplicaciones en Geología y Geomorfología*

La cartografía es un apoyo imprescindible para ciertas ciencias, como por ejemplo pueden ser la Gelogía y la Geomorfología, que necesita plasmar sus datos o resultados en mapas, representaciones gráficas de la superficie estudiada, de ahí la importancia que adquiere la teledetección en estas ciencias, puesto que las imágenes de satélite permiten la realización de gran cantidad y variedad de mapas con aplicación a las mismas. Mediante los distintos tipos de sistemas que se utilizan en Teledetección, se puede realizar la cartografía geológica de zonas sobre las que existan pocos datos o bien sean de difícil acceso (ANANABA y AJAKAIYE,1987; WESTERHOF et al, 1990; RABIE y AMMAR, 1990). Hay que señalar de forma especial la gran utilidad que tiene el radar en el estudio de zonas tropicales donde se presentan dos problemas fundamentales que este tipo de sensor solventa. Uno es la elevada precipitación que se registra en estas zonas (el radar registra imágenes aunque exista una masa nubosa). Otro es que la presencia de una densa vegetación dificulta el reconocimiento de estructuras tectónicas con casi todos los sistemas de Teledetección. Sin embargo, con el radar las estructuras quedan muy realzadas con lo que pueden reconocerse fácilmente.

En Geología se realizan las cartografías de los distintos tipos de rocas; si están alteradas hidrotermalmente, las estructuras que presentan (en especial los lineamientos), etc. De esta manera se conoce de forma general las características de la zona. Pero además, es muy importante en prospección minera ya que determinar las zonas más propicias para hallar yacimientos minerales (como zonas de lineamientos y de rocas alteradas) implica un ahorro considerable de dinero en la búsqueda de los mismos (EOSAT, 1987). Para determinar esta serie de características se utilizan las imágenes multiespectrales, microondas (radar) e infrarrojo térmico. Con las primeras, mediante combinaciones de distintas bandas es posible determinar el contenido en hierro, carbonatos, minerales silicatados, arcillosos, etc que tienen las rocas. Con

el radar se realzan los lineamientos mejor que con las imágenes multiespectrales (hasta un 27% como señala VINCENT, 1980). Mediante el infrarrojo térmico se pueden detectar zonas que presenten fenómenos geotérmicos (áreas volcánicas, fuentes calientes, zonas de emanación de gases). Se realizan cartografías de inercia termal, propiedad que en materiales secos se halla directamente relacionada con su densidad. De esta forma, rocas con elevado contenido en sílice tienen altas inercias termales (granito, riolita) y las pobres en sílice y elevado contenido en hierro bajas inercias termales. Incluso dentro es posible diferenciar entre las rocas silicatadas atendiendo a la emitancia espectral mínima que presentan (VINCENT y THOMSON, 1971).

Desde el comienzo de la utilización de las imágenes de satélites se han aplicado al estudio de zonas volcánicas, tanto para determinar las áreas geotérmicas (como ya se ha indicado), como para observar las características de los edificios volcánicos y de las coladas de lava. Evidentemente resultan sumamente útiles para detectar los cambios producidos en el entorno despues de una erupción volcánica. Ejemplos muy famosos sobre este tema lo constituyen los estudios realizados en Islandia, islas Hawaii y en el Monte Santa Elena (LO, 1986).

En Geomorfología se están utilizando las imágenes obtenidas mediante satélites para distintos fines. Ya hemos mencionado la importancia de la cartografía Geomorfológica, especialmente útil en zonas de difícil acceso. El detalle que se obtiene en la realización de estas cartografías es muy elevado debido a la alta resolución que ofrecen los sensores de los satélites Landsat y SPOT, sobre todo de este último ya que además tiene la posibilidad de analizar estereoscópicamente las imágenes. La introducción en los sensores multiespectrales de bandas que comprendían intervalos de longitudes de onda correspondientes al infrarrojo medio implica que es posible reconocer depósitos superficiales ricos en hierro, hidróxidos y minerales arcillosos, que tienen gran importancia desde el punto de vista geomorfológico y económico.

Sin embargo, la principal utilidad que ofrece a la Geomorfología la Teledetección por satélite consiste en poder comprobar el cambio que experimentan algunos modelados de medios especialmente dinámicos (MILLINGTON y TOWNSHEND, 1986). Este hecho es posible gracias a la frecuencia con que los satélites recubren una misma superficie. Un ejemplo de ello es el estudio de los fenómenos volcánicos que ya mencionamos. Es especialmente útil para determinar las variaciones de las zonas de erosión-depósito que se están produciendo en las áreas costeras y de desembocaduras de las ríos. La principal ventaja que ofrecen al estudiar este tipo de medios, reside en su visión sinóptica, ya que la imagen abarca una extensión grande de terreno lo que, en el caso de estudios que se realizan en zonas afectadas por la marea, es muy importante, pues es conveniente que toda la zona estudiada esté en las mismas condiciones de marea, cosa que no se puede conseguir mediante los fotogramas aéreos tradicionales (BERNAL et al., 1986). Además, los canales del infrarrojo permiten diferenciar entre los medios agua, fango y tierra seca. Son también muy efectivos en la determinación de variaciones en los cauces de los ríos y sobre todo para determinar posibles alteraciones que se producen como

consecuencia de inundaciones (ARBIOL et al., 1984 y YAMAGATA y AKIYAMA, 1988). Esta aplicación es muy importante sobre todo en el caso de ríos muy caudalosos y de gran longitud en los que tarda varios días en producirse la crecida. Sin embargo, incluso en ríos donde ésta dura tan solo unas horas (como es el caso de los de la Península Ibérica) se han estudiado imágenes de un par de meses posteriores a la "riada" y se ha comprobado que se podía limitar bastante bien la zona afectada por ella (ARBIOL et al., 1984). De la misma forma se utiliza en Glaciología para determinar las variaciones de los glaciares y de zonas cubiertas por hielo, en especial en los medios costeros debido al desprendimiento de los icebergs. Incluso en medios que, aparentemente, pudieran considerarse poco dinámicos como los ambientes de playa-lake, se ha apreciado grandes variaciones en el periodo de meses, como consecuencia de lluvias esporádicas (TOWNSHEND, et al., 1989) A escala más detallada, se han realizado trabajos para intentar seguir la evolución de la erosión en áreas de badlands situadas en zonas montañosas, que se degradan sobre todo cuando se producen lluvias torrenciales de baja frecuencia, utilizando imágenes TM. El uso de esta técnica sería muy útil porque facilita los estudios temporales comparados y disminuye los costos de seguimento de la evolución de estos modelados. A partir de zonas estudiadas con gran detalle se pueden aplicar los resultados a otras con características similares probando si el modelo también es válido en estas últimas (SOLE, et al., 1986).

3.2. *Vegetación, clima, productividad. Relaciones*

En este apartado comentaremos brevemente otra aplicación de la telecetección, la investigación sobre vegetación utilizando imágenes de satélite, llevada a cabo por un grupo de investigadores de la Universidad de Maryland, donde la autora de este artículo pudo familiarizarse con el software y técnicas allí empleados. El equipo de investigación del Laboratory for Remote Sensing de la Universidad de Maryland utiliza las imágenes de satélite para el estudio de la vegetación. Este grupo, principalmente, emplea un programa denominado Image System elaborado en el N.A.S.A. Goddard Space Flight Center y modificado por ellos para adaptarlo a uno de los ordenadores de que disponen (un HP 1000 A600 con 4 Mb de memoria interna y dos discos duros de 300 y 80 Mb). Todo ello se complementa con una lectora de cintas y varias pantallas y un procesador de imagen Ramtek de alta resolución.

De manera muy concisa, podemos indicar que las líneas de investigación que ha seguido este grupo de trabajo ha sido establecer las relaciones que existen entre vegetación, clima y productividad de dicha vegetación (siendo éste último el más complejo) a partir de las imágenes de satélite empleando, de manera especial, el satélite meteorológico NOAA, si bien en ocasiones se pueden completar con informaciones de otros, como el LANDSAT. De cualquier modo, la Teledetección está desarrollándose rápidamente e incluso se pueden esperar mayores avances en el nuevo campo de la "bioclimatología de satélite" en un futuro (GOWARD, 1989) A continuación explicamos con algo más de detalle como se llevan a cabo estos trabajos.

Los parámetros climáticos que más directamente están relacionados con la vegetación son la temperatura y el albedo. Mediante las imágenes multiespectrales se obtiene información del estatus foliar. Este depende a su vez de la humedad, temperatura del aire y determina el carácter del cambio de energía entre la tierra y la atmósfera. En general, a mayor vegetación la temperatura de la superficie es más baja (GOWARD et al., 1985a). La relación con el albedo no es tan simple, ya que éste depende también del tipo de vegetación y del sustrato en el que está.

La radiación solar incidente determina la radiación absorbida por las plantas y ésta a su vez influye en la fotosíntesis y transpiración de los vegetales. Como puede verse en la figura 1, la vegetación absorbe mucha radiación en el espectro visible, y poco o nada en el infrarrojo, siendo estas medidas específicas para especies y lugares. Por ello se han utilizado los denominados índices de vegetación en los que se relacionan datos de los canales visible e infrarrojo. Los valores mayores de estos índices se producen cuando aumenta la cantidad de vegetación verde en las zonas observadas. Los índices de vegetación se correlacionan de forma no linear con la biomasa y con los índices de área de hoja verde y linearmente con la radiación activa fotosintéticamente interceptada. Por lo tanto, mediante estos parámetros se puede conocer la productividad de las plantas. Con estos índices se obtiene la vegetación fotosintética activa que cambia según la estación del año considerada. A partir de observaciones periódicas se ve que hay una relación directa entre el valor integrado del índice de vegetación espectral y la acumulación estacional de biomasa.

Todo esto parece sugerir que las observaciones periódicas en las longitudes de onda del visible y del infrarrojo pueden usarse para caracterizar el estatus actual dinámico estacional y magnitud estacional integrada de la actividad fotosintética de la vegetación (GOWARD et al., 1985b). Partiendo de estas ideas y mediante la utilización de imágenes obtenidas por el satélite NOAA que utiliza el sensor AVHRR (advanced very high resolution radiometer) GOWARD et al., (1985b) estudian dos aspectos en Norteamérica, la variabilidad estacional y el área integrada definida por la variación temporal de las medidas sobre la estación de crecimiento. Se aprecia claramente como los valores más elevados de los índices de vegetación se desplazan hacia el norte en la primavera y verano y derivan hacia el sur en el otoño. Del mismo modo, el comportamiento de los índices de vegetación a lo largo del año es distinto si se trata de vegetación natural o bien de áreas cultivadas. En cuanto a las medidas integradas generalmente disminuyen al norte y oeste a traves de norteamérica pero hay más heterogeneidades en el tercio oeste del continente. En algunos lugares aparecen anomalías: valores mayores y menores de lo que cabría esperar. En ambos casos se debe a la influencia de la agricultura (por ejemplo cuando se riega).

Otra línea de trabajo que se ha desarrollado por este equipo es el estudio de la diferenciación de los líquenes de otras plantas superiores mediante el uso de teledetección. Este interés se debe a que éstos son muy abundadntes en latitudes superiores a los 50 °. El espectro

de los líquenes es distinto al de la vegetación típica, ya que presenta mayor reflectancia en la zona del espectro visible que la vegetación normal con lo que no hay un salto tan fuerte hecia el infrarojo próximo (PETZOLD y GOWARD, 1988). Esto permite que a partir de las observaciones multiespectrales se puedan realizar estudios de los ecosistemas árticos y subárticos ya que éstos son los organismos más abundantes en estas zonas. Variaciones en las condiciones climáticas de estos medios se manifestarán rápidamente en los líquenes.

AGRADECIMIENTOS

Al Dr. Goward por las facilidades puestas a disposición de la autora durante su estancia en la Universidad de Maryland. Este trabajo ha sido realizado gracias a la ayuda financiera de la Dirección General de Investigación Ciéntifica y Técnica (BE90-070) que sufragó dicha estancia.

BIBLIOGRAFIA

ANANABA, S.E. y AJAKAIYE, D.E. (1987) Evidence of tectonic control of mineralization in Nigeria from lineament density analysis: a Landsat study. International Journal Remote Sensing vol 8, pp. 1445-1453.

ARBIOL, R.; CALVET, J. y VIÑAS, O. (1984) Detección por el satélite LANDSAT-4 de los efectos de la riadad del 8-XI-82 en el río Segre. Acta Geológica Hispánica t. 19 pp. 235-248.

BERNAL, E.; GUILLEMOT, E. y THOMAS, I.F. (1986) Contribución del sensor TM al conocimiento de las costas oceánicas aluviales: dos ejemplos en el litoral atlántico andaluz. I Reunión científica del Grupo de trabajo en Teledetección. pp.157-178.

CHUVIECO, E. (1990) Fundamentos de Teledetección espacial 453 p. Ed. Rialp

EOSAT (1987) Lisbon valley, Utah. Thematic mapper hydrocarbon study. Eosat. Landsat application notes, vol.2, 4p.

GOWARD, S.N. (1989) Satellite Bioclimatology Journal of Climate vol 2, pp. 710-720.

GOWARD, S.N.; CRUICKSHANKS, G.D. y HOPE, A.S. (1985a) Observed relation between thermal emission and reflected spectral radiance of a complex vegetated landscape. Remote Sensing of Environment 18, pp. 137-146.

GOWARD, S.N.; TUCKER, C.J. y DYE, D.G. (1985b) North American vegetation patterns observed with the NOAA-7 advanced very high resolution radiometer. Vegetatio 64, pp. 3-14.

GRABMAIER, K.; TULADHAR, A.M. y VERSTAPPEN, H. Th. (1988) Stereo mapping with SPOT. ITC Journal 1988 pp. 149-154.

JAKUBAUSKAS, M.E.; LULLA, K.P. y MAUSEL, P.W. (1990) Assessment of vegetation change in a fire altered forest landscape. Photogrametric Engineering and Remote Sensing vol. 56, pp. 371-377.

LO, C.P. (1986) Applied Remote Sensing 393 p. Ed. Longman

MACIAS, M. y SOLE, L. (1988) Técnicas de Teledetección. En: La nieve en el Pirineo español M.O.P.U. 178 p.

MILLINGTON, A.C. y TOWNSHEND, J.R.G. (1986) The potential of satellite remote sensing for geomorphological investigations-an overview. En: Gardiner (ed)International Geomorphology parte II, pp. 331-342. John Wiley&Sons.

PETZOLD, D.E. y GOWARD, S.N. (1988) Reflectance spectra of subartic lichens. Remote Sensing of Environment 24, pp. 481-492.

RABIE, S.I. y AMMAR, A.A. (1990) Pattern of the main tectonic trends from Remote Geophysics, geological structures and Satellite Imagery, central eastern desert, Egypt. International Journal Remote Sensing vol 11, pp. 669-683.

SABINS, F.F. (1986) Remote Sensing. Principles and interpretation 449 p. Ed. Freeman.

SCHOTT, J.R. (1989) Remote sensing of the Earth: a synoptic view. Physics Today 1989, pp. 72-79.

SOLE, L.; CLOTET, N.; GALLART, F. y SALA, I. (1986) Análisis de las posibilidades de las imágenes TM en la detección de áreas degradadas en sectores montañosos. I Reunión científica del Grupo de trabajo en Teledetección. pp. 335-363.

TOWNSHEND, J.R.G.; QUARMBY, N.A.; MILLINGTON, A.C.; DRAKE, N.; READING, A.J. y WHITE, K.H. (1989) Monitoring playa sediment transport systems using Thematic mapper data. Advanced Space Research vol 9, pp. 177-183.

VINCENT, R.K.(1980) The use of radar and Landsat data for mineral and petroleum exploration in the los Andes region,Venezuela. En Radar Geology: an assesment . Jet propulsion Laboratory: Pasadena, California, pp. 367-384.

VINCENT, R.K. y THOMSON, F.J. (1971) Discrimination of basic silicate rocks by recognition maps processed from aerial infrared data. En Proceedings of the seventh International Symposium on Remote Sensing of Environment vol 1, University of Michigan. pp. 247-252.

WESTERHOF, A.B.; ALEVA, G.J.J. y DIJKSTRA, S. (1990) Classification of mineral deposits by host rock lithology: extension, updating and fine-tuning. ITC Journal 1990, pp. 102-110.

YAMAGATA, Y. y AKIYAMA, T. (1988) Flood damage analysis using multitemporal Landsat Thematic Mapper data. International Journal Remote Sensing vol. 9, pp.503-514.

Rev. Acad. Ciencias. Zaragoza 45, (1990)

SIGNIFICADO QUIMICO-ENERGETICO DE LAS ZONAS DE FRICCION Y DE LAS ROCAS MILONITICAS EN LA CADENA PIRENAICA. SU RELACION CON EL ENGROSAMIENTO CORTICAL.

SANCHEZ CELA, V., LAPUENTE, M. P., AUQUE, L. F. & GOMEZ, J.

Dptº Geología, Facultad de Ciencias, Universidad de Zaragoza.

In the Pyrenees there are abundant E-W faulting-friction zones, almost always associated with mylonitic rocks. Such structural-dynamic and petrological features, created under high compressional conditions, during Hercynian-Late Hercynian and also Alpine times, are associated to sialic thickening processes.

The studies on these faulting-mylonitic rocks indicate that such dynamic zones were important chemical and thermal sources, mainly of silica-alkaline elements at high-moderate temperatures (650-300º C).

These active chemical elements, to such temperatures, must be taken into account, since they could take part in various petrogenetic processes, from epimetamorphic to igneous environments.

1. INTRODUCCION.

Los Pirineos constituyen una Cadena Orogénica estructurada en dos ciclos orogénicos, hercínico y alpino, caracterizada por un fuerte engrosamiento cortical. Su núcleo hercínico, que aflora en la Zona Axial y Norpirenaica, comprende materiales sedimentarios y epi-mesometamórficos (precámbricos y paleozoicos) estructurados durante varias fases de deformación. Junto a estos materiales existen abundantes macizos graníticos, cuya entidad en el volumen total del edificio orogénico (deducida a partir de datos geofísicos, estructurales y petrológicos) es de primera magnitud.

Desde el punto de vista estructural, la Cadena Pireanica se caracteriza por la existencia de importantes zonas de cizalla y fracturas longitudinales, paralelas a la dirección general E-W de la Cadena, afectando fundamentalmente a su basamento hercínico.

Asociadas a estas fracturas se desarrollaron diversas rocas miloníticas, y ocasionalmente pseudotaquilitas, originadas en diversas fases compresivas hercínicas-tardihercínicas. Muchas de ellas sufrieron procesos de reactivación durante tiempos alpinos.

A pesar de que gran parte de las paragénesis que constituyen estas rocas miloníticas son de carácter retrógrado, se han reconocido algunas asociaciones minerales que indican condiciones de alta temperatura y moderada presión, como prueba de la actividad térmica y bárica desarrollada durante la formación de estas rocas.

Ultimamente se ha empezado a considerar la importancia de estas fracturas con rocas miloníticas, como zonas de transferencia química, a moderada-alta temperatura, de los componentes geoquímicos más móviles en la corteza, como son los elementos sílico-alcalinos.

En este trabajo pretendemos remarcar el interés y la importancia de estas rocas miloníticas, generadas en zonas de fricción, en cuanto ellas representan relictos de fuentes de energía química y térmica que tuvieron lugar en tiempos hercínicos-tardihercínicos y posiblemente también alpinos. Estas fuentes químicas siálicas, que relacionamos con fenómenos de engrosamiento corticales, pudieron haber participado en diversos procesos pretrogenéticos en los niveles más superficiales de la corteza.

2. ENERGIA DE FRICCION.

Desde hace relativamente pocos años, se ha empezado a considerar el fenómeno del calor friccional o "shear heating", producido por el rozamiento entre bloques siálicos asociados a la dinámica de una falla, como una explicación para la existencia de anomalías térmicas o de procesos petrogenéticos, tanto metamórficos como ígneos, en los que su génesis implica una cantidad de energía adicional a la considerada normalmente. Así últimamente se ha puesto de manifiesto la existencia de ciertas manifestaciones energéticas relacionadas con el funcionamiento de fracturas, principalmente en fallas inversas.

En la zona axial pirenaica y en los macizos norpirenaicos existen abundantes zonas de cizalla a las que se asocian rocas miloníticas (Lamouroux, 1976, 1987; Lamouroux et al, 1980-1981; Soula et al, 1986), e incluso pseudotaquilitas (Passchier, 1982 a, b; 1984), de edad hercínica y/o tardihercínica y gran desarrollo longitudinal, relacionadas con fallas inversas en las que se han reconocido saltos estructurales considerables, incluso de varios km. Esto, añadido a su estructuración principal en dirección E-W, (coincidente con la alineación predominante de las rocas "andesíticas" de esta misma edad), nos ha llevado a relacionar la energía térmica y química liberada en las zonas de cizalla con la petrogénesis de estas rocas volcánicas ligadas al ciclo compresivo hercínico. (Sánchez Cela & Lapuente, en prensa)

No es nuestro objetivo cuantificar el calor generado en estas zonas de cizalla, pero sí resulta interesante comentar algunas de las causas que puedan influir en una mayor o menor eficacia del calor desarrollado en ellas.

La importancia de la energía mecánica como fuente de calor ha sido considerada por muchos autores, entre ellos Price (1970), Sibson (1975, 1977, 1978, 1980), Fyfe et al (1988), Scholz (1980), Lachenbruch (1980), Fleitout & Froidevaux (1980), Brewer (1981), Turcotte & Schubert (1982), etc. ya sea aplicado a la dinámica de fallas o al contraste de propiedades térmicas entre zócalo y cobertera, o como Molnar et al (1983) y Jaupart & Provost (1985), a la dinámica de cabalgamientos.

Aunque son relativamente escasos los trabajos experimentales que cuantifican el calor generado en una zona de falla, los existentes han contribuído, en buena forma, a desarrollar los conceptos teóricos que hay detrás de este mecanismo energético. Así, entre otros, pueden destacarse los trabajos de Teufel & Logan (1978), o los de Lockner & Okubo (1983) y Spray (1987), que extrapolan los resultados a condiciones naturales, verificando los modelos de autores anteriores.

En términos generales, la transformación de la energía mecánica en calor se puede expresar de forma distinta según sea el comportamiento mecánico (frágil o dúctil) de la zona de fallla recibiendo la denominación de "frictional heating" y "viscous strain heating", respectivamente. En los niveles relativamente superficiales rige la deformación frágil, sin embargo a grandes profundidades parece que el proceso más favorable es el de deformación plástica sobre materiales reológicamente dúctiles, y aunque los parámetros físicos que controlan un proceso u otro ("frictional" o "strain"), son distintos, el resultado es el mismo: una

disipación de energía química y térmica con acumulación de calor, que en condiciones favorables puede llegar a la fusión parcial.

De todos los factores condicionantes del desarrollo de calor en estas zonas de falla (McKenzie & Brune, 1972), nosotros pensamos que el componente "compresivo" o "stress tectónico" es de capital importancia, analizándose a continuación la influencia que puede tener.

La energía liberada en una zona de fricción puede expresarse, según McKenzie & Brune (1972), como $Q = \tau . d$, en donde τ es la resistencia a la cizalla y d el desplazamiento.

Entendiendo la resistencia a la cizalla como un criterio friccional, análogo al de resistencia al rozamiento en Física elemental, ésta viene definida por la Ley de Amonton (Turcotte & Schubert, 1982), como: $/ \tau / = \mu \sigma_n$, en donde $/\tau /$ es el valor absoluto de la resistencia a la cizalla, μ el coeficiente de fricción en la superficie de contacto y σ_n el esfuerzo normal.

Aplicando esta expresión a un modelo de zona de fractura entre bloques con un comportamiento reológico frágil:

$\sigma_n = P_l + P_S => / \tau / = \mu (P_l + P_S)$, con $P_l = \rho g z$; $P_S = \rho g z + \Delta P_S$, en donde P_l es la presión litostática, P_S es el esfuerzo horizontal total, ρ la densidad media del cuerpo rocoso, g la aceleración de la gravedad, z la profundidad o espesor y ΔP_S el esfuerzo tectónico.

Turcotte & Schubert (1982) siguiendo la teoría de Anderson (1951) sobre la dinámica de fallas, elaboran un razonamiento mediante el cual expresan el valor del esfuerzo tectónico en función de diversos parámetros en condiciones estáticas o estables. Así llegan a la ecuación:

$$\Delta P_S = \frac{2\mu_S (\rho g z - P_f)}{sen\ 2\theta - \mu_S (1 + cos\ 2\theta)}$$

en donde, θ es el ángulo de la traza del plano de falla con la vertical, P_f la presión de fluídos y μ_S el coeficiente de fricción estático.El criterio de signos utilizado es considerar como positivo ΔP_S en fallas inversas y negativo en fallas normales.

Tanto ΔP_S como μ_S pueden llegar a ser muy elevados si consideramos este esfuerzo como una compresión lateral por intrusión de nuevo material entre bloques rígidos previos.

Además de tener en cuenta todos los parámetros anteriormente citados, con los que el proceso guarda una relación directa (densidad, espesor, coeficiente de fricción...), es muy importante la influencia de la presión de fluídos, ya que actúa reduciendo el esfuerzo normal que se desarrolla sobre el plano de falla principal.Los fluídos interaccionan "contrarrestando" la presión litostática y por lo tanto reduciendo el valor de σ_n..

Según esto la expresión general queda de la forma: $\sigma_n = P_l - P_f + P_S$.

Generalmente, la fracturación que acompaña al plano principal de fractura minimiza la presión de fluidos al interconectar los poros y conductos.

De esta forma la ecuación $\tau = \mu\sigma_n$, pasa a ser $\tau = \mu(\sigma_n - P_f)$

El proceso de fricción produce un desprendimiento de energía térmica provocando en la zona de cizalla un aumento de temperatura que conlleva una elevación de la presión de fluídos. El aumento de P_f determina, siguiendo la ecuación anterior, una disminución del valor de τ. Como se ha visto anteriormente, el calor almacenado (Q) está en relación directa con τ, por lo que una elevación de la presión de fluídos puede producir una disminución de la energía térmica desprendida y una disminución de la temperatura en la zona de fricción.

Pero los procesos de fricción en zonas frágiles dependen mucho del tipo de fractura que se desarrolle. Así, Sibson (1977) destaca la importancia que tiene el tipo de falla, para un valor determinado del coeficiente de fricción, sobre los esfuerzos diferenciales aplicados. De esta manera, por ejemplo, para un coeficiente de fricción de 0.75, los esfuerzos diferenciales aplicados son cuatro veces mayores en las fallas inversas que en las normales. Por otro lado, si la presión de fluídos se hace igual a la presión hidrostática, para una profundidad determinada τ tendrá valores superiores en las fallas inversas.

Cuando la fricción tienen lugar sobre fracturas preexistentes, el comportamiento ante la fractura no será el mismo. La mayor disipación de energía térmica se producirá conforme el valor de τ es grande, por lo que ésta tendrá lugar cuando el proceso de fracturación afecte a un cuerpo rocoso sin planos de discontinuidad preexistentes; y al contrario, en las zonas de fractura en las que las superficies de contacto sean preexistentes, el valor de τ necesario para que se produzcan desplazamientos será menor y por tanto la energía térmica desprendida será menor. De ello se deduce que la energía liberada en una zona de fractura polifásica tenderá a ser menor con el tiempo y por ello el momento de máxima conversión de energía mecánica en calor corresponderá al primer episodio de fracturación.

A partir de algunos trabajos experimentales, como el modelo desarrollado por Fleitout & Froidevaux (1980) para un sistema compuesto por dos materiales con comportamientos mecánicos distintos, se ha observado que si el plano de cizalla corta a ambas capas, el calentamiento es mayor en el material competente, y se pueden llegar a alcanzar temperaturas suficientes para producir la fusión del material incompetente en la interfase. Por ello una zona de cizalla que corte una estratificación, podrá ser en potencia una zona de generación de magmas cuando existe un acusado contraste en las propiedades térmicas de los materiales geológicos estratificados.

England & Thompson (1984) recurren al fenómeno de "shear heating" para comprender cómo pueden originarse temperaturas por encima de 650º C a profundidades relativamente someras, que den explicación a la existencia de leucogranitos. De su estudio se deduce que en la interfase basamento "cristalino"-cobertera sedimentaria, puede llegarse a la fusión debido al contraste de conductividad térmica entre un tipo y otro de rocas .

En resumen, varios factores condicionarán el calor generado en una zona de cizalla o fricción:

- las propiedades térmicas y mecánicas de las rocas
- los tipos de fracturas o fallas
- la estructuración geológica
- el régimen sísmico o asísmico
- el esfuerzo tectónico, el tiempo, la velocidad de desplazamiento, el desplazamiento total, la anchura y el espesor de

la zona, la presión de fluídos y la deshidratación de minerales, etc.

Siendo las condiciones más favorables:

- en zonas de alto contraste térmico, con distinta conductividad térmica (interfase basamento-cobertera)
- en fallas inversas
- en zonas que corten la estratificación, o entre tipos de rocas distintas (interfase rocas graníticas-sedimentarias)
- en cuerpos rocosos homogéneos sin fracturación previa (con alto coeficiente de fricción)
- en zonas sometidas a un elevado esfuerzo tectónico y que han sufrido desplazamientos importantes.

3 . ROCAS MILONITICAS.

3.1.Caracteres generales.

La estructura actual del Pirineo está condicionada por la superposición de las deformaciones hercínicas y alpinas. No siempre es fácil distinguir las estructuras estrictamente alpinas de las hércínicas reactivadas y, si a esto añadimos la dificultad de definir cuando termina un ciclo orogénico y comienza otro, queda explicado el recurrir a las manifestaciones "tardihercínicas" para solapar ambas orogénias.

En relación con este episodio tardihercínico se desarrollan gran número de fracturas, bandas miloníticas y zonas de cizalla; algunas de estas fracturas delimitarán las cuencas estefanienses-pérmicas donde se ubican distintos tipos de manifestaciones volcánicas.

A lo largo de toda la zona axial pirenaica, se han reconocido importantes accidentes subverticales cizallados acompañados de bandas miloníticas,* que se alinean siguiendo fracturas profundas de dirección dominante E-W desarrolladas en el basamento frágil pre-Estefaniense en las últimas fases hercínicas, tardihercínicas e incluso alpinas. Estas bandas miloníticas han sido estudiadas en la zona pirenaica, entre otros por Carreras (1975), Carreras y Santanach (1973), Carreras et al (1977), Saillant (1982), y en el sector centro-occidental por Lamouroux (1976, 1987), Lamouroux et al (1979, 1980-1981), Passchier (1982 a,b, 1984, 1985), McCaig (1983, 1984), McCaig & Miller (1986), Soula et al (1986).

* Los distintos tipos de milonitas quedan definidos según Higgins (1971), por el proceso de milonitización que predomine, sea fracturación o sea granulación con recristalización y / o neoformación, y según se hayan desarrollado, o no, estructuras de flujo.

Todas las variedades quedan englobadas bajo el témino de "roca milonítica" cuando no es posible distinguir cual ha sido el tipo de deformación (frágil o dúctil) de los materiales, según el acuerdo de la Conferencia de Penrose (Tullis et al, 1982).

Atendiendo a su orientación con respecto a la Cadena, se han diferenciado dos familias principales de accidentes, cuyas características específicas son también distintas:

- zonas de accidentes longitudinales mayores, con dirección E-W.

- zonas de accidentes menores, NW - SE, asintóticas a las anteriores

Las zonas longitudinales contienen bandas miloníticas de espesor variable (desde algunos metros a varios cientos de metros e incluso algunas llegan al km); afectan fundamentalmente a los macizos graníticos, donde la densidad de bandas es mayor, y su extensión lateral parece limitada a éstos. Sin embargo, algunas de ellas continúan en los materiales paleozoicos encajantes con caracteres estructurales semejantes, e incluso algunas pueden seguirse prácticamente a lo largo de toda la Cadena, como la "falla nor-pirenaica" (FNP) o la "falla Mérens" (FM).

Los límites de las bandas miloníticas presentan buzamientos variables dentro de cada macizo granítico; así, en el macizo de Néouvielle, varían desde subverticales en las zonas centrales del macizo, hasta los 60 - 80º S en las zonas situadas al N, y 45 - 65º N en las zonas meridionales. El aspecto general es el de una geometría en abanico que es seguida también por los planos de esquistosidad principal hercínica (S_1) en los materiales metamórficos y sedimentarios de la Supraestructura paleozoica. Es posible que ambos fenómenos, la formación de esquistosidad y las bandas miloníticas con geometría en abanico tengan una causa común.

El desarrollo de cada banda milonítica, es generalmente discontínuo con digitaciones de las zonas deformadas, encerrando sectores de forma amigdalar de material no deformado; esta heterogeneidad en la deformación es característica de todas las bandas miloníticas, lo que se traduce en una compartimentación de los macizos graníticos en bloques no deformados separados por bandas de intensa deformación. Los bloques tienen un comportamiento rígido frente a la deformación fuertemente compresiva.

Las zonas de accidentes oblícuos, con directriz general NW - SE contienen bandas miloníticas de espesores menores que las anteriores, variando entre uno y varios metros, y con espaciamiento entre ellas de varios cientos de metros.

Estas zonas delimitan pliegues de igual dirección de amplitud variable, asimétricos y de ejes subverticales. Estos pliegues corresponden a generaciones sucesivas y superpuestas pero de carácter estrictamente local, siendo su geometría y grado de evolución el resultado de un campo de deformación local, que varía de un sector a otro dentro de una misma zona.

La mayoría de las zonas miloníticas han experimentado, además de "saltos" en la vertical con desarrollo de fallas inversas, movimientos de cizalla con desplazamiento longitudinal variable y discontinuo (desde varios metros a varios km). Los movimientos de falla inversa han levantado los bloques centrales de los macizos recubriendo los compartimentos más externos de los macizos, desarrollándose una folición milonítica de flujo entre los bloques.

En cuanto a la edad de la formación de las milonitas, así como a la cronología relativa a la deformación ligada a estas bandas, han sido motivo de controversia por parte de los distintos investigadores:

Carreras (1975), Carreras et al (1980), Saillant (1982) consideran que, al menos en los Pirineos orientales, las zonas de cizalla son tardihercínicas ya que están cortadas por diques de lamprófido datados por K - Ar como Keuper - Lías. Sin embargo para McCaig (1984, 1986) y McCaig & Miller (1986), las milonitas de los Pirineos centrales son alpinas (entre 100 m.a. y 40 m.a. por dataciones radiométricas en micas), o al menos han sido activas térmica y químicamente en tiempos alpinos.

Según Lamouroux et al (1980-1981), Soula et al (1986), Lamouroux (1987), las bandas milonitas longitudinales se generaron en el ciclo hercínico (ya que observan que las cuencas estefaniense-pérmicas se alinean según estas mismas directrices) y fueron reactivadas en distintos episodios tardihercínicos y alpinos, como fallas inversas e incluso como cabalgamientos. Sin embargo las bandas oblícuas se originarían en episodios alpinos, pudiendo actuar algunas de ellas como rampas oblícuas de cabalgamientos en distintos episodios posteriores.

Tampoco hay acuerdo a cerca de la cinética de la cizalla asociada a las bandas miloníticas. Algunos autores consideran que se formaron por desplazamientos de falla inversa, o inversa dextal (Carreras et al, 1980; Saillant, 1982; Guitard, 1970; McCaig, 1983, 1986). Sin embargo otros las asocian con desplazamientos sinestrales (Lamouroux et al, 1980-1981), o sinestrales combinados con un componente perpendicular a la Cadena de carácter extensional (transtensión), al menos desde el Pérmico hasta el Cretácico Superior (Soula et al, 1986).

Coincidiendo con el desarrollo de la fracturación y la formación de bandas miloníticas en los macizos graníticos y metamórficos de la zona axial pirenaica, y siguiendo las mismas directrices, se formaron en los bordes septentrional y meridional de esta zona, pequeñas cuencas sedimentarias en donde se localizan diversas rocas volcánicas andesíticas.

Los datos aportados en lo que respecta al carácter polifásico de las bandas miloníticas (con reactivaciones tectónicas),la diversidad de opiniones en cuanto a la dinámica de cizalla que los ha generado, así como la escasa precisión de los datos disponibles sobre su formación y tiempo de funcionamiento, hace que la evaluación de cualquier modelo energético de fricción, resulte tremendamente difícil si no imposible.

No obstante, el análisis de las paragénesis minerales formadas en estas bandas miloníticas puede suministrar cierta información sobre los gradientes térmicos y báricos alcanzados en algún momento de su dinámica.

3.2. Condiciones de temperatura y presión. Caracterización a través de paragénesis minerales.

El carácter polifásico ya mencionado de las bandas miloníticas condiciona los resultados térmicos y báricos deducibles de las paragénesis minerales que presenten ya que normalmente, las paragénesis de alta P y T son sustituídas por asociaciones retrógradas de menor gradiente.

Las rocas miloníticas asociadas a las dos familias principales de accidentes definidos en el apartado anterior para el Pirineo, presentan macroestructuras y asociaciones minerales distintas (Lamouroux et al, 1980- 1981; Lamouroux, 1987).

Así las rocas miloníticas longitudinales de los macizos graníticos y gneissicos están construídas por una alternancia de bandas micáceas (biotita y moscovita) y porfidoblastos de cuarzo y feldespatos (microclina y oligoclasa), en distinto grado de recristalización, deformación dúctil o fragmentación mecánica.

Estas bandas longitudinales son el resultado de la superposición de varias fases de deformación sucesivas en condiciones retrógradas, que producen efectos diferentes a lo largo de una misma zona dinámica. Así, Lamouroux (1987) distingue dos paragénesis distintas en distintos puntos de una misma banda milonítica:

- Una paragénesis de alta temperatura con Biotita ± Microclina ± Sillimanita (± Granate, según McCaig,1984), en condiciones de 650-700º C y 3-4 kb.

- Una paragénesis de baja temperatura con Moscovita ± Clorita ± Epidota ± Albita ± ± Opacos, en condiciones de 200-250º C y 1-2 kb.

Esta última paragénesis reemplaza a la anterior, evidenciando la movilización de diversos elementos en los minerales iniciales, con abundante circulación de fluídos.

En las bandas miloníticas oblícuas de los macizos graníticos y gneíssicos, se desarrollan lentículas alargadas de cuarzo y feldespato, con tamaño cristalino heterogéneo, presentando abundantes lamelas de deformación, en una "matriz" micácea (moscovita y clorita) con microcristales de cuarzo y opacos. Las fracturas, de forma sigmoidal, contienen neoformaciones de cuarzo, moscovita y clorita, pero no de biotita. (Lamouroux et al, 1980-1981). La paragénesis mineral que desarrollan es la de baja temperatura y la gran cantidad de cuarzo neoformado no puede explicarse sólo por recristalización de cuarzos anteriores (Lamouroux, 1987), siendo necesario un aporte adicional de SiO_2.

Passchier (1985), al analizar las bandas de rocas miloníticas del Macizo de Saint-Barthélemy, define para las últimas etapas de funcionamiento de estas bandas, condiciones de 450-550º C y 2-3 kb; además interpreta las venas pseudotaquilíticas asociadas a estas bandas como generadas por esfuerzos desviatorios locales de gran magnitud dentro de la deformación milonítica.

Esta asociación entre milonitas y pseudotaquilitas no es extraña ya que, en definitiva, las pseudotaquilitas constituyen la máxima expresión energética de un proceso dinámico de "shear" que alcanza la etapa de fusión anatéctica. (Sibson, 1975).

Los gradientes definidos en otras zonas y sobre distintas rocas miloníticas concuerdan con los definidos en el Pirineo, tanto en sus estimaciones máximas como en los efectos producidos por fenómenos retrógrados de las paragénesis iniciales. De esta manera, Theodore (1970) define unas temperaturas de 580-660ºC en rocas miloníticas de California, petrográficamente muy similares a las del Pirineo.

Vauchez (1978), en unas milonitas de Argelia, con asociaciones minerales de Moscovita ± Biotita ± Granate ± Clinozosita, deduce unas condiciones de temperatura de 500-650º C.

En los Alpes, similares a los Pirineos en muchos aspectos geológicos, existen abundantes rocas miloníticas asociadas a zonas de cizalla de alta temperatura (650-500º C), aunque también existen otras muchas, en donde rocas miloníticas claramente retrógradas, indican condiciones de más baja temperatura (< 300º C). (Zingg et al, 1990).

En relación a las condiciones físicas en la formación de las pseudotaquilitas debemos de decir que es una temática controvertida, principalmente en lo referente al papel jugado por la presión. De los resultados obtenidos por Maglouhlin (1986, 1989), de 600 º C y 8 kb, parece deducirse que la presión es el factor determinante desde la generación de procesos en estado sólido (milonitas), hasta otros que conllevan a la fusión parcial de los materiales (pseudotaquilitas), como ha observado, entre otros, Passchier (1985).

4. EL ENGROSAMIENTO CORTICAL Y SU SIGNIFICADO EN LA CADENA PIRENAICA.

A través de estudios de reflexión y refracción sísmica profunda, (ECORS Pyrenees Team, 1988), se ha caracterizado la Cadena Pirenaica como un órogeno con un acusado engrosamiento cortical, que puede alcanzar en algunos sectores de la zona central axial hasta 50 km.

La causa de este engrosamiento cortical ha sido y es diferentemente interpretada, debido a que puede ser analizado desde distintos puntos de vista y no siempre se tienen en cuenta la convergencia de fenómenos diversos de carácter dinámico, petrogenético y geofísico que se desarrollaron en tiempos hercínicos y tardihercínicos.

Este engrosamiento de la corteza se corresponde con anomalías gravimétricas negativas muy acusadas, del orden de -100 mGal en la zona axial al sur de la Falla Norpirenaica (Torne et al, 1989). Como es sabido, estas anomalías negativas indican un déficit de masa, pero también puede interpretarse como ocasionadas por la existencia de material cortical o infracortical que está ascendiendo. Este movimiento de materia puede resultar de una simple actuación de mecanismo isostático, con reequilibrio de las zonas de la Cordillera que más se están erosionando, pero también pueden ser indicativo de la presencia de masas de menor densidad en ascenso. Gallart (1982) establece un modelo gravimétrico para el Pirineo Oriental, comparando los valores de anomalía de Bouguer calculados y reales. Consigue la mejor concordancia mediante un conjunto de bloques corticales de diferente espesor y llega a la necesidad de introducir bloques o masas graníticas en determinados puntos. Una extensión de este modelo a toda la Cordillera parece indicar la presencia en profundidad de abundantes masas graníticas, en gran parte no aflorantes.

En este sentido, es importante tener en cuenta la gran cantidad de macizos graníticos hercínicos-tardihercínicos que afloran en la zona axial (recordemos que superficialmente suponen casi la mitad de la extensión de esta zona). De esta observación podemos preguntarnos ¿qué significado o qué relación pudo existir entre el origen y el emplazamiento de estas rocas graníticas y dicho engrosamiento?

Superficialmente la zona axial pirenaica está constituída por materiales paleozoicos diferentemente metamorfizados y replegados y por un gran número de afloramientos graníticos, desde batolitos hasta "stocks", que parecen converger en profundidad en una gran unidad "granítica". Pero ¿cuál es la composición de la corteza en profundidad?. A partir de los datos de velocidad de propagación de las ondas sísmicas longitudinales (Vp) se ha podido llegar a la consideración de que, al menos la corteza superior y media presentan valores propios de una corteza de características"graníticas s.l.", con un lento aumento de los valores de la velocidad (Vp) con la profundidad.

En cuanto a la naturaleza litológica de la corteza inferior, no parece existir un acuerdo general. Por la morfología que presentan los reflectores sísmicos se ha definido como una corteza "en lecho". Este tipo de corteza se ha reconocido en la mayoría de los perfiles sísmicos profundos realizados recientemente en distintas zonas orogénicas de Europa, Norteamérica, Australia (para recopilaciones de artículos sobre este tema ver Barazangi & Brown, 1986 a y b, Matthews & Smith, 1987 y Leven et al, 1990).

Según nuestra interpretación, ya esbozada por Sánchez Cela et al (1985), el fenómeno de engrosamiento cortical que comenzó en tiempos hercínicos, está íntimamente relacionado con el origen de las rocas graníticas a expensas del manto superior. Dicho fenómeno, de carácter episódico, parece haber sido activo hasta tiempos geológicos relativamente recientes. Aunque no se han datado en la Cadena Pirenaica rocas graníticas de edades recientes, consideramos que el engrosamiento en profundidad está causado por el aporte en tiempos alpinos de nueva corteza siálica a partir del manto superior.Esta consideración podría estar de acuerdo con los datos de reflexión sísmica de la corteza inferior, caracterizada como una corteza "en lecho".

Sobre el significado de esta corteza "en lecho" se han barajado hipótesis tan distintas como: a) intrusiones máficas en capas, b) diferenciación metamórfica "granulítica", c) zonas fuertemente cizalladas (miloníticas), d) zonas muy fracturadas rellenas de fluídos...; pero

también se ha sugerido que esta reflectividad "en lecho" puede ser interpretada como una zona de transición física entre la corteza y el manto (Meissner, 1973; Hale & Thompson, 1982). En este mismo sentido, estas estructuras en la corteza inferior podrían estar íntimamente relacionadas con la segregación siálica o "transformación" de la materia del manto (de características más densas) en materia siálica, dentro de un nuevo concepto físico-químico para el manto superior (Sánchez Cela, 1990).

A pesar de no existir un consenso sobre el significado de esta corteza "en lecho", sí parece existir un acuerdo en considerar que la reflectividad de la corteza inferior está relacionada con zonas de elevado flujo térmico y que es una característica más reciente que las que presentan la corteza superior y media, (Klemperer et al, 1986, 1987, 1990; Klemperer, 1987, 1989; Allmendinger et al, 1987; Bois et al, 1987 a), que parece avalar que el engrosamiento cortical está relacionado con un crecimiento de la corteza en profundidad.

Independientemente de la naturaleza de la corteza inferior, es evidente que la contribución de los materiales sedimentarios paleozoicos en el engrosamiento cortical no fue suficiente, aun teniendo en cuenta las potencias máximas de estos sedimentos (en total menos de 10 km). Si además tenemos en cuenta que la cronología de las masas graníticas es coetánea con este fenómeno y que, como ya hemos apuntado, la corteza presenta constantes físicas que coinciden con las de las rocas graníticas, es obvio que tales rocas tuvieron que jugar un papel muy importante en el origen del engrosamiento cortical.

Sin embargo, aunque todos los geólogos están de acuerdo en definir la corteza como de características "graníticas" no ocurre lo mismo para determinar como tuvo lugar este fenómeno, obviando, en su mayoría, el carácter "granítico" en profundidad. Según nuestra interpretación, consideramos que estos macizos graníticos, superficialmente individualizados (circunscritos, domáticos, etc., según su evolución espacial y temporal), corresponden en profundidad a una única unidad granítica, constituyendo la "columna vertebral de la Cadena".Por el contrario, muchos autores consideran que las rocas graníticas constituyen son masas diapíricas desenraizadas con una escasa representación en el volumen total de la corteza. Según esta última perspectiva, se plantearían problemas tan importantes como ¿cuál es el material que constituye el resto de corteza siálica engrosada?.

Aquí subyace cuál es el origen y formación de estas rocas graníticas. Aunque no es nuestro objetivo tratar aquí esta importante problemática, es difícil explicar la formación de un volumen importante de rocas graníticas a través de la anatexia cortical, ya que por una parte, no existía el "protolito" adecuado (la columna litológica paleozoica puede caracterizarse en conjunto de composición "margosa", con un espesor total inferior a los 10 km) y, por otra es difícil de explicar cuáles serían los mecanismos energéticos que pudieran ser capaces de provocar una fusión tan generalizada. Es por ello por lo que el origen de las rocas graníticas debería contemplarse desde otra perspectiva geológica.

En relación al engrosamiento y acreción cortical es también interesante analizar cómo se manifiesta la morfología estructural de la zona axial en profundidad. A partir del perfil ECORS-Pirineos (ECORS Pyrenees Team, 1988) se ha puesto de manifiesto que la estructura general superficial con morfología "en abanico" de la zona axial puede prolongarse en profundidad bastantes km, hasta la corteza media . Sin embargo, en las zonas subpirenaicas, en ambas vertientes de la zona axial, no hay evidencias de trazas de deformación bajo los materiales mesozoicos y cenozoicos, al igual que el engrosamiento cortical no es tan patente en estas zonas.

Entonces, ¿qué significado petroestructural puede tener esta morfología "en abanico" de la zona axial con fallas inversas en la corteza?. Todo hace pensar que precisamente el fenómeno de

engrosamiento cortical puede estar relacionado con esta morfología especial, y que justamente las bandas miloníticas, que siguen esta misma estructuración, sean una manifestación de la dinámica de este fenómeno.

Debemos de reseñar también que el engrosamiento cortical en la Cadena Pirenaica tuvo lugar en varias épocas geológicas, al menos hercínica-tardihercínicas. Esto se deduce a partir de los estudios estructurales detallados de algunos macizos metamórficos de la zona axial. En ellos se ha determinado que la estructuración de la Infraestructura de la Cadena (constituída por rocas paleozoicas metamórficas y abundantes rocas graníticas) es una característica "más reciente" que la que presenta la Supraestructura (constituída por rocas epimetamórficas del Paleozoico Superior). Así lo han deducido, entre otros, Verhoef et al (1984) en el macizo de Aston; de Bresser et al (1986) en la parte occidental del macizo de Lys-Caillaouas y van den Eeckhout (1984, 1986) en el macizo de Hospitalet.

Esta Infraestructura, más joven, aflora en estructuras domáticas metamórficas, con esquistosidades subhorizontales o ligeramente inclinadas de tendencia concéntrica, que son más modernas que las esquistosidades subverticales de las rocas epimetamórficas que constituyen la Supraestructura. Estas estructuras domáticas, y las esquistosidades subhorizontales que contienen, constituyen los "domos metamórficos" asociados al emplazamiento de las rocas graníticas en profundidad.

Además, a través de dataciones absolutas en algunos de estos macizos, como en el Canigou, se han establecido edades hercínicas más recientes (~ 330 m.a.) en los granitos biotíticos "profundos" que en los "ortogneisses" supuestamente precámbricos (~ 580 m.a.) suprayacentes, habiéndose observado que durante el emplazamiento de los granitos profundos se originó una esquistosidad domática que afectó a los ortogneisses. (Soliva et al, 1989; Gibson, 1989).

Tanto los datos estructurales como la cronología de formación de los macizos graníticos, hacen pensar que el engrosamiento cortical se realizó en varias fases petro-estructurales.

Todo ello, añadido a las interpretaciones geofísicas, parece indicar que el engrosamiento cortical está relacionado con el aporte en profundidad de nueva materia granítica y que a causa de la isostasia se manifiesta también en superficie a través de una elevación del terreno y la formación del orógeno. Así, la corteza va engrosándose por el aporte más o menos episódico de material "granítico" (en gran parte juvenil) que se emplaza en diferentes niveles estructurales deformando y abombando las estructuras superiores y alcanzando, en algunos casos, la cobertera sedimentaria. Esta nueva materia siálica origina una dinámica cortical fuertemente compresiva debido al aumento del volumen en profundidad.

Este abombamiento cortical compresivo puede responder fracturándose en algunas zonas, en especial en las de borde (zonas de ruptura estructural). Es precisamente en estas zonas de "debilidad" estructural y fuerte componente compresivo donde pueden tener lugar importantes fenómenos petrogenéticos y estructurales, relacionados con la liberación de energía térmica y química a través de zonas de falla.

5. CONSIDERACIONES QUIMICAS EN LAS ZONAS DE FRICCION Y ROCAS MILONITICAS.

5.1. Sobre la movilidad de los elementos químicos.

Es sabido que cualquier cambio en las condiciones físicas de un sistema natural en equilibrio, podrá originar, en su entorno, un estado de energía potencial manifestándose, entre

otros aspectos, por la creación de potenciales químicos entre sus componentes que podrán formar campos de fuerzas capaces de movilizar cargas y transferirlas de una posición a otra. Estos potenciales tenderán a equilibrarse por medio de reacciones químicas entre los propios minerales del sistema o a través de una migración metasomática entre los minerales y la fase fluída movilizada.

Este proceso de movilización selectiva de sustancias se desarrollará fundamentalmente en fracturas mecánicas, fisuras y cavidades, pero también puede desarrollarse a través de los espacios intergranulares, ya sea mediante infiltración favorecida por la presión de fluídos, o por difusión a través de poros, entre los límites de granos, o entre los iones de una red cristalina. La migración de sustancias a grandes distancias se ve influenciada por la dinámica a lo largo de planos de falla, y por la actividad de soluciones fluídas, siendo en este sentido más efectivo (incluso en distancias de algunos km), el proceso de infiltración que el de difusión; sin embargo el mecanismo de difusión controla muchos procesos geológicos y puede ser importante sobre todo en condiciones de alta temperatura; así las modificaciones físicas en la interfase entre la corteza inferior y manto superior, podrían favorecer la relación de migraciones atómicas a través de las redes cristalinas y los potenciales creados en los fundidos silicatados naturales tenderán a equilibrarse a través de la difusión de las especies en migracion (Holloway & Wood, 1988).

La investigación empírica sobre procesos naturales ha demostrado que existe una movilidad diferencial de los componentes iónicos de un sistema, influenciada por las propiedades de los elementos, las condiciones ambientales y los propios mecanismos de transferencia iónica. Es asumido que esta movilidad está controlada por el tamaño iónico y por la carga, siendo en general los cationes menos móviles que los aniones, (Ramberg, 1952; Rankama & Sahama, 1952; Povarenmych, 1954) y entre todos los iones metálicos, los más fácilmente transportables son los alcalinos y alcalino-térreos; así las actividades de Na y K favorecen su migración hasta distancias del orden del km, mientras que la transferencia de otros componentes como Fe, Al, Ti es extremadamente limitada. Sin embargo la actividad de la fase fluída juega un papel importante, tanto en la igualación de los potenciales químicos creados, favoreciendo las transformaciones mineralógicas, como en el de ser un agente de transferencia de calor y de masa capaz de transportar grandes cantidades de solutos (SiO_2, Ca, Al, Fe, Mg, S,...) sobre distancias considerables (Norton & Knight, 1977, Holloway & Wood, 1988).

Pero además, el aumento de la temperatura favorece la solubilidad de algunas especies como el cuarzo, siendo transferida gran cantidad de SiO_2, (Anderson & Burnham, 1965), en especial en las zonas de fracture, (Korzhinskii, 1970). La influencia de la T como factor de equilibrio en los procesos de transferencia iónica queda reflejada a través de la variación del coeficiente de difusión; así a 25º C este coeficiente para algunos iones, es $K^+ = 1{,}96.10^{-5}$ cm^2 s^{-1}; $Na^+ = 1{,}33.10^{-5}$ cm^2 s^{-1}; $Ca^{++} = 0{,}79.10^{-5}$ cm^2 s^{-1}. Estos valores se ven directamente incrementados en función de la temperatura, según la ecuación:

$$D = RT \frac{\Omega}{V_i}$$

en donde R es la constante de los gases, T la temperatura absoluta, Ω la conductividad eléctrica de los iones a una T dada y V_i la valencia del ion, llegando a aumentar 20 veces su valor, en el intervalo de 0º a 250º C.

Como ejemplo, las variaciones del coeficiente de difusión del K^+, a distintas temperaturas son (Gray, 1957): a 5º C —> $D_{(K^+)} = 1{,}16 \cdot 10^{-5}$ cm^2 S^{-1}; a 55º C —> $D_{(K^+)} = 3{,}49 \cdot 10^{-5}$ cm^2 S^{-1}; a 500ºC —> $D_{(K^+)} = 109 \cdot 10^{-5}$ cm^2 S^{-1}.

Aunque la tendencia natural en todo sistema está dirigida a alcanzar el equilibrio químico, las investigaciones y controversias entre científicos americanos (Weill & Fyfe, 1964, Thompson, 1959, 1970;..) y soviéticos (Korzhinskii, 1953, 1965, 1970; Nikolaev, 1957; ..) han contribuído a establecer las bases sobre la migración de sustancias, pero también a concluir que es muy probable que en muchos casos los sistemas de rocas naturales no hayan alcanzado el equilibrio químico, aunque se aproximen estadísticamente a él.

De cualquier modo, esta tendencia al equilibrio podrá plasmarse a través de transformaciones mineralógicas reversibles, ya sea por recristalización en estado sólido, o haciendo intervenir la fase fluída movilizada.

En resumen, la dinámica hercínica efectiva en fracturas entre bloques siálicos en la Cadena Pirenaica pudo favorecer tanto la creación de altos gradientes térmicos como los procesos de migración selectiva de materia.

5.2. Significado químico de las zonas de fricción y rocas miloníticas.

De los apartados anteriores se deduce que en las zonas de alta energía mecánica desarrolladas sobre materiales siálicos (graníticos) va a tener lugar una gran transferencia de la materia más móvil hacia las capas más superficiales, es decir hacia zonas menos energéticas, principalmente de aquellos elementos más móviles y abundantes en la Corteza como son sílice, y en menor proporción, sodio, potasio y otros elementos geoquímicamente análogos pero minoritarios en la corteza.

Aunque en base a consideraciones químicas y energéticas se puede deducir la movilización de la materia en las zonas de fricción de la corteza, la existencia de rocas miloníticas puede ayudar a corroborar dicha movilidad, principalmente la de los elementos químicos citados anteriormente.

Como hemos comentado, las rocas miloníticas incluyen un amplio grupo de rocas en las cuales se pueden diferenciar diversos tipos estructurales-texturales y mineralógicos, aunque todas ellas responden a tipologías graníticas, desde las de mayor gradiente T-P, con sillimanita-granate, a las de menor gradiente, retrógradas, con abundantes minerales arcillo-micáceos.

En relación a las milonitas de alta T-P debemos hacer resaltar que estas rocas tienen menor proporción de cuarzo modal que las rocas graníticas encajantes; y siempre como un mineral tardío. Esto parece indicar que durante la génesis de estas rocas miloníticas hubo una gran movilidad de SiO_2, favorecido por la alta T y la existencia de fluídos (minerales hidratados). Aunque no se aprecian diferencias acusadas en el porcentaje de feldespatos se puede predecir también la movilidad de los álcalis, junto a la sílice, durante la formación de dichas rocas.

Es en las abundantes rocas miloníticas polifásicas y retrógradas en donde se pueden obtener interesantes datos geoquímicos sobre la movilidad de algunos elementos. Estas rocas al igual que las de alta T-P tienen menor proporción de cuarzo, también de generación tardía, en relación a los granitos encajantes. Pero también tienen menor proporción de feldespatos (alcalinos y calcosódicos) y de biotita; por el contrario tienen abundante moscovita, minerales arcillo-micáceos y cloritas.

De los estudios petrológicos se deduce fácilmente que la mayor parte de la moscovita y minerales químicamente análogos (sericitas) provienen de la alteración de los feldespatos alcalinos y calcosódicos. Estas alteraciones asociadas a regímenes dinámicos bajo un ambiente rico en fluídos (minerales hidratados), que podrían definirse en la literatura petrológica como sericitizaciones, sausuritizaciones, etc., conllevan balances químicos interesantes en donde los rasgos más sobresalientes radican en la liberación de sílice, sodio y potasio y otros elementos móviles minoritarios de menor significado químico.

Así en la "alteración" de los feldespatos alcalinos y calcosódicos se libera sílice y álcalis:

$$\text{Ortosa} + \text{Plagioclasa} + H_2O \longrightarrow \text{Moscovita-Illita} + SiO_2 + K^+ + Na^+.$$

En la cloritización de las biotitas también se liberan elementos sílico-alcalinos:

$$\text{Biotita} \longrightarrow \text{Clorita} + SiO_2 + K^+$$

Estos elementos a mayor o menor temperatura se liberarán hacia zonas más superficiales en donde pueden participar en diversos procesos petrogenéticos.

5.3. Significado químico del engrosamiento cortical

En muchas zonas orogénicas como el Pirineo parece existir una relación entre dinámica cortical y procesos petrogenéticos con el engrosamiento cortical. Este engrosamiento, aunque definido principalmente por datos geofísicos, puede ser también deducido a partir de datos geológicos, en especial petrológicos.

También parece existir una relación entre el engrosamiento cortical y la reflectividad sísmica de la corteza inferior. Estas reflexiones son muy comunes en muchas zonas de la corteza (Klemperer, 1989; Klemperer et al, 1986, 1990; Bois et al, 1987 b; Allmendinger et al, 1987, etc.). La interpretación de esta corteza "en lecho", con Vp de 7 a 6 km / s es muy diversa, aunque la relación con zonas de "cizalla" parece ser la más defendida tanto a partir del estudio sísmico de complejos metamórficos de alto grado (Fountain et al, 1987; Hurich et al, 1985, Green et al, 1990), como a partir de la construcción de sismogramas sintéticos por modelización sísmica (Reston, 1987; Blundell, 1990). En el aspecto petrológico los valores de Vp podrían corresponder a rocas "granulíticas" que encajarían con el carácter "estructural-metamórfico" de dicha zona.

Desde nuestro punto de vista es difícil entender:

a) La existencia de una diferencia muy marcada entre el carácter sísmico de la corteza superior (transparente) y la inferior (reflectora).

b) Cómo procesos estructurales-metamórficos pueden tener lugar a tales profundidades entre 20 y 40 km).

Una posible explicación para tales reflexiones podría ser que ellas son consecuencia de fenómenos físico-químicos que tienen lugar en la transición manto superior-corteza y ligados con el fenómeno de engrosamiento siálico a expensas del manto superior. En este sentido tales reflexiones corresponderían más a cambios de fase físicos que químicos, y la interpretación, por muchos geofísicos, de que las reflexiones sísmicas de la corteza inferior corresponden a estructuras más recientes que las de la corteza superior parecen indicar que el engrosamiento cortical se realiza fundamentalmente a expensas del manto superior.

En relación con estas zonas de fracturación compresiva más o menos profunda tienen lugar los fenómenos energéticos y químicos considerados en este trabajo. Si a esto añadimos el carácter siálico del engrosamiento, no es difícil entender que la naturaleza del material que se va a movilizar es principalmente sílico-alcalina.

6. IMPLICACIONES PETROGENETICAS DE LAS ZONAS DE FRICCION Y ROCAS MILONITICAS.

La relación entre procesos petrogenéticos diversos y la generación de calor asociado a zonas dinámicas, ha sido considerada por algunos autores desde hace bastantes años. Así Scott & Breyer (1953) fueron unos de los primeros autores en considerar procesos de fusión parcial debidos a efectos friccionales en el Himalaya. La génesis de pseudotaquilitas por fusión parcial inducida por el calor friccional generado en la falla Alpina (Nueva Zelanda) es considerada por Sibson (1975) y Wallace (1976).

Posteriormente se han reconocido diversos procesos petrogenéticos asociados a "shear heating" en distintas zonas de la Tierra: Graham & England (1976), Barton & England (1979), England & Thompson (1984).

En relación con zonas de fricción se considera la generación de diversas rocas calcoalcalinas, desde verdaderos granitos (Strong & Hanmer, 1981; Molnar et al, 1983), hasta andesitas. Por lo que respecta a estas últimas rocas son abundantes los autores que, dentro o no del esquema de Tectónica de Placas, relacionan directa o indirectamente la generación de magmas andesíticos con la existencia y el funcionamiento de grandes fallas inversas. (Nicolas et al, 1977; Scholz et al, 1979; Scholz, 1980; Spohn, 1980; Brun & Cobbol, 1980; Turcotte & Schubert, 1982,etc.).

Este mecanismo energético ha sido utilizado por nosotros para relacionar las rocas andesíticas del Pirineo con el fenómeno de engrosamiento cortical (Sánchez Cela et al, 1985) y para relacionar los procesos ígneos y los metamórficos en las rocas volcánicas del SE de España, que suelen contener enclaves metamórficos de alta T y P (Sánchez Cela et al, en prensa).

Aplicado al caso pirenaico, para comprender la efectividad del proceso dinámico de fricción como fuente térmica y química del volcanismo "andesítico" calcoalcalino, es obvio que hay que estudiarlo en el contexto espacial y temporal tardihercínico. Como vestigios de esta actividad dinámica están las importantes fracturas E-W, asociadas en muchos casos a rocas miloníticas.

Estas rocas miloníticas aparecen asociadas a los macizos graníticos de la zona axial pirenaica y a los macizos metamórficos norpirenaicos. (Lamouroux, 1976, 1987; Lamouroux et al, 1980-1981; Soula et al, 1986). También existen pseudotaquilitas (Passchier, 1982 a, b; 1984) asociadas a extensas zonas de cizalla y fallas inversas de directriz principal E-W. Debido al carácter polifásico de la mayor parte de las rocas miloníticas, a causa de la superposición de varias fases de deformación, es difícil de precisar los valores P y T máximos alcanzados en estas zonas dinámicas. A pesar de este carácter polifásico retrógrado, en algunos sectores de la falla de Mérens y en las zonas miloníticas de algunos macizos metamórficos, como en el de Saint Barthélemy, se han reconocido paragénesis (Biotita ± Microclina ± Sillimanita ± Cordierita ± Almandino) que corresponden a 650º - 750º C y 4-5 kb, coetáneos con el metamorfismo regional de alta T y P hercínicos (Passchier, 1982, a, b; 1985; Lamouroux, 1987).

En muchas rocas volcánicas andesíticas son frecuentes los enclaves metamórficos con paragénesis de Corindón ± Espinela ± Sillimanita ± Granate..., como ocurre en algunas rocas volcánicas del SE de España. Nuestros estudios petrológicos (Sánchez Cela et al, en prensa) parecen indicar la existencia de una relación petrogenética entre el proceso metamórfico y el volcánico. Así el fenómeno volcánico parece corresponder a un estado de descompresión subsiguiente al proceso metamórfico de alta T y P (stress) previo.

Ambos procesos, el metamórfico y el volcánico, parecen estar relacionados con zonas de fricción que se asocian al fenómeno de engrosamiento cortical.

Como ya hemos citado las zonas de fricción y las rocas miloníticas constituyeron, durante su generación, importantes fuentes de energía térmica y química (elementos siálicos). Estas fuentes térmicas y químicas pueden tener mucha importancia en la petrogénesis de muchas rocas ígneas en la zona pirenaica como parece ocurrir con las rocas volcánicas del Estefaniense.

Pensamos que los aportes térmicos y químicos necesarios para la formación de dichas rocas volcánicas proceden precisamente de las zonas de fricción y las rocas miloníticas que hemos descrito. Las siguientes observaciones apoyan esta interpretación:

1º) - Ausencia de rocas volcánicas en el basamento preEstefaniense, existiendo por el contrario, en análogas o iguales alineaciones estructurales, rocas miloníticas o zonas de fricción de evidente significado térmico y químico.

2º) - Existencia, por el contrario, de una correlación entre los tipos petrológicos volcánicos y el nivel estratigráfico al que afectan. Las rocas volcánicas del Estefaniense son diferentes de las del Pérmico y ambas no aparecen en otros niveles estratigráficos inferiores.

3º) - Existencia en muchos casos, de evidentes relaciones petrogenéticas entre las rocas volcánicas y los materiales encajantes sedimentarios: relaciones de contacto,datos petrológicos, mineralógicos, químicos, etc.

Todos ellos, junto a otros datos, deducciones y consideraciones, nos han llevado a considerar el origen de las rocas volcánicas del Estefaniense en la zona pirenaica como íntimamente relacionado con las manifestaciones químicas y energéticas asociadas a zonas de fricción y rocas miloníticas, junto a la mayor o menor participación de la cobertera sedimentaria (Sánchez Cela & Lapuente, en prensa).

7. BIBLIOGRAFIA.

Allmendinger, R.W., Nelson, K.D., Potter, C.J., Barazangi, M., Brown, L.D. & Oliver, J.E. (1987). Deep seismic reflection characteristics of the continental crust.*Geology,* 15, 304-310.

Anderson, E.M. (1951). *The dynamics of faulting.* Oliver & Boyd (eds). London.

Anderson, G.M. & Burnham, C.W. (1965). The solubility of quarzt in supercritical water. *Am.J. Sci.*, 263, 494-511.

Barazangi, M. & Brown, L. (eds.) (1986) (a). Reflection Seismology: a global perspective. *Am. Geophys. Union Geodyn. Ser.*, 13, 311 pp.

Barazangi, M. & Brown, L. (eds.) (1986) (b). Reflection Seismology: the continental crust. *Am. Geophys. Union Geodyn. Ser.*, 14, 339 pp.

Barton, C.M. & England, P.C. (1979). Shear heating at the Olympus (Greece) thrust and the deformational propierties of carbonates at geological strain rates.*Geol.Soc. Amer. Bull.*, 90, 438-492.

Blundell, D. (1990). Seismic images of continental lithosphere. *J. Geol. Soc. London,* 147, 895-913.

Bois, C., Damotte, B., Mascle, A., Cazes, M., Hirn, A. & Biju-Duval, B. (1987).(a) Operations and main results of the ECORS project in France. *Geophys. J. R. Astr. Soc.*, 89, 279-286.

Bois, C., Cazes, M., Hirn, A., Matte, P., Mascle, A., Montadert, L. & Pinet, B.(1987).(b). Crustal laminations in deep seismic profiles in France and neighbouring areas. *Geophys. J. R. Astr. Soc.*, 89, 297-304.

Brewer, J. (1981). The thermal effects of thrust faulting. *Earth. Planet. Sci. Lett.*, 56, 233-244.

Brodie, K.H., Rex, D. & Rutter, E.H. (1989). On the age of deep crustal extensional faulting in the Ivrea zone, northern Italy. In: Eward et al (eds.) ."Alpine Tectonics". *Geol.Soc. London Spec. Pap.*, 45, 203-210.

Brun, J.P., & Cobbold, P.R. (1980). Strain heating and thermal softening in the continental shear zones: a review. *J. Struct .Geol.*, 2, 149-158.

Carreras, J. (1975). Las deformaciones tardi-hercínicas en el litoral septentrional de la península, Cabo de Creus (Prov. de Gerona, España). La génesis de las bandas miloníticas.*Acta Geol. Hisp.*, 10, 109-115.

Carreras, J. & Santanach, P. (1973). Micropliegues y movimiento en los cizallamientos profundos del Cabo de Creus (Prov. Gerona, España).*Estudios Geol.*, 29, 439-450.

Carreras, J., Estrada, A. & White, S. (1977). The effects of folding on the c-axis fabrics of a quartz mylonite.*Tectonophysics,* 39, 3-24.

Carreras, J., Julibert, M. & Santanach, P. (1980). Hercynian mylonite belts in the eastern Pyrenees: an example of shear zones associated with late folding.*J . Struct. Geol.*, 2, 5-9.

De Bresser, J.H.P., Majoor, F.J.M., & Ploegsma, M. (1986). New insights in the structural and metamorphic histoty of the western Lys-Caillaouas massif (Central Pyrenees, France).*Geol. Mijnbouw*, 65, 177-187.

ECORS Pyrenees team (1988). The ECORS deep reflection seismic survey across the Pyrenees. *Nature.*, 331, 508-511.

England, P.C.&Thompson, A.B.(1984). Pressure-temperature-time paths of regional methamorphim. I. Heat transfer during the evolution of regions of thickened continental crust.*J. Petrol.*, 25, 894-928.

Fleitout, L. & Froidevaux, C. (1980). Thermal and mechanical evolution of shear zones, role of shear heating, effect of non-newtonian law of deformation and possible mechanism for melting. *J.Struct. Geol.*, 2, 159-164.

Fountain, D. McDonaught, D. & Gorham, J. (1987). Seismic reflection models of the continental crust based on metamorphic terranes. *Geophys. J. R. Astr. Soc.*, 89, 61-67,

Fyfe, W.S., Price, N.J. & Thompson, A.B. (1978). *Fluids in the Earth's Crust. Developments in Geochemistry* 1. Elsevier, 383 pp.

Gallart, J. (1982). Aportación de la geofísica al conocimiento geodinámico de los Pirineos. *Rev. de Geofísica,* 38, 13-30.

Gibson, R.L. (1989). The relationship between deformation and metamorphism in the Canigou Massif, Pyrenees: a case study. *Geol. Mijnbouw,* 68, 345-356.

Graham, C.M. & England, P.C. (1976). Thermal regimes and regional metamorphism in the vicinity of overthrust faults: An example of shear heating and inverted methamorphic zonation from southern California.*Earth. Planet. Sci. Lett.*, 31, 142-152.

Gray, D.E. (Ed). (1957). *American Institute of Physics Handbook.* McGraw-Hill.

Green, A., Milkereit, B., Percival, J., Davison, A., Parrish, R., Cook, F., Geis, W., Cannon, W., Hutchinson, D., West, G., & Clowest. (1990). Origin of deep crustal reflections: seismic profiling across high-grade metamorphic terranes in Canada. In: Laven et al (eds.)" Seismic Probing of Continents and their Margins".*Tectonophysics,* 173, 627-638.

Guitard, G. (1970). Le métamorphisme hercynien mésozonal et les gneiss oeillés du massif du Canigou (Pyrénées Orientales).*Mém. B.R.G.M.*, 63, 349 pp.

Hale, L.D. & Thompson, G.A. (1982). The seismic reflection character of the continental Mohorovicic discontinuity.*J. Geophys. Res.*, 87, 4625-4635.

Higgins, M.W. (1971). Cataclastic rocks. *Prof. Pap. U.S. Geol. Serv.*, 687, 1-97.

Holloway, J.R. & Wood, B.J. (1988). Simulating the Earth: experimental geochemistry. Unwin Hyman (ed.), London, 196 pp.

Hurich, C., Smithson, S., Fountain, D. & Humphreys, M. (1985). Seismic evidence of mylonite reflectivity and deep structure in the Kettle dome metamorphic core complex, Washingtong. *Geology,* 13, 577-580.

Jaupart, C. & Provost, A. (1985). Heat focussing, granite genesis and inverted metamorphic gradients in continental collision zones.*Earth.Planet.Sci.Lett.*, 73, 385-397.

Klemperer, S.L. (1987). A relation between continental heat flow and the seismic reflectivity of the lower crust.*J. Geophys. Res.*, 61, 1-11.

Klemperer, S.L. (1989). Deep seismic reflection profiling and the growth of the continental crust: in Ashwal (Edit.) "Growth of the Continental crust".*Tectonophysics*, 161, 233-244.

Klemperer, S.L. & B.I.R.P.S. Group. (1987). Reflectivity of the crystalline crust: hypothesis and test.*Geophys. J. R. Astr. Soc.*, 89, 217-222.

Klemperer, S.L., Hauge, T.A., Hauser, E.C., Oliver, J.E. &Potter, C.J. (1986). The Moho northern Basin and Range province, Nevada, along COCORP 40ºN seismic reflection transect.*Geol. Soc. Am. Bull.*, 97, 603-618.

Klemperer, S.L., Hobbs, R.W. & Freeman, B. (1990). Dating the source of lower crust reflectivity using BIRPS deep seismic profiles across the Iapetus suture. in: Leven et al (eds.) "Seismic Probing of Continents and their margins."*Tectonophysics*, 173, 445-454.

Korzhinskii, D.S. (1953). Survey of metasomatic processes. In: Main Problems of Study of the Magmatic Ore Deposits (in Russian).*AN SSSR, Moscow,* 332-452.

Korzhinskii, D.S. (1965). The theory of systems with perfectly mobile components and processes of mineral formation.*Am. J. Sci.,* 263, 193-205.

Korzhinskii, D.S. (1970).*Theory of Metasomatic Zoning.* Clarendon. Oxford. 162 pp.

Lachenbruch, A.H. (1980). Frictional heating, fluid pressure, and the resistance to fault motion.*J. Geophys. Res.,* 85, 6097-6112.

Lamouroux, C. (1976). Les mylonites dans le massif du Néouvielle "(Textures, déformations intracristalllines). *Thèse 3º cycle, Univ. Toulouse.,* (Inédite) 148 pp.

Lamouroux, C. (1987). Les mylonites des Pyrénées: classification, mode de formation et évolution.*Thèse. Sci. Toulouse.,* 553 pp.

Lamouroux, C., Debat, P., Deramond, J. & Majeste-Menjoulas, C. (1979). Influence de massifs plutoniques hercyniens dans l´evolution des structures pyrénéennes: exemple du massif du Néouvielle. *Bull. Soc. Géol. France,* 21, 213-220.

Lamouroux, C., Soula, J.C. & Rodaz, B. (1980-1981). Les zones mylonitises des massifs du Bassies et d´Aston (Haute Ariège). *Bull. B.R.G.M.,* 1, 103-111.

Leven, J. H., Finlavson, D. M., Wright, C., Dooley, J. C. & Kennett, B.L.N. (eds.) (1990). Seismic probing of continents and their margins.*Tectonophysics,* 173, 1-641.

Lockner, D.A. & Okubo, P.G. (1983). Measurements of frictional heating in granite. *J.Geophys. Res.,* 88, 4313-4320.

Magloughlin, J.F. (1986). A new occurrence of pseudotachylite, Wenatchee Ridge area, North Cascades, Washington.*Geol. Soc. Am. Abs.w. Prog.,* 18, 153.

Magloughlin, J.F. (1989). The nature and significance of pseudotachylite from the Nason terrane, North Cascade Mountains, Washington.*J. Struct. Geol.,* 11, 907-917.

Matthews, D.H. & Smith, C.A. (eds.) (1987). Deep seismic reflection profiling of the continental lithosphere. *Geophys. J. R. Astr. Soc.,* 89, 1-447.

Mc Caig, A.M. (1983). Kinematics, age and geochemistry of shear zones in the Aston-Hospitalet Massif, Pyrenees. *Ph. D. Thesis. University of Cambridge,* (Unpublished).

Mc Caig, A.M. (1984). Fluid-rock interaction in some shear zones from the Central Pyrenees.*J.Metamor. Geol.,* 2, 129-141.

Mc Caig, A.M. (1986). Thick-and thin-skinned tectonics in the Pyrenees.*Tectonophysics.* 129, 319-342.

Mc Caig, A.M. & Miller, J.A. (1986). 40Ar-39Ar age of mylonites along the Mérens fault, Central Pyrenees.*Tectonophysics.* 129, 149-172.

Mc Kenzie, D. & Brune, J. (1972). Melting of fault planes during large earthquakes. *Geophys. J. R. Astr. Soc.,* 29, 65-78.

Meissner, R. (1973). The "Moho" as a transition zone,*Geophysical Surveging,* 1,195-216.

Molnar, P., Chen, W.P. & Padovani, E. (1983). Calculated temperatures in overthrust terrains and possible combinations of heat sources responsible for the Tertiary granites in the greater Himalaya. *J.Geophys. Res.,* 88, 6415-6429.

Nicolas, A., Bouchez, J.L., Blaise, J. & Poirier, J.P.(1977). Geological aspects of deformation in continental shear zones. *Tectonophysics,* 42, 55-73.

Nikolaev, V.A. (1957). Application of thermodynamics on some petrological processes. (In Russian).*Zap. Vses. miner. Obschez.* 86, 223-237.

Norton, D. & Knight, J. (1977). Transport phenomena in hydrothermal sytems: Cooling plutons. *Amer. J. Sci.,* 277, 937-981.

Passchier, C.W. (1984). Mylonite-dominated footwall geometry in a shear zone,Central Pyrenees.*Geol. Mag.,* 121, 5, 429-436.

Passchier, C.W. (1985). Water-deficient mylonite zones-An example from the Pyrenees. *Lithos,* 18, 115-127.

Passchier, C.W. (1982) (a). Pseudotachylites and the development of ultramylonite bands in the Saint Barthélemy Massif. French Pyrenees.*J. Struct. Geol.,* 4, 69-79.

Passchier, C.W. (1982) (b). Mylonitic deformation in the Saint Barthélemy Massif, French Pyrenees, with emphasis on the genetic relationship between ultramylonite and pseudotachylite.*Geol. Inst. Univ. Amsterdam, Pap. Geol. Ser.,* 1, 173 pp.

Povarenmych, A.S.(1954). Some problems of alterations of granites along dykes. (In Russian)*Trudy Krivojr. gorm. Instituta.,* 1.

Price, N.J. (1970). Laws of rock behaviour in the earth´s crust. In : W.H. Somerton (ed.), "Rock Mechanics Theory and Practice". pp.1-23. Proc. 11 th. Symp. on Rock Mechanics, Berkeley, Calif.*Am. Inst. Min. Metall. Pet. Eng.*

Ramberg, H.(1952).*The Origin of Metamorphic and Methasomatic Rocks.* Chicago.Univ.Press, 317 pp.

Rankama, K. & Sahama, T.K. (1952). *Geochemistry.* Chicago. Univ. Press, 912 pp.

Reston, T. (1987). Spatial interference, reflection character and the structure of the lower crust under extension. *Ann. Geophys,* 5B, 339-348.

Saillant, J.P. (1982). La faille de Mérens (Pyrenées Orientales): microstructures et mylonites.*Ph.D. thesis, 3rd. cycle. Univ. Paris VI.* (Inédite).

Sánchez Cela, V. (1990). Energy and geochemical-geophysical data as critical aspects of the Plate Tectonics Theory. In:"Critical aspects of the Plate Tectonics Theory". *Theophrastus Publ. S.A. Athens,* 14-43.

Sánchez Cela, V., Ortiga, M. & Lapuente M.P. (1985). The Pyrenean orogenic belt. Petrological processes in relationship with the granitic crustal evolution .In "The Crust-The Significance of Granites-Gneisses in the Lithosphere".*Theophrastus Publ. Athens.,* 95-130.

Sánchez Cela, V., Auqué L.F. & Lapuente M.P. (en prensa). Petrological significance of High T-P Metamorphic enclaves in dacitic-andesitic rocks. In: "High Grade Metamorphism".*Theophrastus Publ. S.A. Athens.*

Sánchez Cela, V. & Lapuente, M.P. (en prensa). A petrogenetic model for the Late Hercynian Pyrenean andesitic rocks. (Enviado a *Modern Geology*).

Scholz, C.H. (1980). Shear heating and the state of stress on faults.*J. Geophys. Res.,* 85, 6174-6184.

Scholz, C.H., Beavan, J. & Hanks, T.C. (1979). Frictional methamorphism, argon deplection and tectonic stress on the Alpine fault, New Zealand.*J. Geophys. Res.*, 84, 6770-6782.

Scott, J.S. & Breyer, H.I. (1953). Frictional fusion along a Himalayan thrust. *Proc. R. Soc. Edimburgh,* Sect. B. 65, 121-135.

Sibson, R.H. (1975). Generation of pseudotachylite by ancient seismic faulting. *Geophys.J.R. Astron. Soc.*, 43, 775-789.

Sibson, R.H.(1977). Fault rocks and fault mechanism.*J. Geol. Soc. Lond.*, 133, 191-213.

Sibson, R.H. (1978). Radiant flux as a guide to relative seismic efficiency. *Tectonophysics,* 51,39-46.

Sibson, R.H. (1980). Power dissipation and stress levels on faults is the upper crust. *J. Geophys. Res.*, 85, 6239-6247.

Soliva, J., Salel, J.F. & Brunel M. (1989). Shear deformation and emplacement of the gnessic Canigou thrust nappe (Eastern Pyrenees) *Geol. Mijnbouw,* 68, 357-366.

Soula, J.C., Debat, P., Deramond, J., Gucherau, J.Y., Lamouroux, C.& Pouget. F. (1986). Evolution structurale des ensembles métamorphiques des gnesis et des granitoides dans les Pyrénées Centrales.*Soc. Geol. France Bull.*, 8, 79-93

Spohn, T. (1980). Orogenic volcanism caused by thermal runaways? *Geophys. J. R. Astr. Soc.*, 62, 403-419.

Spray, J.G. (1987). Artificial generation of pseudotachylite using friccion welding apparatus:simulation of melting on a fault plane. *J. Struct. Geol.*, 9, 49-60.

Strong, D.F. & Hanmer, S.K. (1981). The leucogranites of southern Britany: origin by faulting, frictional heating, fluid flux and fractional melting.*Canadian Miner.*, 19, 163-176.

Teufel, L.W. & Logan, J.M. (1978). Effect of displacement rate on the real area of contact and temperatures generated during frictional sliding of Tennessee sandstone. *Pure. Appl. Geophys.*, 116, 840-865.

Theodore, T.G. (1.970). Petrogenesis of mylonites of high metamorphic grade in the peninsular ranges of southern California. *Geol. Soc. Amer. Bull.*, 81, 435-450.

Thompson, J.B. (1959). Local equilibrium in metasomatic processes.Researches in Geochemistry. *Wiley,* 427-451.

Thompson, J.B. (1970). Geochemical reaction and open systems. *Geochim. Cosmochim. Acta.*, 34, 529-551.

Torne, M., de Cabissole, B., Bayer, R., Casas, A., Daignieres, M. & Rivero, A. (1989). Gravity constraints on the deep structure of the belt along the ECORS profile.*Tectonophysics,* 165,105-116.

Tullis, J., Snoke, A.W. & Todd, V.R. (1982). Significance and petrogenesis of mylonitic rocks. *Geology*. 10, 227-230.

Turcotte, D.L. & Schubert, G. (1982).*Geodynamics. Aplication of Continuun Physics to Geological Problems.* Wiley. 450 pp.

van den Eeckhout, B. (1984). The Hospitalet mantled gneiss antiform (Central Pyrenees).*Swansea, England, Tectonic Studies Group.*, 88pp.

van den Eeckhout, B. (1986). A case study of a mantled gneiss antiform, the Hospitalet massif, Pyrenees (Andorra, France). *Geol. Ultraiectina,* 45, 196 pp.

Vauchez, A. (1978).Déformation naturelle par cisaillement d'un granite de Grande Kabylie (Algérie).*Tectonophysics*, 51, 57-81.

Verhoef, P.N.W., Vissers, R.L.M. & Zwart. H.J.(1984). A new interpretation of the structural and metamorphic history of the western Aston massif (Central Pyrenees, France).*Geol. Mijnbouw,* 63, 399-410.

Wallace, R.C. (1976). Partial fusion along the Alpine fault zone, New Zealand.*Geol.Soc. Am. Bull.*, 87, 1225-1228.

Weill, D.F. & Fyfe, W.S. (1964). The solubility of quartz in H2O in the range 1000-4000 bars and 400-500º C.*Geochim. Cosmochim. Acta,* 28, 1243-1255.

Zingg. A., Handy, M.R., Hunziker, J.C. & Schmid, S.M. (1990). Tectonometamorphic history of the Ivrea Zone and its relationship to the crustal evolution of the Southern Alps.*Tectonophysics,* 182, 169-192.

Rev. Acad. Ciencias. Zaragoza 45, (1990)

ORIGEN DE LOS GRANATES EN ROCAS VOLCANICAS INTERMEDIAS Y ACIDAS. REVISION DE LOS CRITERIOS DE DISCRIMINACION GENETICA.

AUQUE, L.F. y SANCHEZ CELA, V.

Area de Petrología y Geoquímica. Departamento de Geología. Facultad de Ciencias. Universidad de Zaragoza. 50009 ZARAGOZA.

ABSTRACT

Almandine garnets in dacitic-andesitic rocks are currently interpreted as from magmatic, xenolithic or restitic origin. In this work the textural-compositional criteria used for their genetical interpretation are analyzed and discussed.

Different nowadays interpretation on garnets of volcanics appear rather to be related to the ambiguity in the use of such criteria, that in many cases results in the interpretation either xenolith or magmatic origin by different authors for garnets of a same volcanic rock. On the other hand, papers indicating the coexistence of both genetical garnet types in the same rock are more and more frequent.

From all that, the use of such textural-compositional criteria ougth to be complemented and checked through overall geological-petrological study of volcanics.

1.INTRODUCCION.

El granate de composición predominantemente almandínica aparece asiduamente como constituyente mineralógico en distintos tipos de rocas ígneas intermedias y ácidas, y puede considerarse como uno de los elementos clave dentro del planteamiento de los modelos petrogenéticos correspondientes a estas rocas.

Un breve repaso a la bibliografía existente sobre el tema permite comprobar la existencia de este mineral en numerosas variedades de rocas ígneas calcoalcalinas, desde andesitas, dacitas, riodacitas y riolitas, hasta granitos, granodioritas y otras rocas plutónicas. Esta dispersión se traduce, por un lado, en el establecimiento de un amplio rango de condicionamientos físicos, químicos y geológicos que posibiliten la presencia de este mineral en tan variados grupos petrológicos; y por otro, en la definición de improntas genéticas muy similares para granates asociados a distintos procesos de formación.

Tres son las hipótesis genéticas planteadas para explicar la presencia de granates en rocas volcánicas calcoalcalinas:

a) La que propugna un origen magmático de este mineral y lo asocia a un proceso de cristalización durante la evolución del fundido ígneo.

b) La que considera un origen xenolítico, ajeno al proceso evolutivo de las rocas volcánicas, por incorporación de granates previamente formados en el encajante atravesado por el fundido durante su ascenso.

c) La que mantiene un origen restítico del granate como residuo refractario del proceso de fusión parcial de materiales corticales (pelíticos) al que se asocia el origen de las rocas ígneas consideradas.

La acepción para un caso concreto de una de las hipótesis que acabamos de señalar, se basa en una serie de caracteres de tipo estructural, textural y geoquímico que, aisladamente y a menudo también en conjunto, pueden resultar bastante ambiguos.

Esta situación ha llevado a frecuentes controversias, no ya a un nivel meramente conceptual, sino en lo que se refiere al estudio de zonas y afloramientos muy concretos. Además, es cada vez más frecuente la definición de granates genéticamente distintos coexistiendo en la misma roca (véase p.e. Clemens & Wall, 1984; Barley, 1987).

Esta posible coexistencia introduce una nueva visión de la cuestión, no exenta de problemas, y que está en relación directa tanto con el hecho de que el crecimiento magmático de los granates no está limitado a los dominios de altas presiones en la corteza, como con que en la mayoría de los fundidos, en los que geoquímicamente pueden crecer estos minerales, es factible también la conservación sin desequilibrios pronunciados de granates xenolíticos.

A lo largo de este trabajo nos referiremos a la problemática del granate en lo que se refiere a su presencia específica en rocas volcánicas. La discusión sobre el origen de este mineral en rocas graníticas puede realizarse en términos bastante similares a los que a continuación pasaremos a describir. De hecho, muchos de los criterios de discriminación son utilizables en ambos tipos de rocas y, además, los datos teóricos y experimentales correspondientes a los dos sistemas petrogenéticos pueden ser mutuamente extrapolados en determinadas condiciones.

Sin embargo, el análisis particular de la génesis de granates en rocas graníticas posee suficientes caracteres diferenciales respecto al de las rocas volcánicas como para constituir un trabajo con entidad propia, actualmente en vías de realización.

2.CARACTERISTICAS GENERALES DE LAS ROCAS VOLCANICAS ENCAJANTES.

Es bastante frecuente la presencia de granate con predominio de la molécula de almandino en distintas variedades de rocas volcánicas calcoalcalinas. Su distribución abarca desde términos composicionales andesíticos (tanto piroxénicos como anfibólicos o biotíticos) hasta riolíticos.

Pese a esta aparente aleatoriedad, un examen de algunos de los trabajos publicados hasta la fecha (Tabla 1) permite constatar la repetición, más o menos completa, de una serie de rasgos comunes asociados a este tipo de rocas volcánicas granatíferas.

En primer lugar, se trata de rocas con altas proporciones de K_2O y claramente peralumínicas (con corindón normativo), con la única excepción de las andesitas granatíferas de Checoslovaquia, con contenidos medios en K_2O y diópsido normativo. Estos caracteres se de las rocas volcánicas calcoalcalinas se repiten en todos los casos considerados (Tabla 1).

Otro de los rasgos comunes a este tipo de rocas es la existencia de enclaves metamórficos que incluyen, a su vez, granate en su asociación mineral constitutiva. La distribución, tipo y abundancia de estos enclaves es muy variable, y existen casos en los que no se evidencia esta relación (p. ej. las andesitas granatíferas de Checoslovaquia o las manifestaciones dacíticas y riolíticas del Pirineo). Sin embargo, este carácter constituye uno de los puntos fundamentales a considerar al hablar de la génesis de los cristales aislados de granate en la roca volcánica.

Tabla 1. Recopilación de trabajos sobre granates en rocas volcánicas intermedias-ácidas.

Situación	Tipo de roca volcánica	Edad	Caracteres geoquímicos	Enclaves	(Sr^{87}/Sr^{86})	Referencias
Canterbury (Nueva Zelanda)	andesitas dacitas y riolitas	Mioceno Cretácico	Rocas calcoalcalinas con alta proporción de K con corindón normativo (las dacitas con hiperstena normativa)	SI SI	 >.705	Batrum (1937) Wood (1974) Barley (1987)
SW de Japón	andesitas y dacitas	Mioceno	Rocas calcoalcalinas con alta proporción de K, peralumínicas (con corindón normativo).	SI		Tagiri et al. (1975) Ujike & Onuki (1976)
California (USA)	andesitas ácidas	Pleistoceno	Rocas calcoalcalinas con alta proporción de K, peralumínicas (con corindón normativo).	SI	>.705	Gill (1981)
Checoslovaquia	andesitas básicas y ácidas	Terciario	Rocas calcoalcalinas con proporción media de K, (con diópsido normativo).	NO	>.705	Brousse et al. (1972)
Crimea (USSR)	andesitas básicas y ácidas	Mesozoico	Rocas calcoalcalinas con alta proporción de K, peralumínicas (con corindón normativo).	SI		Makarov & Suprychev (1964) Gill (1981)
Kamchatka (USSR)	andesitas básicas y ácidas	Mesozoico	Rocas calcoalcalinas con alta proporción de K, peralumínicas (con corindón normativo).	SI		Gill (1981)
Victoria (Australia)	riolitas	Devónico	Rocas calcoalcalinas con alta proporción de K, peralumínicas (con corindón normativo).	SI	>.705	Green & Ringwood (1972) Birch & Gleadow (1974) Irving & Frey (1978) Clemens & Wall (1984)
Borrowdale (Inglaterra)	andesitas ácidas y dacitas	Ordovícico	Rocas calcoalcalinas con alta proporción de K, peralumínicas (con corindón normativo).	NO		Oliver (1956) Fitton (1972) Fitton et al. (1982) Thirlwall & Fitton (1983)
Macizo Central (Francia)	Riodacitas y riotraquitas	Carbonífero	Rocas calcoalcalinas con alta proporción de K, peralumínicas (con corindón normativo).	SI		Bertaux (1982)
Pirineos (Francia)	dacitas y riolitas	Estefaniense-Pérmico	Rocas calcoalcalinas con alta proporción de K, peralumínicas (con corindón normativo).	NO		Bixel (1988)
SE de España	dacitas y riolitas	Neógeno	Rocas calcoalcalinas con alta proporción de K, peralumínicas (con corindón normativo).	SI	>.705	Zeck (1968) López Ruiz et al. (1977) Massare (1979) Munksgaard (1984)
Cordillera Ibérica (España)	andesitas y dacitas	Estefaniense-Autuniense	Rocas calcoalcalinas con alta proporción de K, peralumínicas (con corindón normativo).	SI		Aparicio & García Cacho (1984) Auqué (1986) Lago et al. (1989) Sánchez Cela et al. (1990)

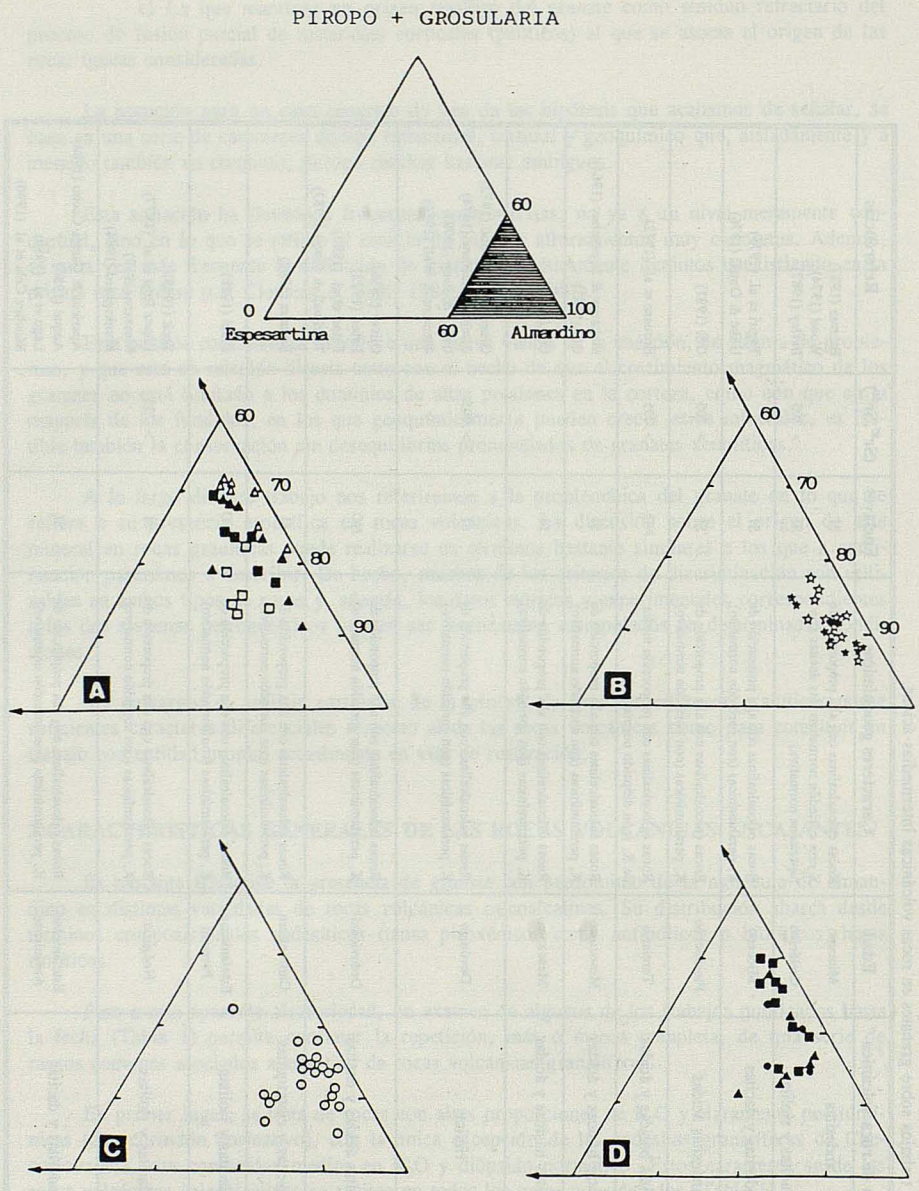


Figura 1. Representación del quimismo de distintos tipos de granates en diagramas piropo + grosularia-almandino-espesartina. A.-Granates en vulcanitas (△ borde; ▲ centro) y rocas metamórficas de la serie regional (□ borde; ■ centro) del Macizo Central Francés (Bertaux, 1982). B.- Granates aislados en andesitas y riolitas (☆) y en enclaves metamórficos (★) de esas mismas rocas (López Ruiz et al., 1977) en el SE de España. C.- Granates de la serie metamórfica del Sistema Central español (Aparicio y García Cacho, 1984). D.- Granates en andesitas (■) y en enclaves metamórficos (●) de Atienza; granates en enclaves (▲) de rocas daciandesíticas de Noguera (Teruel). Datos de la Cordillera Ibérica (Aparicio y García Cacho, 1984; Auqué, 1986).

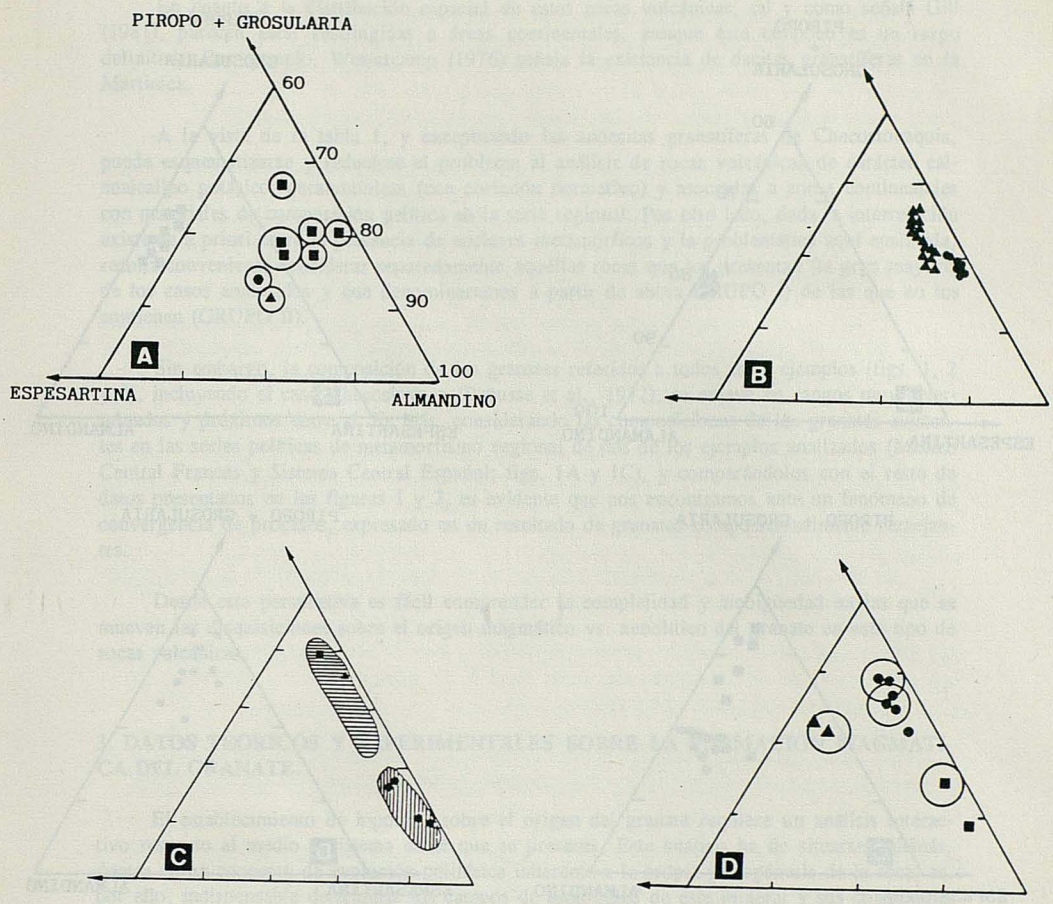


Figura 2. A.- Granates en rocas plutónicas (Allan & Clarke, 1981): ■ tipo I, cristales aislados de granate de origen xenolítico; ▲ tipo II, granates de origen magmático; ● tipo III, granates magmáticos tardíos. B.- Granates en dacitas (▲) y riolitas (●) pirenaicas (Bixel, 1988). C.- Granates en dacitas ▤ y riolitas ▥ (Barley, 1987): ● de origen magmático; de origen xenolítico: ■ granate en enclave incluído en dacitas; ▲ cristal aislado de granate, de origen xenolítico, en riodacita. D.- Granates en riolitas (Clemens & Wall, 1984): ● granates tipo A y D (magmáticos tempranos y xenolíticos con recrecimientos magmáticos); ■ granate tipo C, magmático tardío; ▲ granate en enclaves xenolíticos.

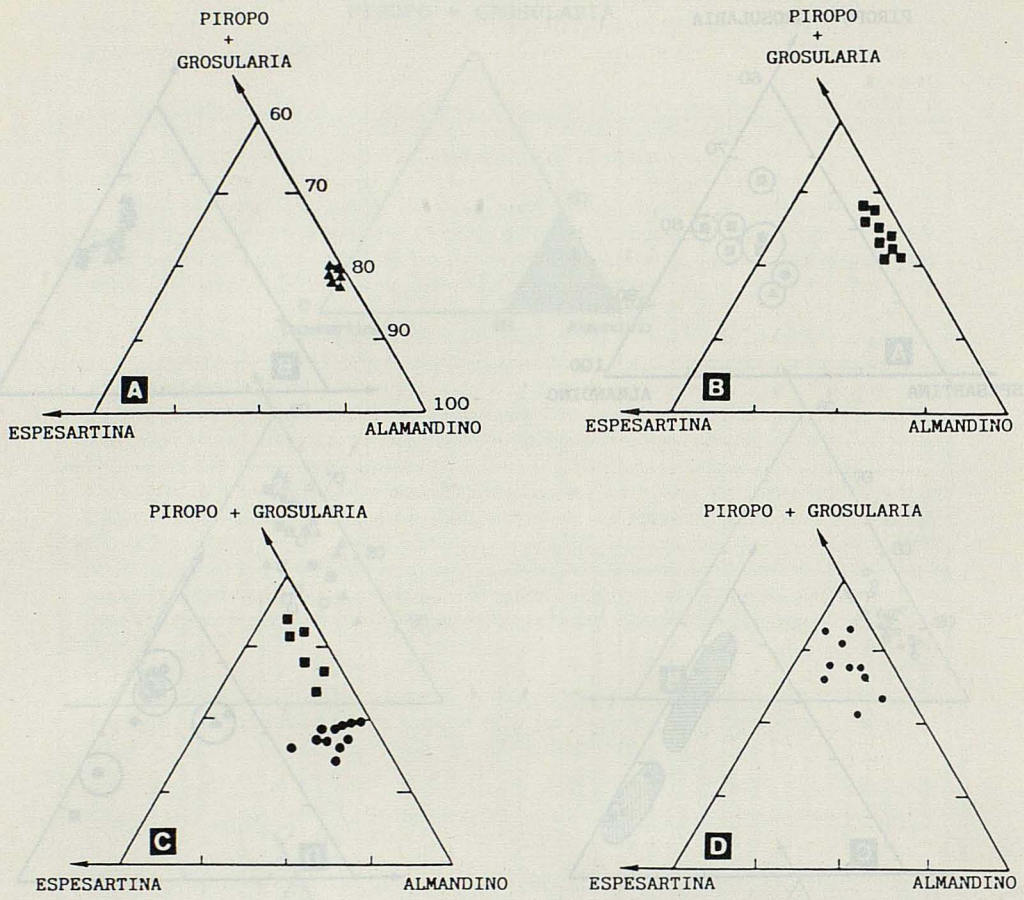


Figura 3. A.- Granates aislados en riolitas (Wood, 1974). B.- Granates en riodacitas (Green & Ringwood, 1968). C.- Granates en andesitas (■) y dacitas (●) (Oliver, 1956; Fitton, 1972). D.- Granates en andesitas-riolitas de Checoslovaquia (Brousse et al., 1972).

Independientemente de la problemática acerca de la influencia que la asimilación de enclaves ha podido tener en la modificación del quimismo de la roca volcánica encajante, los datos isotópicos de Sr^{87}/Sr^{86} que han podido obtenerse para las rocas en cuestión (Tabla 1) son siempre mayores de 0.705, lo cual parece sugerir que el carácter peralumínico puede proceder de la asimilación de materiales pelíticos (Gill, 1981). Por otro lado, en los estudios de Clemens & Wall (1984), Munksgaard (1984), Barley (1987), etc., se concluye que las dacitas y riolitas granatíferas de distintas zonas proceden de fenómenos anatécticos en los materiales pelíticos de las respectivas series regionales.

En cuanto a la distribución espacial de estas rocas volcánicas, tal y como señala Gill (1981), parecen estar restringidas a áreas continentales, aunque éste tampoco es un rasgo definitivo. Por ejemplo, Westercamp (1976) señala la existencia de dacitas granatíferas en la Martinica.

A la vista de la tabla 1, y exceptuando las andesitas granatíferas de Checoslovaquia, puede esquematizarse y reducirse el problema al análisis de rocas volcánicas de carácter calcoalcalino potásico, peralumínicas (con corindón normativo) y asociadas a zonas continentales con materiales de composición pelítica en la serie regional. Por otro lado, dada la interrelación existente a priori entre la presencia de enclaves metamórficos y la problemática aquí analizada, resulta conveniente considerar separadamente aquéllas rocas que los presentan (la gran mayoría de los casos analizados y que denominaremos a partir de ahora GRUPO I) de las que no los contienen (GRUPO II).

Sin embargo, la composición de los granates referidos a todos estos ejemplos (figs. 1, 2 y 3), incluyendo el caso Checoslovaco (Brousse et al., 1972), se mueve en rangos muy determinados y próximos entre sí. Es más, considerando las composiciones de los granates existentes en las series pelíticas de metamorfismo regional de dos de los ejemplos analizados (Macizo Central Francés y Sistema Central Español; figs. 1A y 1C), y comparándolos con el resto de datos presentados en las figuras 1 y 2, es evidente que nos encontramos ante un fenómeno de convergencia de procesos, expresado en un resultado de granates composicionalmente semejantes.

Desde esta perspectiva es fácil comprender la complejidad y ambigüedad en las que se mueven las disquisiciones sobre el origen magmático vs. xenolítico del granate en este tipo de rocas volcánicas.

3. DATOS TEORICOS Y EXPERIMENTALES SOBRE LA FORMACION MAGMATICA DEL GRANATE.

El establecimiento de hipótesis sobre el origen del granate requiere un análisis interactivo respecto al medio o sistema en el que se presenta. Este análisis ha de situarse, además, dentro de un concepto de evolución polifásica inherente a la propia petrogénesis de la roca; es, por ello, indispensable determinar los campos de estabilidad de este mineral y sus condiciones limitantes para una adecuada evaluación de su génesis.

Desde esta perspectiva, la evolución en el conocimiento de los sistemas teóricos y experimentales ha influído de manera determinante en los sucesivas extrapolaciones realizadas en el análisis de sistemas naturales.

Inicialmente, los datos experimentales de Green & Ringwood (1968, 1972) indicaban que el granate almandínico aparecía como fase cercana al líquidus en magmas de composición andesítica y dacítica (en un rango de P de 9-27 Kbars). Consideraban, por tanto, que tales fundidos se derivaban de procesos de fusión parcial en la corteza inferior o manto superior, a profundidades mayores de 25 km. Los procesos de desestabilización, asociados frecuentemente a granates, se contemplaban como producto de reequilibrio al cambiar las condiciones de presión del fundido durante su ascenso. De esta manera se explicaban los recrecimientos de cordierita e hiperstena (Miyashiro, 1955), biotita-feldespato (Oliver, 1956) o clorita-magnetita (Fitton, 1972).

Sin embargo, los granates cristalizados experimentalmente en las condiciones arriba señaladas resultaban sistemáticamente más ricos en Mg y Ca que los encontrados en las rocas volcánicas (Gill, 1981; Bertaux, 1982).

Los datos experimentales aportados por Green (1976 y 1977), al realizar fundidos a partir de vidrios de composición pelítica, comenzaron a sugerir que cierta proporción de Mn proviniente del material pelítico original podría estabilizar la formación de granate a bajas presiones. De esta manera, este autor estimó que granates con un 20-25% de espesartina podrían estabilizarse a presiones correspondientes a profundidades menores de 12 km.

Posteriormente, Miller & Stoddard (1981) definieron condiciones de formación menores de 3 Kb para granates espesartínicos con una proporción superior al 10% de esa molécula en su composición. Por otro lado, Clemens & Wall (1981), al estudiar la evolución de fundidos de tipo S, indicaron la posibilidad de cristalización del granate a P de hasta 1 Kb dependiendo de la relación Mg/(Mg+Fe) y del contenido en Mn del fundido.

El proceso de cristalización del granate abarca, pues, un amplio rango de condiciones fisicoquímicas y geológicas. Los caracteres diferenciales de los distintos condicionamientos genéticos van a manifestarse en pequeñas variaciones del quimismo de este mineral y en el desarrollo o no de un zonado composicional.

Sin embargo, y desde un punto de vista estrictamente magmático, el granate, una vez que cristaliza, está sujeto a la evolución posterior del fundido en el que se encuentra y, por tanto, a posibles desestabilizaciones o procesos retrógrados que modifiquen sus caracteres originales (p. ej. Green & Ringwood, 1968).

Por otro lado, si un proceso ígneo satisface en un momento determinado las condiciones necesarias para la cristalización de granate, también las posee para la conservación de un posible granate xenolítico de similar composición (Bertaux, 1982). Las condiciones de estabilidad diferenciales entre el fundido, granate magmático y granate xenolítico, pueden dar lugar a una compleja gama de situaciones que serán función del quimismo de los granates y de la evolución posterior del proceso magmático.

Por tanto, los estudios teóricos y experimentales, por sí mismos, no son un argumento de peso a la hora de decidir la génesis de los granates existentes en un afloramiento concreto.

4. REVISION DE CRITERIOS PARA LA DEFINICION GENETICA DEL GRANATE.

La revisión bibliográfica realizada sobre el tema pone de manifiesto la existencia de distintos tipos de criterios empleados en la definición de la génesis del granate. Básicamente podemos agruparlos en tres tipos fundamentales: criterios estructurales, texturales y composicionales.

En el primer grupo, definido a escala de afloramiento, podemos incluir la distribución (homogénea o heterogénea) de los cristales de granates en el cuerpo ígneo aflorante, la existencia o ausencia de litologías con granate en la serie regional atravesada por el mismo, la presencia o ausencia de enclaves (especialmente de tipo metamórfico) y la existencia o no de una relación espacial entre los granates y los posibles enclaves existentes en esas rocas.

El segundo grupo de criterios, los texturales, incluye el tipo de morfologías cristalinas de los granates existentes en la roca volcánica (anhedral/euhedral), la existencia o no de bordes de reacción en estos minerales, presencia o ausencia de inclusiones, tipo de inclusiones existentes, etc.

En los composicionales hemos incluído tanto los criterios que hacen referencia a la constitución química propia de los granates (composición media, borde-centro del cristal, existencia o no de zonado mineral y tipos específicos de este zonado), como a las pautas geoquímicas relacionales entre estos minerales y las roca volcánica que los engloba (acomodación o no de la constitución del granate a las pautas evolutivas Fe-Mg, La-Y, La-Lu del proceso volcánico), cuyo carácter calcoalcalino hemos señalado anteriormente.

Del análisis más o menos completo de este conjunto de criterios se han deducido hasta cuatro posibles situaciones en cuanto a la génesis de los granates en distintos tipos de rocas volcánicas intermedias y ácidas:

- Un origen magmático (véase p. ej. Green, 1976, 1977; López Ruiz et al., 1977).

- Un origen xenolítico (p. ej. Bertaux, 1982; Aparicio y García Cacho, 1984)

- Un origen restítico (p. ej. White & Chapell, 1977; Clemens & Wall, 1984).

- Varias de las génesis anteriores coexistiendo en la misma roca (p. ej. Birch & Gleadow, 1974; Clemens & Wall, 1984; Barley, 1987).

La tercera de las génesis propuesta, la restítica, supone un origen del granate como residuo refractario de un fenómeno de fusión parcial de materiales corticales, que da lugar a su vez al fenómeno volcánico. Esta hipótesis ha sido propuesta en casos en los que la existencia de enclaves metamórficos ha sido interpretada como restos (restitas) de ese proceso de fusión parcial y, por tanto, se enlaza directamente con la controversia existente entre la relación genética de estos enclaves y el proceso ígneo (Auque et al., 1987). Por ello, y desde un punto de vista estrictamente descriptivo, consideraremos esta génesis dentro del carácter xenolítico, ya que su definición requiere la utilización de criterios más ligados con la relación petrogenética entre el enclave metamórfico y la roca volcánica que con la específicamente referida al granate (Sánchez Cela et al., 1990).

4.1. Criterios estructurales.

Tal y como acabamos de ver, estos criterios suponen un análisis del problema a escala regional o de afloramiento. En general, la existencia de enclaves metamórficos con granate en las rocas volcánicas consideradas es indicativa, cuando menos, de la necesidad de un estudio muy detallado para la definición de la génesis de los granates aislados presentes en las vulcanitas.

La controversia fundamental de este tema se ha centrado casi siempre en rocas volcánicas que presentan este tipo de enclaves (Grupo I). Si nos fijamos en la tabla 1, en los ejemplos en los que no se manifiesta este carácter (únicamente existe cristales de granate aislados en la roca volcánica) ha sido definido un carácter magmático para el granate y, en general, son los casos donde los criterios composiconales (aquéllos de relación geoquímica entre granate y roca volcánica) parecen confirmar claramente esta hipótesis. La aplicación del resto de criterios a estos casos carece de sentido, bien por su ambigüedad intrínseca, bien por la ausencia de otros posibles caracteres que permitan inferir un caracter xenolítico y, por lo tanto, el otro término de comparación requerido.

Si a la presencia de enclaves metamórficos sumamos la existencia de relaciones en la distribución espacial y proporcional entre los granates y este tipo de xenolitos (Auqué, 1986; Lago et al., 1989) deberemos comenzar a considerar como hipótesis de trabajo que, al menos, parte de los granates existentes en la roca volcánica puedan tener un origen xenolítico.

En general, la existencia de granate en la serie regional atravesada por el proceso ígneo en su emplazamiento suele coincidir con la presencia de enclaves con granate. Sin embargo, e independientemente de la existencia de este tipo de xenolitos, es interesante comprobar la mineralogía de la serie regional, puesto que en el caso de encontrar litologías con granate será necesario determinar su composición para la aplicación de criterios más específicos (composicionales).

Como puede deducirse de todo lo anterior, los criterios estructurales suelen tener un valor indicativo importante, al menos en cuanto a la planificación y planteamiento previo de hipótesis de trabajo.

4.2. Criterios texturales.

Dentro de este tipo de criterios se incluyen caracteres con un valor discriminatorio importante junto a otros más ambigüos.

La abundancia de inclusiones minerales dentro de los cristales aislados de granate (por ejemplo, texturas poiquilíticas con biotita) en la roca volcánica ha sido calificada como más típica de crecimientos en estado sólido y por tanto indicativa de un origen xenolítico (Allen & Clarke, 1981). El criterio, así definido, no posee un carácter excluyente respecto al resto de posibles orígenes. En todo caso, su confirmación requeriría un análisis composicional de los minerales incluídos en el granate, los existentes en la roca volcánica o los presentes en la asociación mineral de los enclaves metamórficos si están también presentes.

Un criterio mucho más definitorio es la presencia en un cristal aislado de granate de inclusiones minerales específicamente relacionadas con la mineralogía existente en los enclaves metamórficos asociados (por ejemplo, sillimanita), o bien de texturas de tipo metamórfico (helicíticas); situaciones ambas indicativas de un origen xenolítico de ese granate (Clemens y Wall, 1984).

Otro criterio indicativo de un carácter xenolítico del granate puede venir definido por la persistencia de este mineral en los bordes de reacción que suelen presentar los enclaves metamórficos, y en los que el resto de la asociación mineral metamórfica típica ha sido destruída o transformada en el proceso de asimilación del enclave por el fenómeno ígneo (Auqué, 1986; Lago et al., 1989).

La morfología de los cristales y la presencia o ausencia de bordes de reacción constituyen criterios que únicamente deben servir de apoyo a otros más definitorios, ya que intrínsecamente no poseen más valor que el descriptivo. En general, los caracteres anhedrales y presencia de bordes de reacción en los granates existentes en una roca volcánica han sido interpretados de dos maneras:

- Si existen otros criterios que permiten presumir un carácter xenolítico de este mineral, ambos caracteres se relacionan con fenómenos de desestabilización producidos por los procesos ígneos al englobar un granate de origen, por tanto, xenolítico (p. ej. Bertaux, 1982).

- Si por el contrario se ha definido previamente la génesis magmática del granate por cualquier otra circustancia, las morfologías anhedrales y coronas de reacción se interpretan como desequilibrios generados por la evolución del fundido sobre los granates previamente cristalizados (p. ej. Green, 1976).

Evidentemente, puede llegarse a una situación similar en cuanto a indiscriminación al interpretar los cristales de granate euhedrales y sin bordes de reacción (ver, p. ej., la discusión de Hernán et al., 1981).

Pese a esta ambigüedad, la utilización de este tipo de criterios ha sido considerada especialmente válida en aquellos casos en los que se ha definido la coexistencia de dos génesis distintas de granate en la misma roca ígnea (tanto en estudios referidos a rocas volcánicas como graníticas; ver p. ej. Vennum & Meyer, 1979; Allan & Clarke, 1981; Clemens & Wall,

1984; Barley, 1987; etc.) Aunque, si bien es cierto que esta diferenciación se ha realizado de manera interactiva con criterios composicionales, no deja de presentar ciertas deficiencias tal y como veremos posteriormente.

4.3. Criterios composicionales.

Se trata de los argumentos a los que se ha otorgado una mayor capacidad de definición en los trabajos analizados. Tal y como se ha señalado al introducir este apartado, bajo la denominación de criterios composicionales incluímos tanto los que hacen referencia específica a la composición o pautas de zonado del granate como aquellos relativos a la mutua relación geoquímica con la evolución del proceso ígneo.

La utilización de estos dos grandes grupos de caracteres posee unos aspectos diferenciales a la hora de aplicarlos a casos concretos: mientras que los datos composicionales del granate o sus pautas de zonado adquieren su valor cuando existen términos comparativos distintos (granates aislados, en la serie regional, en enclaves metamórficos, etc., es decir, en los casos que hemos definido como Grupo I), los datos de caracteres geoquímicos interrelacionales granate/roca volcánica son válidos incluso en el caso de existir únicamente cristales aislados de granate en la roca volcánica (Grupo II).

De hecho, este último tipo de criterio (que podemos subclasificar como geoquímico) ha sido utilizado casi exclusivamente en las situaciones señaladas como pertenecientes al Grupo II, y no conocemos resultados de su aplicación a casos más complejos con evidente presencia de granates no magmáticos.

Los criterios geoquímicos consisten, básicamente, en el análisis comparado de las pautas evolutivas interactivas existentes entre el granate y el proceso ígneo como resultado de un fenómeno de cristalización fraccionada. Entre los caracteres incluídos en la aplicación de estos criterios se encuentra el análisis de la evolución La/Y de la serie magmática a la que se asocian las rocas volcánicas con granate (p. ej. Fitton, 1972; Fitton et al., 1982), el análisis del coeficiente de reparto de tierras raras (REE) a partir de los contenidos La-Lu del granate y de la roca que lo incluye (p. ej. Irving & Frey, 1978; Gilbert & Rogers, 1989), relación Fe-Mg en granates y biotitas de la roca volcánica (Aparicio & García Cacho, 1984), etc.

La relación La/Y de la serie considerada tiende a aumentar en los líquidos residuales si se produce una cristalización fraccionada con términos granatíferos, ya que los coeficientes de reparto de estos elementos en relación con el granate indican que este mineral va a concentrar REE pesadas e Y, a la vez que rechaza REE ligeras (Fitton et al., 1982). De manera similar, y en base al mismo principio, pueden considerarse las relaciones La-Lu del granate y de la roca volcánica encajante para comprobar si los coeficientes de reparto de estos elementos (Irving & Frey, 1978; Gilbert & Rogers, 1989) indican una génesis por cristalización fraccionada del granate.

Si bien la existencia de un enriquecimiento en REE ligeras y empobrecimiento en REE pesadas de la roca respecto al granate , es indicativa del origen magmático de este mineral, el caso contrario no necesariamente conlleva a un origen xenolítico. Simplemente lo que puede deducirse en esta situación es que el granate no procede de un proceso de cristalización fraccionada a partir del fundido original, pero puede derivar de modificaciones posteriores del mismo por fenómenos de asimilación o contaminación y, por tanto, poseer un caracter magmático (Fitton, 1972; Fitton et al., 1982).

Por otro lado, a partir del estudio de los coeficientes de reparto La-Lu entre roca volcánica y granate, puede llegar a inferirse la intervención de este mineral como residuo refractario (restita) del proceso de fusión generador del fenómeno volcánico. Este es el caso del estudio de Gilbert & Rogers (1989) en riolitas pirenaicas donde, curiosamente, sólo aparece una agrupación composicional de granate y no existen enclaves metamórficos. Aunque los granates aislados en el seno de la riolita pueden proceder de un proceso de cristalización fraccionada, existe la posibilidad de que este magma hubiera incorporado granates restíticos (y por lo tanto asimilables a xenolíticos), por lo que habría que suponer que el punto de partida original del proceso y su posterior evolución han conseguido separarse espacialmente.

La aplicación de este tipo de criterios, tal y como se ha señalado anteriormente, se ha restringido hasta el momento a casos asimilables a los que hemos denominado de Grupo II (p. ej. Irving & Frey, 1978). No conocemos el resultado de su utilización en casos más complejos como los presentados, por ejemplo, por Clemens & Wall (1984) o Barley (1987), en los que se indica la coexistencia de granates xenolíticos y magmáticos.

Otro argumento de similar filosofía a los anteriores puede utilizarse en el caso de existencia de granates en distintos términos composicionales de una serie magmática (naturalmente en una determinada zona). En esta situación, si los granates de cada tipo rocoso, por ejemplo en dacitas y riolitas o en andesitas y dacitas, presentan caracteres composicionales distintivos, en elementos mayores y/o menores (ver fig. 3C), puede deducirse que este rasgo se encuentra ligado directamente a los procesos genéticos diferenciales de andesitas, dacitas y/o riolitas, y por tanto, poseer un carácter magmático.

Esta misma situación puede visualizarse en los casos recogidos por Gill (1981, fig. 6.4., pág. 187) y con algo menos de claridad en los granates existentes en dacitas y riolitas del Pirineo (fig. 2B) según los datos presentados por Bixel (1988). En todos ellos puede observarse que la composición de los granates contenidos en andesitas son más ricos en Mg y más pobres en Fe y Mn que los existentes en dacitas y riolitas, debido a las mayores relaciones Mg/Fe y mayores temperaturas de las rocas andesíticas frente a los términos más diferenciados (Brousse et al., 1972).

De la misma manera que los anteriores criterios, éste último ha sido referido a situaciones del grupo II o bien a casos en los que no se analizan expresamente los granates existentes en enclaves metamórficos (Gill, 1981). Es de destacar que en el estudio realizado por Barley (1987) aparece una tendencia similar entre granates existentes en dacitas y riolitas pero con la salvedad de que los granates en las dacitas tienen un carácter xenolítico (definido por su similitud composicional con los granates presentes en los enclaves metamórficos incluídos en estas rocas, fig. 2C) mientras que en las riolitas presenta un carácter magmático.

Los criterios composicionales aplicados a casos más complejos que los hasta ahora señalados (en general con presencia de enclaves metamórficos) se limitan a la comparación de los análisis químicos realizados por microsonda de granates en distintas situaciones o con distintos caracteres, obteniéndose resultados dispares en cuanto a la génesis de los granates considerados.

Los distintos autores que han trabajado sobre el tema han deducido de este tipo de análisis comparativo tanto orígenes xenolíticos (Bertaux, 1982; Aparicio y García Cacho, 1984; Auqué, 1986; ver fig. 1) como coexistencia de granates de origen xenolítico y magmático (Clemens & Wall, 1984; Barley, 1987; ver figs. 2C y 2D) interrelacionando criterios composicionales y texturales.

En cuanto al zonado de los granates como criterio discriminatorio de su origen existe una importante controversia en cuanto a su interpretación. Desde el punto de vista de que tanto los granates magmáticos como los metamórficos pueden poseer cualquier tipo de zonado (o no poseerlo) o, incluso, a menor escala, ser el Mn un elemento crítico y variable tanto en zona-

dos metamórficos como magmáticos (Yardley 1977; Gill, 1981), puede comprenderse la dificultad de aplicar este planteamiento a casos complejos con granates aislados, enclaves metamórficos y granates en la serie regional.

Repasando la bibliografía puede verse cómo zonados diferenciales entre los granates aislados y los presentes en enclaves metamórficos y/o serie regional han servido para definir tanto:

- Un origen magmático de los granates aislados (ver p. ej. López Ruiz et al., 1977) o bien la coexistencia de granates magmáticos y xenolíticos en la misma roca (p. ej. Birch & Gleadow, 1974).

- Un origen enteramente xenolítico, con modificaciones más o menos acusadas en los granates aislados respecto a los presentes en los enclaves metamórficos o serie regional debido a la asimilación de aquéllos por el proceso volcánico (Bertaux, 1982; fig. 1A).

Gran parte de la polémica se centra en que los rasgos más diferenciales del zonado de granates en distintas situaciones se encuentran asociados a la zona periférica del mineral y, por tanto, sujetos a interpretaciones contrapuestas semejantes a las mencionadas al hablar de las coronas de reacción en los granates (véase la discusión de Bertaux, 1982).

Si bien esta situación puede pensarse que alcanza el máximo grado de complejidad en el caso de que todos los granates (los que aparecen como cristales aislados y como constituyentes de enclaves o de la serie regional) estén zonados diferencialmente, tampoco presenta una mayor facilidad de interpretación el que aparezca el zonado únicamente en alguna de las formas de manifestarse este mineral. De esta manera:

- Si los granates aislados no están zonados y sí lo están los existentes en los enclaves o en la serie regional, puede interpretarse que los primeros son xenocristales que al ser englobados por el proceso volcánico, y debido a la influencia que tiene la temperatura sobre las propiedades de difusión química en este mineral (Yardley, 1977), sufren un proceso de homogeinización (ver p. ej. Aparicio & García Cacho, 1984). Sin embargo, también es factible pensar que los granates aislados no zonados hayan cristalizado en condiciones de P constante y que el proceso de ascenso posterior del fundido haya sido rápido (ver p.ej. Gilbert & Rogers, 1989).

- En el caso de que los granates aislados estén zonados y no lo estén los incluídos en enclaves o en la serie regional puede interpretarse un origen magmático de los primeros ya que no es factible una reorganización química del granate xenolítico al ser incluído en el magma para explicar la adquisición del zonado. Sin embargo, puede suceder que un granate xenolítico haya servido de núcleo para un crecimiento posterior de tipo magmático con desarrollo de zonado (ver p. ej. Clemens & Wall, 1984).

5. CONSIDERACIONES FINALES.

En esta revisión sobre la problemática del origen del granate en rocas volcánicas calcoalcalinas (intermedias-ácidas) hemos pretendido situar al lector frente al mayor número de criterios y caracteres, así como a las dificultades inherentes a su interpretación. Quizás por ello alguno de ellos no haya sido tratado en la extensión necesaria, aunque esperamos que las abundantes citas bibliográficas subsanen este defecto.

De este análisis general pueden obtenerse algunas consideraciones acerca de la situación actual de la cuestión. En general, los rangos composicionales obtenidos para los distintos tipos posibles de granate han de tratarse siempre de una manera interrelacionada con otro tipo de caracteres. Esta situación suele llevar asociada una cierta subjetividad a la hora de dividir o clasificar grupos composicionales, mayor cuanto menor sea el número de datos analíticos disponibles.

Por otro lado, la utilización de distintos tipos de criterios de manera interactiva ha hecho aparecer frecuentemente análisis en los que se obtiene como resultado la existencia de granates con distintos orígenes en la misma roca. Esta situación se hace perfectamente comprensible si pensamos que el problema fundamental es que estamos tratando con procesos convergentes superpuestos, que pueden dar lugar incluso a situaciones de recrecimientos magmáticos de granate a partir de cristales xenolíticos de este mineral (granates tipo D de Clemens & Wall, 1984 y tipo B de Barley, 1987; figs. 2C y 2D).

Esta superposición de procesos con resultados similares es la causa fundamental de las dificultades existentes en la aplicación de los distintos criterios definidos. El mismo efecto puede obtenerse a partir de distintas causas y, por tanto, la hipótesis planteada debe estar basada en el conocimiento del mayor número de parámetros posible del sistema involucrado.

En este sentido no conviene olvidar que algunos de los argumentos utilizados (por ejemplo, la mayoría de los geoquímicos) han sido comprobados parcialmente en casos muy determinados y aparentemente sencillos. Por ello necesitan una validación y análisis más amplios en orden a determinar sus condicionamientos reales de aplicación general.

Por otro lado, el resto de criterios composicionales (composición global, zonado, etc.), así como muchos de los texturales, resultan intrínsecamente ambiguos, y gran parte de la controversia existente en este tema ha nacido de su exclusiva aplicación a determinados casos.

La presencia o ausencia de enclaves metamórficos, como elemento asociado a esta problemática, constituye un argumento fundamental a la hora de plantear la correspondiente hipótesis genética del granate. Es sintomático que en aquellos casos en los que no aparecían este tipo de enclaves se ha concluído un origen magmático para este mineral. Y de manera opuesta, en situaciones con presencia de enclaves metamórficos, se ha establecido un origen parcial o totalmente xenolítico.

La depuración en la metodología de estudio y en los criterios utilizados ha provocado frecuentes redefiniciones en las interpretaciones de los mismos casos analizados por diferentes autores. Este hecho puede observarse si comparamos los trabajos de Wood (1974) y Barley (1987) en Mt. Somers (Nueva Zelanda): mientras que para el primero los granates existentes en las riolitas eran de origen magmático, para el segundo coexisten granates magmáticos y xenolíticos.

Un ejemplo todavía más claro lo tenemos en la interpretación de los granates existentes en las rocas calcoalcalinas intermedias y ácidas del SE de España. Así Zeck (1968) considera los granates existentes en las rocas dacíticas como de origen xenolítico; López Ruiz et al. (1977) como de origen magmático, y por último, Massare (1979) y Bertaux (1982) como de origen xenolítico con recrecimientos magmáticos.

A la vista de estos casos es aconsejable, al abordar esta problemática, el análisis del mayor número de criterios posibles. Esta necesidad viene directamente definida por el carácter ambiguo de muchos de estos criterios analizados aisladamente. Por otro lado, es necesario un mayor análisis del carácter discriminatorio de criterios fundamentalmente geoquímicos, en orden a establecer sus capacidades en los casos más complejos. Las rocas dacítico-andesíticas con enclaves metamórficos (como las del SE de España y Cadena Ibérica) en las que puede establecerse una comparación entre los granates de la roca volcánica y los de los enclaves, pueden constituir una buena base de testificación.

BIBLIOGRAFIA.

ALLAN, B.D. & CLARKE, D.B. (1981). Occurence and origin of garnets in the South Mountain batholith, Nova Scotia, Canada. *Canadian Mineral., 19, 25-34.*

APARICIO, A. y GARCIA CACHO, L. (1984). Quimismo de los principales componentes minerales de las rocas volcánicas paleozoicas del Area de Atienza (Prov. de Guadalajara). *Bol. Geol. Min., t. XCV-I, 80-89.*

AUQUE, L.F. (1986). *Las rocas volcánicas de Noguera de Albarracín (Teruel) y sus enclaves metamórficos.* Tesis de Licenciatura, Universidad de Zaragoza, 315 pp. (No publicada).

AUQUE, L.F.; SANCHEZ CELA, V. y APARICIO, A. (1987). Enclaves con espinela-corindón--sillimanita en rocas andesítico-dacíticas (Noguera, Sierra de Albarracín, Teruel). *Estudios Geol., 43, 139-147.*

BARLEY, M.E. (1987). Origin and Evolution of Mid-Cretaceous, Garnet-bearing, Intermediate and Silicic Volcanics from Canterbury, New Zealand. *J. Volcanol. Geother. Res., 32, 247-267.*

BATRUM, J.A. (1937). Interesting xenoliths from Whangarei Heads, Auckland, New Zealand. *Trans. R. Soc. New Zealand, 67, 251-280.*

BERTAUX, J. (1982). Origine métamorphique des grenats des volcanites acides d'âge viséen supérieur dans le nord-est du Massif Central Francais. *Bull. Minéral., 105, 212-222.*

BIRCH, W.D. & GLEADOW, A.J.W. (1974). The genesis of garnet and cordierite in acid volcanics rocks: evidence from the Cerberean Cauldron, Central Victoria, Australia. *Contrib. Mineral. Petrol., 45, 1-13.*

BIXEL, F. (1988). Le volcanisme stéphano-permien des Pyrénées Atlantiques. *Bull. Centres Rech. Explor-Prod. Elf-Aquitaine, 12, 661-706.*

BROUSSE, R.; BIZOUARD, H. et SALAT, J. (1972). Grenats des andésites et des rhyolites de Slovaquie, origine des grenats dans les séries andésitiques. *Contrib. Mineral. Petrol., 35, 201-213.*

CLEMENS, J.D. & WALL, V.J. (1981). Origin and crystallization at some peraluminous (S-type) granitic magmas. *Canadian Mineral., 19, 111-131.*

CLEMENS, J.D. & WALL, V.J. (1984). Origin and evolution of peraluminous silicic ignimbrite suite: the Violet Town Volcanics. *Contrib. MIneral. Petrol., 88, 354-371.*

FITTON, G. (1972). The genetic significance of almandine-pyrope phenocrysts in the calc-alkaline Borrowdale volcanic group, Northern England. *Contrib. MIneral. Petrol., 36, 231-248.*

FITTON, J.G.; THIRLWALL, M.F. & HUGHES, D.J. (1982). *Volcanism in the Caledonian orogenic belt of Britain.* In: Andesites. Orogenic andesites and related rocks (R.S. Thorpe, ed.). John Wiley & Sons, pp. 611-638.

GILBERT, J.S. & ROGERS, N.W. (1989). The significance of garnet in Permo-Carboniferous volcanic rocks of the Pyrenees. *J. Geol. Soc. London, 146, 477-490.*

GILL, J. (1981). *Orogenic Andesites and Plate Tectonics.* Springer-Verlag, 390 pp.

GREEN, T.H. (1976). Experimental generation of cordierite-or garnet-bearing granitic liquids from a pelitic composition. *Geology, 4, 85-88.*

GREEN, T.H. (1977). Garnet in silicic liquids and its possible use as a P-T indicator. *Contrib. Mineral. Petrol., 65, 59-67.*

GREEN, T.H. & RINGWOOD, A.E. (1968).Origin of garnet phenocrysts in Calc-Alkaline rocks. *Contrib. Mineral. Petrol., 18, 163-174.*

GREEN, T.H. & RINGWOOD, A.E. (1972). Crystallization of garnet-bearing rhyodacite under high pressure hydrous conditions. *J. Geol. Soc. Aust., 19, 203-212.*

HERNAN, F.; PERNI, A. y ANCOCHEA, E. (1981). El volcanismo del Area de Atienza. Estudio Petrológico. *Estudios Geol., 37, 13-25.*

IRVING, A.J. & FREY, F.A. (1978). Distribution of trace elements between garnet megacrysts and host volcanic liquids of kimberlitic to rhyolitic composition. *Geochim. Cosmochim. Acta, 42, 771-787.*

LAGO,M.; TORRES, J.A.; AUQUE, L.F.; BAMBO, C.; HIDALGO, M.A. y POCOVI, A. (1989). Caracteres composicionales de granates en rocas calcoalcalinas, stephaniense-pérmicas, del sector de la Depresión Axial del Cámaras y el Anticlinal de Montalbán (prov. de Zaragoza y Teruel). *Estudios Geol., in press.*

LOPEZ RUIZ, J.L.; BADIOLA, E.R. et GARCIA CACHO, L. (1977). Origine des grenats des roches calco-alcalines du Sud-Est de l'Espagne. *Bull. Volcanol., 40, 1-12.*

MAKAROV, N.N. & SUPRICHEV, V.A. (1964). Xenogenic garnet (pyrope-almandine) from volcanic rocks of the Crimea. *Dokl. Akad.Nauk. SSR, 157, 64-67.*

MASSARE, D. (1979). *Etude des inclusions vitreuses de quelques minéraux de roches volcaniques acides: thermométrie, barométrie, compositions chimiques et éléments volatifs dissous.* Thèse 3 cycle, Paris.

MILLER, C.F. & STODDARD, E.F. (1981). The role of manganese in the paragenesis of magmatic garnet: an example from the Old Woman-Puite Range, California. *J. Geol., 89, 233-246.*

MIYASHIRO, A. (1955). Pyralspite granets in volcanic rocks. *J. Geol. Soc., Japan, 61, 463-470.*

MUNKSGAARD, N.C. (1984). High ^{18}O and possible pre-eruptional Rb-Sr isochrons in cordierite-bearing Neogene volcanics from SE Spain. *Contrib. Mineral. Petrol., 87, 351-358.*

OLIVER, R.L. (1956). The origin of garnets in the Borrowdale Volcanic series ans associated rocks, English Lake District. *Geol. Mag., 93, 121-139.*

SANCHEZ CELA, V.; AUQUE, L.F. & LAPUENTE, M.P. (1990). Petrological significance of high T-P metamorphic enclaves in dacitic-andesitic rocks. In L. Farrell (ed.): *High Grade Metamorphics.* Theophrastus Publ., Athens, 1990 (in press).

STONE, M. (1988). The significance of almandine garnets in the Lundy and Dartmoor granites. *Mineral. Mag., 52, 651-658.*

TAGIRI, M.; ONUKI, H. & YAMAZAKI, T. (1975). Mineral paragenesis of argillaceous xenoliths in andesite rocks from Nijo-san and Amataki-yama districts, SW Japan. *J. Jpn. Assoc. Mineral. Pet. Econ. Geol., 70, 305-314.*

THIRLWALL, M. & FITTON, J.G. (1983). Sm-Nd garnet age for the Ordovician Borrowdale Volcanic Group, English Lake District. *Jour. Geol. Soc. London, 140, 511-518.*

UJIKE, O. & UNUKI, H. (1976). Phenocrystic hornblende from Tertiary andesites and dacites, Kagawa Prefecture, Japan. *J. Jpn. Assoc. Mineral. Pet. Econ. Geol., 71, 389-399.*

VENNUM, W.R. & MEYER, C.E. (1979). Plutonic granets from the Werner batholith, Lassiter Coast, Antartic Peninsula. *Am. Mineral., 64, 268-273.*

WESTERCAMP, D. (1976). Pétrologie de la dacite á grenat de Gros Ilet, Martinique, Petites Antilles francaises. *Bull. B.R.G.M., 2 serie, Section IV, 4, 253-265.*

WHITE, A.J.R. & CHAPPEL, B.W. (1977). Ultrametamorphism and granitoid genesis. *Tectonophysiscs, 43, 7-22.*

WOOD, C.P. (1974). Petrogenesis of Garnet-bearing Rhyolites from Canterbury, New Zealand. *N.Z.J. Geol. Geophys., 17, 759-788.*

YARDLEY, B.W.D. (1977). An empirical study of diffusion in garnet. *Am. Mineral., 62, 793-800.*

ZECK, J.P. (1968). *Anatectic origin and further petrogenesis of almandine-bearing biotite-cordierite-labradorite-dacite with many inclusions of restite and basaltic material. Cerro del Hoyazo, SE Spain.* Thesis. Amsterdam.

Rev. Acad. Ciencias. Zaragoza 45, (1990)

NOTAS ECOLOGICAS DEL RIO GRIO

H. Marco

Departamento de Bioquímica y Biología Molecular
y Celular. Facultad de Ciencias Químicas.
Ciudad Universitaria. 50009 ZARAGOZA (España).

Abstract.: In this paper is exposed a study physical-chemical and bacteriological of Grio river water, subjected to great fluctuations on its fluvial regimen, even completely dry sometime during the year. It is described the microflora and microfauna detected in Grio river.

1. INTRODUCCION.

El presente trabajo es continuación de un estudio ecológico que estamos realizando en cursos fluviales y aguas estancadas de Aragón.

El río Grío es un afluente del río Jalón, que tiene su nacimiento en la Sierra de Algairen, en el pico Atalaya, con una cota de 1.235 metros sobre el nivel del mar.

Este río ofrece normalmente un caudal muy reducido, con estiajes durante los meses de verano, que en ocasiones dejan su cauce totalmente seco, que discurre a lo largo de una depresión tectónica situada en la citada sierra, así como en la Sierra de Vicort. Pasa por un estrecho valle, llegando al norte de la Almunia de Doña Godina, vertiendo finalmente sus aguas en el río Jalón, en las afueras de Ricla (Fig. 1). Como puede apreciarse en el plano, el río Grío inicia su recorrido en el Campo de Daroca, atraviesa los Campos de Cariñena, Calatayud, terminando en el Campo de la Alumnia de Doña Godina.

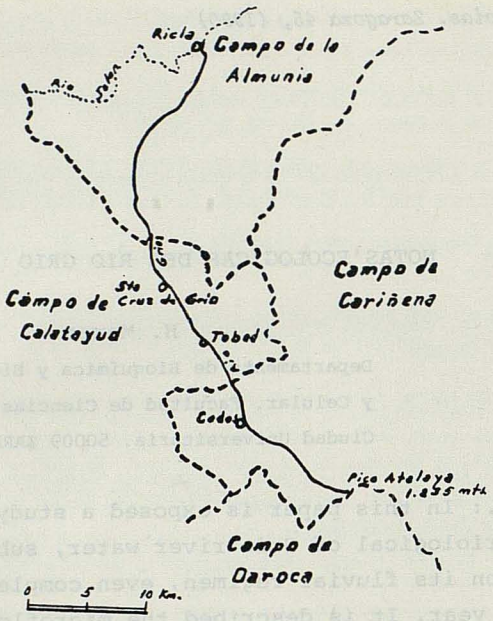


Fig. 1.- Ubicación hidrográfica del río Grío.

2. MATERIAL Y METODOS DE TRABAJO.

2.1.- Campo de trabajo.

Se han elegido como estaciones de muestreo las que a continuación se citan:

Estación de Codos

En las cercanías del pueblo, a una altitud de 751 metros, con una precipitación anual de 751 mm. La población vegetal predominante está representada por el Quercus ilex, Quercus pirenaica, Pinus Pinaster y Pinus halepensis. Es zona de cultivo de cereales, así como de frutales: almendro, manzano y peral.

Estación de Tobed

Este pueblo se encuentra situado al pie de la Sierra de Vicor, se ubica a 637 metros de altitud, con una precipitación anual de 500 mm. Vegetación y cultivos semejantes a los indicados en Codos.

Estación de Santa Cruz de Grío.

Esta localidad se encuentra situada igualmente al pie de la Sierra de Vicor, a una altitud de 712 metros y una precipitación

media anual estimada en 500 mm. La vegetación y cultivos son semejantes a los de las estaciones anteriores, salvo un cultivo preferente de fresones.

Estación de Mula Roya

Se trata de un Parque, protegido por la Dirección forestal, con una repoblación dominante de Pinus halepensis. Las aguas del río Grío están controladas por una presa recién construida, que permite conservar el agua durante los estiajes y muy especialmente mediante el aporte de un manantial cercano.

2.2.- Técnicas de trabajo

Durante los años 1989 y 1990 se han practicado muestreos mensuales, para así obtener los valores medios correspondientes a parámetros físico-químicos, basteriológicos y determinanción de la microflora y microfauna reinantes.

Los análisis químicos se han realizado en los Laboratorios CONTA S.A., de Zaragoza.

En la identificación de la microflora y microfauna, se han consultado los trabajos de Margalef (1953, 1955, 1983), Dosset y Monzon (1888), Cámara Niño (1951), Kiefer y Fryer (1978) y Bick (1972).

3. OBSERVACIONES Y RESULTADOS.

3.1.- Parámetros físico-químicos.

Por los datos expuestos en la Tabla I, se aprecia un elevado contenido de oxígeno, explicable por el hecho de que,en las épocas de riadas relativamente intensas, las aguas se airean en forma apreciable, puesto que en un tramo de unos 41 kms. se desciende desde 1.200 metros en su nacimiento hasta 347 metros en la localidad de Ricla, lugar en que el río Grío vierte sus aguas en el río Jalón.

Temperatura en Cº. . . .	13
DBO	<10 mg/l
Oxígeno disuelto. . . .	9 mg/l
pH	7

Tabla I.-Valor medio de los parámetros físico-químicos del río Grío.Años 1989-90

El análisis químico de estas aguas, nos ha proporcionado los siguientes valores medios, que se exponen en la Tabla II.

CO_3Ca	208 mg/l
Sulfatos	107 mg/l
Cloruros	28,5 mg/l
Nitratos	71,4 mg/l
Nitritos	0,094 mg/l
Bicarbonatos	179,4 mg/l
Calcio	60,9 mg/l
Magnesio	13,6 mg/l
Potasio	1,8 mg/l
Sodio	9,2 mg/l
Conductividad eléctrica	337 $\mu S\ cm^{-1}$

Tabla II.- Valor medio de la composicion química de las aguas del río Grío.Años 1989-90

Como consecuencia de estos resultados, se observa un elevado porcentaje en carbonatos y sulfatos, indicativos del tipo de terrenos por los que discurre el cauce del río Grío, que son depósitos yesíferos del terciario así como materiales calcáreos del cuaternario, por lo que son aguas de alta dureza, corroborada por el elevado valor de la conductividad eléctrica, del orden de 337 $nS.cm^{-1}$ puesto que cuando se trata de aguas muy puras, los valores de la conductividad son siempre bajos, como pudimos apreciar en un trabajo publicado acerca del río Huecha (Marco Moll, 1988).

Teniendo en cuenta la posible contaminación, que por desgracia aqueja en la actualidad a la mayor parte de los ríos españoles, especialmente por el abusivo empleo de productos químicos en la agricultura, se analizó la posible presencia de determinados oligoelementos que pueden ser perjudiciales y que se exponen en la Tabla III:

Hierro	$<0,5$ mcg/l
Cadmio	<1 mcg/l
Mercurio	$<0,1$ mcg/l
Zinc	<5 mcg/l
Manganeso	$<0,5$ mcg/l

Tabla III.- Valor medio de oligoelementos detectados en las aguas del río Grío. Años 1989-90

Como puede apreciarse, el contenido de estos metales se encuentran por debajo de los valores establecidos como marcadamente contaminantes.

3.2.- Observaciones bacteriológicas.

En el aspecto bacteriológico, según los datos aportados en la Tabla IV, nos indica que el número de coliformes totales sobrepasan el valor máximo, establecido según normas de la Sanidad de un 10 % ml., ya que aparecen con una cifra de 92 % ml., que justamente con el Streptocccus fecalis, sobrepasan el valor máximo

establecido, que es de un 10 % ml., deduciéndose con ello que estas aguas se encuentran fuertemente contaminadas, no debiendo ser utilizadas para el consumo humano.

	resultado	unidades
Recuento aerobios a 37 Cº. . .	291	1 ml
Coliformes totales	92	100 ml
Coliformes fecales	9	100 ml
Streptococcus fecalis.	21	100 ml
Clostridium sulfito-reductores	2	20 ml

Tabla IV.- Analisis bacteriologico de las aguas del río Grío. Valores medios correspondientes a los años 1989-90

3.3.- Microflora característica.

A continuación se describen las formas vegetales pertenecientes al grupo de las Algas, que han sido destectadas a lo largo de dos años de observaciones.

Cianoficeas

Género ANABAENA

Muy esporadicamente detectada la Anabaena flos-aquae.

Género CHROOCOCCUS

Durante el mes de septiembre aparecen masas constituidas por agrupaciones de ocho células correspondientes a la especie Chroococcus minimus.

Género GLOEOCAPSA

Aparece como muy abundante la especie Gloeocapsa juliana, con células de 4 micras. En menor presencia Gloeocapsa sp., en colonias de dos células.

Género MERISMOPEDIA

Ha sido detectada la presencia de Merismopedia punctata en colonias de 14 x 10 micras.

Género OSCILLATORIA

Este género es bastante abundante, habiéndose detectado las siguientes especies: O. splendida, O.irrigua, O. brevis y O. tenuis.Las especies O. splendida y O. irrigua aparecen con mucha constancia en las cuencas fluviales que han sido estudiadas en Aragón.

Género SPIRULLINA

Muy esporadicamente se detecta la presencia de la Spirullina minor.

Diatomeas

Este grupo se encuentra ampliamente representado a lo largo

del curso del río Grío. La relación de especies identificadas son practicamente las mismas que hemos descrito en el trabajo que publicamos acerca del río Huecha (Marco Moll, 1988).

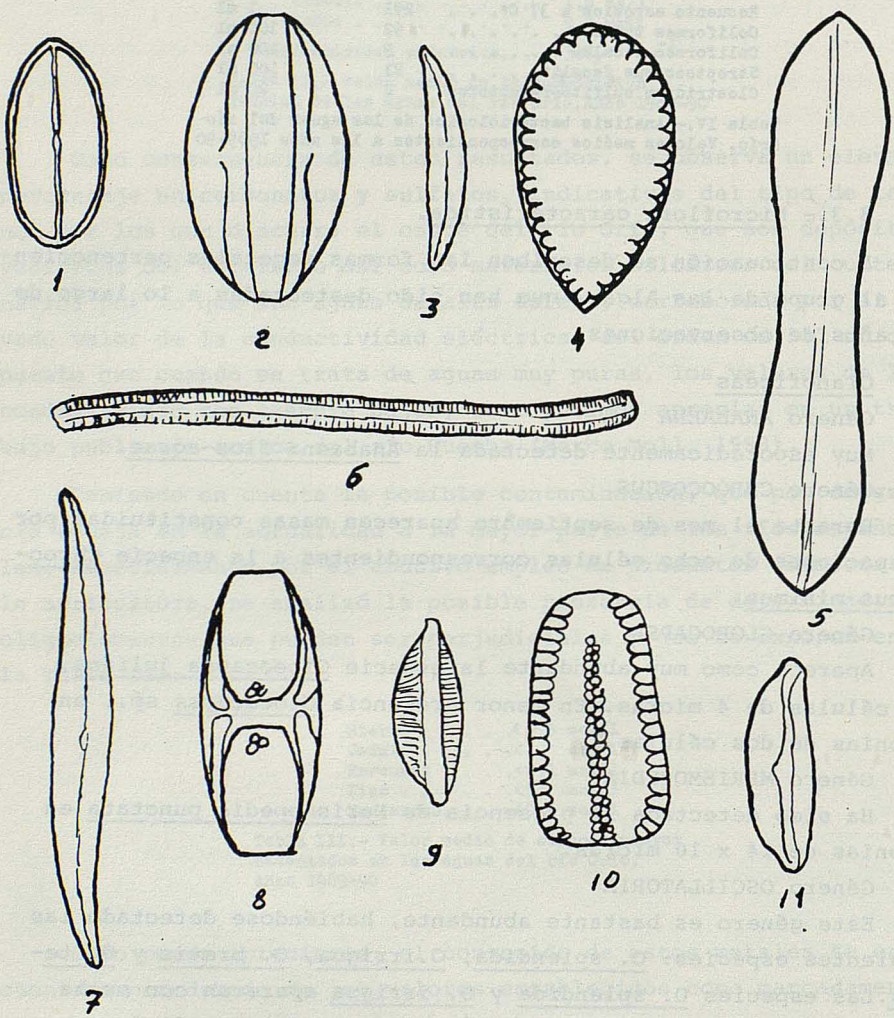


Lámina I.- 1) *Diploneis ovalis*; 2) *Amphora ovalis*; 3) *Cymbella tennuis*; 4) *Surirella patella*; 5) *Cymatopleura solea*; 6) *Nitzschia genuina*; 7) *Gyrosigma attenuatum*; 8) *Amphora ovalis*; 9) *Cymbella naviculiformis*; 10) *Surirella* sp.; 11) *Cymbella prostrata*.

Debemos resaltar la presencia de algunas especies en cantidades abundantes, que exponemos en la lámina I, como son Diploneis ovalis, Amphora ovalis, Cymbella tennuis, Surirella patella, Cymatopleura solea, Nitzschia genuina, Gyrosigma attenuatum, Cymbella naviculiformis y Cymbella prostrata.

Cloroficeas

Género ANKISTRODOSMUS

Durante los meses de mayo y junio se detecta la presencia de Ankistrodosmus falcatus.

Género CLADOPHORA

El género Cladophora se encuentra poco extendido, unicamente en remansos, que no son muy abundantes. Aparece muy esporadicamente Cladophora glomerata y Cladophora fracta.

Género CRUCIGENIA

Muy esporadicamente la especie Crucigenia irregularis, en cenobios de 14 micras.

Género PEDIASTRUM

Este género se encuentra ampliamente representado a lo largo de todo el curso fluvial, presentándose en cantidades masivas durante el mes de noviembre, hasta tal punto que constituyen las formas dominantes de algas, con las siguientes especies:

Pediastrum duplex.- Formas coloniales de 10+5+1, de 42 micras de diámetro, siendo ésta la única especie de este género que se mantiene durante el invierno (Fig. 2)

Pediastrum sp.- Colonias de 6+2, en ocasiones de 6+1, con un diámetro de 42-56 micras. Las células periféricas aparecen múticas (Fig. 3).

Pediastrum sp.- Formas coloniales de 8+1 células. Las periféricas aparecen con dos ángulos cortos.

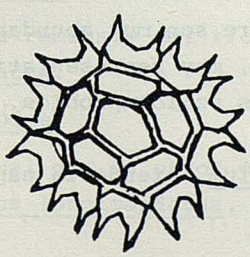


Fig 1.-Pediastrum duplex

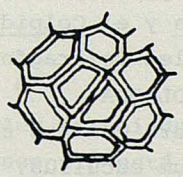


Fig 2.-Pediastrum sp.

Género SCENEDESMUS

Este género se encuentra representado por las siguientes especies:

Scenedesmus Westi, formas coloniales de ocho células, que portan un cloroplasto bien definido.

Scenedesmus acutiformis, colonias de cuatro células.

Scenedesmus cuadricauda, colonias de cuatro células, las extremas terminadas en una larga espina encorvada.

Scenedesmus acuminatus, en colonias de ocho células.

Conjugadas

Género CLOSTERIUM

Las aguas del río Grío son abundantes en especies del género Closterium, habiéndose identificado las siguientes especies:

Closterium acerosum, con sus células rectas en forma de puro, que miden 280 micras.

Closterium pusillum. muy abundante.

Closterium Ehrenbergi, abundante.

Closterium littorale.

Closterium moniliforme, muy abundante.

Closterium beibleni.

Género COSMARIUM

La especie más corriente es el Cosmarium bipunctatum. Muy esporádico el Cosmarium cucurbita al igual que el Cosmarium vesatum.

3.4.- Microfauna característica.

Infusorios

Género PARAMECIUM

Especialmente durante el mes de noviembre se detecta una abundante población de paramecios, representados el primer lugar por el Paramecium caudatus y con menor presencia el Paramecium bursaria. En los restantes meses del año aparecen muy esporadicamente.

Género COLPIDIUM

En los meses de octubre y noviembre son muy abundantes el Colpidium colpoda y el Colpidium campylum, muy representativos en los cursos fluviales de Aragón, indicadores poliosaprobios.

Género COLPODA

Este género convive con el Colpidium, siendo la especie típica el Colpidium cucullus.

Género COLEPS

Este género está representado por el Coleps hirtus, cuya pre-

sencia es muy abundante durante el otoño.

Género EUPLOTES

Se ha identificado el Euplotes charon. Sobre restos vegetales en descomposición el Euplotes patella, infusorio indicador betamesosaprobio. Muy esporádico el Euplotes affinis.

Género LEMBADION

Unicamente durante el mes de septiembre se ha identificado el Lembadion lucens, que vive sobre los filamentos del alga Cladophora fracta.

Género STYLONYCHIA

Sobre filamentos de Cladophora aparecen ejemplares de Stylonychia mytilus. Es poco abundante, siendo considerado por Kolwitz (1969) como indicador alfa/beta mesosaprobio.

Género TRACHELIUS

La especie Trachelius aovum se identificó durante el mes de noviembre, considerado por Kolwitz (1969) como infusorio indicador beta-mesosaprobio de alimentación carnívora.

Género VORTICELLA

Se detectaron sobre filamentos de Cladophora fracta, la Vorticella convallaria y la Vorticella similis, propia de medios oligosaprobios.

Rotíferos

Los Rotíferos no son muy abundantes en en río Grío. Se han identificado las siguientes especies: Philodina roseola, Philodina citrina, Lepadella acuminata, Monostyla lunaris y Monostyla closterocerca.

Copépodos

Es corriente la presencia de una especie sin identificar del género Eucyclops.

Ostracodos

En los remansos del río Grío se ha identificado la especie Cypris monacha.

Nematodos

Al igual que en otros cursos fluviales estudiados, aparece el Mononchus longicaudatus, aunque su presencia está muy restringida.

Finalmente debemos citar la presencia de larvas pertenecien-

tes al grupo de los Chironómidos y que no ha sido posible identificar.

4. DISCUSION.

De todo lo expuesto anteriormente, se puede sacar como consecuencia que el río Grío constituye un ecosistema sumamente inestable, debido a las grandes variaciones a que se encuentra sometido su régimen fluvial, muy especialmente cuando en los meses de verano, su cauce queda totalmente seco.

Esto determina una pobreza en algas filamentosas, a excepción del género Cladophora. Los géneros Ulothrix y Spyrogira, abundantes en el río Huerva (Marco Moll, 1979, 1988), así como en el río Huecha (Marco Moll, 1988), no se desenvuelven en estas aguas.

En cambio, formas unicelulares como Closterium y Cosmarium, así como las del tipo colonial Pediastrum y Scenedesmus, se encuentran ampliamente adaptados a las condiciones ambientales del río Grío, muy especialmente durante los meses de noviembre y diciembre, coincidentes con con un mayor régimen fluvial y una mayor mineralización de sus aguas, representada por su elevada conductividad eléctrica, que alcanza una media de 337 nS cm^{-1}.

Todos estos factores determinan una precaria representación de la microfauna, que no encuentra medio estable y permanente para su supervivencia.

5. RESUMEN.

Las aguas del río Grío muestran una elevada mineralización, con altas dotas de carbonatos y susfatos, con intervalos de estiaje total. Las aguas aparecen fuertemente contaminadas, especialmente en Streptococcus fecalis.

Hay una marcada pobreza en algas filamentosas y existe una preponderancia en especies de los géneros Pediastrum, Scenedesmus, Cosmarium y Closterium, muy particularmente durante los meses en que el caudal resulta apreciable.

6. REFERENCIAS.

BICK, H. (1972): Ciliated Protozoa. World Health Org. Geneve.

CAMARA NIÑO, F. (1951): Diatomeas de las aguas minerales de Aragón. Rev. Acad. Ciencias. Zaragoza, t. VI, 103-16.

DOSSET y MONZON, J.A. (1888): Datos para la sinopsis de las Diatomeas de Aragón.

KIEFER, F. and FRYER, G. (1978): Das Zooplankton der Binnengewasser. Stuttgart.

KOLWITZ, R. and MARSSON, M. (1969): Oecologie der Tierischen Saprobien. Int. Rev. ges Hidrobiol. Hydrogr. 2, 126-52.

MARCO, H. y GASPAR, P. (1979): Estudio de la influencia de las aguas polucionadas sobre la población viviente del río Huerva. An. Est. Exp. Aula Dei. V. 14 (3/4), 606-26.

MARCO, H. (1988): Estudio ecológico del río Huecha. Rev. Acad. Ciencias Zaragoza. 43, 257-283.

MARCO, H. (1988): Contribución a la algología del curso inferior del río Gállego. Rev. Acad. Ciencias. Zaragoza. 43, 285-301.

MARGALEF, R. (1953): Los crustaceos de las aguas continentales ibéricas. Inst. Forets. Inv. y Exp. Madrid.

MARGALEF, R. (1955): Los organismos indicadores en la limnología. Inst. Forets. Inv. y Exp. Madrid.

MARGALEF, R. (1983): Limnología. Ed. Omega. Barcelona.