Received: 10 May 2018 | Revised: 2 November 2018 | Accepted: 6 January 2019

DOI: 10.1111/pirs.12424

FULL ARTICLE

Modelling commuting time in the US: Bootstrapping techniques to avoid overfitting

José Ignacio Gimenez-Nadal[1,2,3] | José Alberto Molina[1,2,4] | Jorge Velilla[1]

[1] Department of Economic Analysis, University of Zaragoza, Spain

[2] Institute for Biocomputation and Physics of Complex Systems (BIFI), Zaragoza, Spain

[3] Centre for Time Use Research (CTUR), Oxford, UK

[4] Institute of Labor Economics (IZA), Bonn, Germany

Correspondence
Jose Ignacio Gimenez-Nadal, Department of Economic Analysis, Faculty of Economic and Business Studies, University of Zaragoza, C/Gran Vía 2, 50005 Zaragoza, Spain.
Email: ngimenez@unizar.es

Funding information
Government of Aragón, Grant/Award Number: project S32_17R "Economía de la población, mercado de trabajo y economía industrial

JEL Classification: J22; J32; R41

Abstract

The research on commuting has emerged in recent decades, but the issue of whether the empirical techniques used are appropriate has not been analysed. Thus, results from prior research could be based on non-accurate models, leading to misleading conclusions. We apply an algorithmic approach based on bootstrapping, variable selection, and mean absolute prediction errors, which is designed to avoid overfitting. Using the American Time Use Survey, we find that models with a reduced set of explanatory variables have similar accuracy to standard econometric models. Our results shed light on the importance of determining whether models can be overfitted.

KEYWORDS

American Time Use Survey, Bootstrap, Commuting time, Overfitting

1 | INTRODUCTION

The time individuals devote to commuting is an important factor in daily activity. In the US, employed workers devote around 38 minutes per day, on average, to commuting (Gimenez-Nadal, Molina, & Velilla, 2018a), and recent studies show that commuting time has increased in many developed countries (Gimenez-Nadal et al., 2018a; Kirby & LeSage, 2009; Mckenzie & Rapino, 2009; Susilo & Maat, 2007). On the other hand, Kahneman, Krueger, Schkade, Schwarz, and Stone (2004) and Kahneman and Krueger (2006) show that commutes are among the lowest daily activities in terms of "instant enjoyment," and others have shown that commuting is associated with high levels of stress and low levels of well-being (Frey & Stutzer, 2004; Hennessy & Wiesenthal, 1999; Novaco & Gonzalez, 2009; Schaeffer & Street, 1988; Stone & Schneider, 2016; Wener, Evans, Phillips, & Nadler, 2003). Thus, the analysis of what factors are related to more time in commuting is important (Liu, Zhang, & Yang, 2017; Rosales-Salas & Jara-Díaz, 2017), and the techniques used to analyse these factors are crucial in determining the research results.

Commuting has been extensively analysed (see Ma & Banister, 2006, for a review), and a certain level of consensus about the factors that are related to commuting has been achieved; factors that can be microeconomic, geographical, or macroeconomic.[1] Several socio-economic characteristics of workers have been found to be important determinants of commuting trips. First of all, commutes can be considered as shocks to time endowments (Ross & Zenou, 2008), and some uses of time, such as leisure, market work, child care or home production, are significantly correlated with commuting (Gimenez-Nadal & Molina, 2016; Gimenez-Nadal, Molina, & Velilla, 2018b). The type of employment is also an important determinant of commutes, as prior research has found significant differences between employees and self-employed workers (Gimenez-Nadal et al., 2018a; Van Ommeren & Van der Straaten, 2008), between full and part-time employees (McQuaid & Chen, 2012), and between workers with different occupations (McQuaid, 2009; McQuaid & Chen, 2012; Walks, 2014). The relationship between wages and commutes has been previously studied and, in general, higher wages are associated with longer commutes (Crane, 2007; Gimenez-Nadal et al., 2018b; Leigh, 1986; Mulalic, Van Ommeren, & Pilegaard, 2014; Ross & Zenou, 2008; Rupert, Stancanelli, & Wasmer, 2009; White, 1999; Zax, 1991). Education has also been found to be positively correlated to commuting (Dargay & Clark, 2012; Dargay & Van Ommeren, 2005; Rouwendal & Nijkamp, 2004; Sandow & Westin, 2010). Commutes have been found to be significantly determined by a range of family variables, such as family structure or car ownership (Dargay & Clark, 2012; McQuaid & Chen, 2012). Furthermore, prior research has pointed to gender as an important determinant of commuting, from different perspectives, as women tend to have shorter commutes than men (e.g., Dargay & Clark, 2012; Gimenez-Nadal & Molina, 2016; Hanson & Hanson, 1993; McQuaid & Chen, 2012; Sandow, 2008; Sandow & Westin, 2010; Waldfogel, 2007). In addition, urban forms, specific geographical characteristics (e.g., population and job density, or housing attributes), and the mode of transport (e.g., by private vehicle, public transport, or active commuting) have been found to be important determinants of commuting, (Cropper & Gordon, 1991; Sandow & Westin, 2010; Deding, Filges, & Van Ommeren, 2009; Gimenez-Nadal et al., 2018a; Manning, 2003; McQuaid & Chen, 2012; Rodriguez, 2004; Ross & Zenou, 2008; Rouwendal & Nijkamp, 2004; Small & Song, 1992; Susilo & Maat, 2007).

Prior research on commuting has yielded certain methodological conclusions. For instance, commuting can be considered in terms of time or of distance, but it is important to know how it is measured and reported (e.g., diaries, stylized questions, aggregated flows). The evidence suggests that commuting times are, in general, less biased than commuting distances (Small & Song, 1992), and that surveys based on diaries may represent a more accurate source of information than stylized question surveys (Gimenez-Nadal & Molina, 2016; Jara-Díaz, Bhat, & Tudela, 2015; Jara-Díaz & Rosales-Salas, 2015; Kitamura, Fujii, & Pas, 1997). Also, commuting times depend on the type of commute (active commuting, commuting by public transport, or commuting by private vehicle), and also on exogenous and stochastic factors (e.g., traffic congestion). Furthermore, commuting is endogenously related to several other variables, such as income, other time allocations, or specialization.

Most of the research on commuting behaviour is based on techniques that rely on strong assumptions (e.g., little or no multicollinearity, homoscedasticity, or normality of errors in the significance of parameters in regressions, among others), but it often omits to check whether these assumptions are fulfilled or not. In this sense, empirical analyses of commuting are characterized by an absence of statistical proofs about model assumptions (Gimenez-Nadal & Molina, 2016; Gimenez-Nadal et al., 2018a; Kimbrough, 2016; Mulalic et al., 2014; Ross & Zenou, 2008; Rupert et al., 2009; Van Ommeren, Rietveld, & Nijkamp, 1999; Van Ommeren & Van der Straaten, 2008). In that context, a criterion such as model accuracy may become more important than the fulfilment of several assumptions. Friedman (1953) notes that the performance of a model is a key point in determining whether results and conclusions are accurate, and Breiman (2001) presents a review of the importance of model accuracy. If a model is accurate, we can assume that results and conclusions are reliable, and we do not have to worry about whether model assumptions

[1] The literature addressing the relationship between commuting and macroeconomic aspects is scarce, in comparison with the literature studying commutes, microeconomic characteristics, and geographical structure (e.g., Johansson, Klaesson, & Olsson, 2002; Osth & Lindgren, 2012).

are fulfilled or not. Thus, in this paper, we analyse commuting time by focusing on model accuracy rather than on the fulfilment of assumptions. In doing so, we apply an algorithmic approach developed in Gimenez-Nadal, Lafuente, Molina, and Velilla (2019) that is based on bootstrapping, variable selection, and mean absolute prediction errors over test sets, designed to avoid overfitting, not to fulfil strong assumptions.

The risk of overfitting arises when empirical models use large sets of explanatory variables, which may not be so relevant in explaining the dependent variable. Classical measures of accuracy, such as the R^2 or the mean squared error, tend to overestimate the accuracy of models, as they always increase with the number of explanatory variables. Hence, when the accuracy of models is estimated using these measures, we may be relying on non-relevant explanatory variables that increase the difficulty of the model, but do not add useful information, which may increase the risk of being overfitted. As other measures that penalize the number of variables, such as information criteria or the adjusted R^2, are not appropriate to examine overfitting—as they still overperform model accuracy (Gimenez-Nadal et al., 2019)—a common approach to avoid the problem is to estimate the accuracy of empirical models using observations that have not been used to estimate the model (test sets). [2] To the best of our knowledge, prior research has not investigated this issue, and this is the first applied research to study overfitting in commuting modelling.

We analyse the time devoted to commuting using the American Time Use Survey for the years 2003 to 2014, applying a forward selection algorithmic process based on bootstrapped linear regression models, and mean prediction errors over test sets. Using mean prediction errors over test sets as an indicator of model accuracy, rather than classical statistics (*t*-tests, *F*-tests, R^2, adjusted R^2, or information criteria), we can deal with overfitting and find the best models in terms of model accuracy, without relying on strong assumptions. In such a context, bootstrapping is important, as it allows us to avoid biases from the selection of training and test sets. Bootstrapping techniques have rarely been used in commuting analysis (sometimes in the presence of predicted regressors; Fingleton, 2006; Gimenez-Nadal & Molina, 2016) but, to the best of our knowledge, they have not previously been used to estimate the goodness of fit of empirical models via prediction errors. Our results show that this algorithmic approach performs better (using fewer regressors), in terms of model accuracy, in comparison with standard techniques, such as the standard forward stepwise selection process. Further, it performs like other techniques in contemporary statistics, such as the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), using a considerably smaller set of regressors, and then with simpler models. This may consequently reduce the risk of overfitting. From our algorithmic approach, we conclude that population density at the urban level, gender, race, and civic status (i.e., living in couple) are major predictors of commuting time. The type of dwelling, working hours, and certain occupations and industries are also strong predictors of commuting time.

We contribute to the literature by the application of a statistical technique that avoids overfitting in the analysis of commuting patterns and behaviours. Our algorithmic approach shows a similar or better performance, in terms of accuracy, than standard econometric techniques, but using significantly smaller set of explanatory variables. Thus, we contribute to the field of methodological tools to analyse commuting time, by its application to a field where these tools have been overlooked. We also contribute to the analysis of the factors related to the commuting behaviour of workers in the US, using a method that has not previously been applied to this kind of study. Our conclusions about the factors related to commuting behaviour are consistent with prior results.

The rest of the paper is organized as follows. The data and variables are described in Section 2. Section 3 presents the empirical strategy, and Section 4 shows the results. Our conclusions are drawn in Section 5.

[2] When we estimate the predictive power of a model using the same set of individuals as are used to estimate the parameters of the model, we may overestimate its accuracy. The more variables, the better the model, but adding explanatory variables to a model cannot generally be associated with better performance (the problem of overfitting). When we estimate the performance of a model using test sets, we find that the performance of a model vs. the number of explanatory variables has a U-shaped relationship (Gimenez-Nadal et al., 2019). Then, we can identify the point at which more variables do not improve the accuracy of the model, thus dealing with overfitting.

2 | EMPIRICAL STRATEGY

The accuracy of empirical models must be considered when doing applied research, as accurate models lead to accurate conclusions, while inaccurate models can lead to results with a high risk of error. Then, performance and inference cannot be considered separate concepts. Traditional measures of model accuracy, such as the R^2, or the mean squared error, tend to overestimate the accuracy of models, as they are estimated together with the parameters of models using the same individuals (training sets), and thus they monotonously increase with the number of explanatory variables, whether or not these variables are relevant to the model. Information criteria (e.g., BIC, AIC) and the adjusted R^2 are also biased, and do not correctly deal with overfitting (Gimenez-Nadal et al., 2019), even when they are designed to take into account and penalize the number of explanatory variables, given that they are also estimated over training sets. In particular, the addition of non-relevant explanatory variables increases the difficulty and the levels of noise of the model, perhaps leading to confounding estimates. This problem, known as overfitting, can be avoided by estimating the relevance of the explanatory variables of an empirical model using prediction errors over test sets (individuals not used to estimate the parameters of the model). From that perspective, we propose the forward stepwise selection algorithm of Gimenez-Nadal et al. (2019) in the analysis of commuting time (e.g., see point #2 of the code in their Appendix), to obtain the best regression model in terms of predictions, with the primary objective of studying overfitting in commuting models.

Following this approach, for each explanatory variable X_i, $i = 1, \ldots, M$, with M being the total number of potential explanatory variables, we estimate a bootstrapped simple linear regression model of the log of commuting time (C), in terms of X_i, $C = \beta_0 + \beta_1 X_i$. In each iteration of the bootstrap, parameters are estimated over the bootstrapped sample and the absolute prediction error is calculated over the observations not included in the bootstrapped sample. When the bootstrap ends, the mean absolute error (mae) is defined as the mean of all the average absolute errors of the bootstrap process. With a large number of bootstrap iterations, the mae does not depend on the training and test sets random selection, and it will be an unbiased estimate of the relevance of the explanatory variable X_i, and an indicator of the accuracy of the model.

We repeat this bootstrap process for each of the M explanatory variables, and we then obtain M mae values, each associated with a variable. We retain the explanatory variable with the lowest mae, X_{opt}^1, that would be the "best" regressor in terms of predictions.

Now, for each of the $M - 1$ remaining explanatory variables, we repeat analogously the bootstrap process, but now including in the models the variable selected in the first step, X_{opt}^1. Then, for each of the remaining X_i explanatory variables, we estimate $C = \beta_0 + \beta_1 X_{opt}^1 + \beta_2 X_i$. That is to say, we run $M - 1$ bootstrapped regression models and estimate $M - 1$ mae values, each associated with a variable. We retain the variable with the lowest mae, X_{opt}^2. It must be noted that the inclusion of X_{opt}^1 is important to avoid multicollinearity issues. If an explanatory variable X_k is highly collinear with X_{opt}^1, the mae of the model, including both variables, X_{opt}^1 and X_k, will either increase or not decrease meaningfully, in comparison with the mae associated only with X_{opt}^1, since the extra information collected by X_k will be considered noise, once X_{opt}^1 is included in the model.[3]

The process continues iteratively, and ends when the optimum mae of the K step is higher than the optimum mae of the $K - 1$ step. This indicates that the addition of the best explanatory variable, in terms of its predictive power, of the K iteration does not improve the model of the $K - 1$ step. Then, the optimum commuting linear regression model, according to its predictive power, would contain $K - 1$ independent variables, according to the sample and data used, and will not be overfitted, as no relevant information will have been included in the model.

[3] This is the key point of the forward stepwise selection approach proposed, since the average prediction error estimated through training sets will always decrease, oppositely to the mean of the average prediction errors over test sets, that will only decrease if the variable added to the model is not correlated with the previously-added regressors, and contributes to the explanation of the dependent variable.

Furthermore, conclusions derived from it would be the most accurate, independently of model assumptions, since, if the model is accurate, so are the results (Breiman, 2001). It must be noted that this iterative process is based on predictions, and can be used with empirical specifications other than linear regressions.

The difference between this forward stepwise selection approach and the classical one, based on the addition of the "most significant" explanatory variables, arises from the potential biases of such significance. First, significance levels are based on *t*-tests, which are designed to work over normally-distributed variables (which is not standard, and is often ignored in applied research in microeconomics). Second, and more importantly, the addition of variables according to their significance could lead to confounding issues. It could be that the effect and significance of a variable is conditioned by the presence of others. For instance, important features may not be included using significances, while others that represent repetitive information may be (i.e., there may be problems of unobserved heterogeneity and multicollinearity). Finally, there is no guarantee that variables whose coefficients are significant at standard levels are relevant to models, in terms of predictions (Gimenez-Nadal et al., 2019). Thus, a secondary benefit of the algorithmic approach proposed, is that of not relying on strong assumptions (which are a problem only if they are not satisfied), since if the empirical models are accurate, we can expect that results and conclusions are accurate. Given that most of the applied research in commuting ignores model assumptions, this secondary contribution is worth mentioning.

3 | DATA AND VARIABLES

We use the American Time Use Survey (ATUS) for the years 2003 to 2014. The ATUS is a database that provides individual time use based on diaries, in which respondents are asked to report what they did (i.e., activities, time spent on these activities) during the 24 hours of the day. The ATUS then identifies a set of primary activities, such as paid work, leisure, or TV watching, among others, and thus this information can be used to add up the time devoted to any of these activities. An advantage of the ATUS over other time use surveys, that are based on stylized questions rather than on diaries, is that prior research has found that diary-based estimates are more reliable and show lower measurement error than estimates based on stylized questions (Bianchi, Milkie, Sayer, & Robinson, 2000; Bonke, 2005; Yee-Kan, 2008). The database also includes a wide range of socio-demographic, family, and labour variables. The ATUS, administered every year since 2003 by the US Bureau of Labor Statistics, is considered the official time use survey of the country. More information, and data, can be found at http://www.bls.gov/tus/.

We restrict the sample to private sector employees between 21 and 65 years old who commute by private vehicle (car, truck, or motorbike), to minimize the role of time-allocation decisions, such as education and retirement, and for the sake of comparison with prior studies (Aguiar & Hurst, 2007; Gimenez-Nadal et al., 2018a; Gimenez-Nadal & Molina, 2016).[4] That way, results can be interpreted as being "per working-age adults" (Gimenez-Nadal & Sevilla, 2012). Further, given that individuals may have completed the diaries during non-working days, and then we would compute zero commuting for those workers, we restrict the analysis to working days. In doing so, we define working days as days when workers devote 60 minutes or more to market work activities, excluding commuting. That way, we avoid computing zero commuting for individuals who filled out the ATUS diary on a non-working day. These restrictions leave us with a final sample of 27,439 individuals, 14,373 males and 13,066 females.

We define the time of commuting of individuals as the time spent in the activity "commuting to/from work" (with activity code "180501" in the ATUS). Table 1 shows summary statistics of commuting time, by gender. Male workers devote, on average, 48.2 minutes per day to commuting, in contrast with 37.9 minutes by female workers, which is consistent with Gimenez-Nadal and Molina (2016). Figure 1 shows k-density estimates of commuting and log of

[4] Self-employed workers are eliminated from the sample, as prior research on commuting has found that they show different structural behaviours related to both commuting distance (Van Ommeren & Van der Straaten, 2008), and commuting time (Gimenez-Nadal et al., 2018b).

TABLE 1 Summary statistics

	Male		Female		
Variables	**Mean**	**S.Dev.**	**Mean**	**S.Dev.**	***p*-value**
Commuting time	48.276	43.234	37.967	33.536	(<0.001)
Age	41.201	11.185	41.629	11.618	(0.002)
Age squared	18.226	9.360	18.679	9.806	(<0.001)
Primary education	0.095	0.293	0.068	0.252	(<0.001)
Secondary education	0.303	0.459	0.290	0.454	(0.024)
University education	0.603	0.489	0.642	0.480	(<0.001)
Naturalized citizen	0.887	0.317	0.924	0.265	(<0.001)
Being white	0.852	0.355	0.807	0.395	(<0.001)
Being American	0.810	0.392	0.849	0.358	(<0.001)
Being Asian	0.039	0.194	0.038	0.191	(0.519)
Being Pacific islander	0.002	0.046	0.002	0.047	(0.813)
Father born in US	0.774	0.418	0.811	0.391	(<0.001)
Mother born in US	0.775	0.418	0.812	0.391	(<0.001)
Live in couple	0.672	0.470	0.549	0.498	(<0.001)
Couple work	0.447	0.497	0.469	0.499	(<0.001)
Have children	0.540	0.498	0.549	0.498	(0.145)
Family size	3.061	1.517	2.879	1.434	(<0.001)
Family total income	62787	43775	56144	42421	(<0.001)
Being a supervised worker	0.558	0.497	0.707	0.455	(<0.001)
Agreed working hours	43.750	14.548	37.455	13.443	(<0.001)
Hourly wage	20.561	15.086	16.330	19.209	(<0.001)
Leisure time	106.450	85.467	96.565	80.175	(<0.001)
Market time	509.175	142.052	464.085	135.928	(<0.001)
Leisure at work time	29.823	35.054	26.775	31.155	(<0.001)
Personal care time	104.421	56.626	113.578	57.231	(<0.001)
Live in the center of the MSA	0.213	0.410	0.220	0.414	(0.163)
Live in the fringe of the MSA	0.635	0.482	0.625	0.484	(0.090)
Live in a non-metropolitan area	0.152	0.218	0.155	0.226	(0.089)
Dwelling: Family home	0.727	0.446	0.707	0.455	(<0.001)
Dwelling: Room rented	0.263	0.440	0.282	0.450	(<0.001)
Other dwelling	0.010	0.103	0.011	0.104	(0.290)
Live in a house/apartment/flat	0.956	0.206	0.954	0.209	(0.545)
Live in a hotel	0.002	0.026	0.001	0.024	(0.369)
Mobile home	0.039	0.193	0.038	0.192	(0.834)
Other housing	0.003	0.032	0.007	0.081	(0.405)
MSA size	3.606	2.515	3.563	2.509	(0.155)
N. Observations	14,373		13,066		

Notes: The sample comes from the ATUS 2003–2014 and is restricted to private sector employees who work the diary day and commute by private vehicle. Time-use variables are measured in minutes. Wages are measured in $ per hour. Family total income is measured in $ per year. *T*-test *p*-values for the differences in parentheses.

commuting time. We observe that most of the density mass is concentrated around the median of 30 minutes, and shows a long tail on the right side. Given that the log of commuting time has a similar density to that of the normal distribution, we will estimate the empirical models in terms of log-of-commutes.

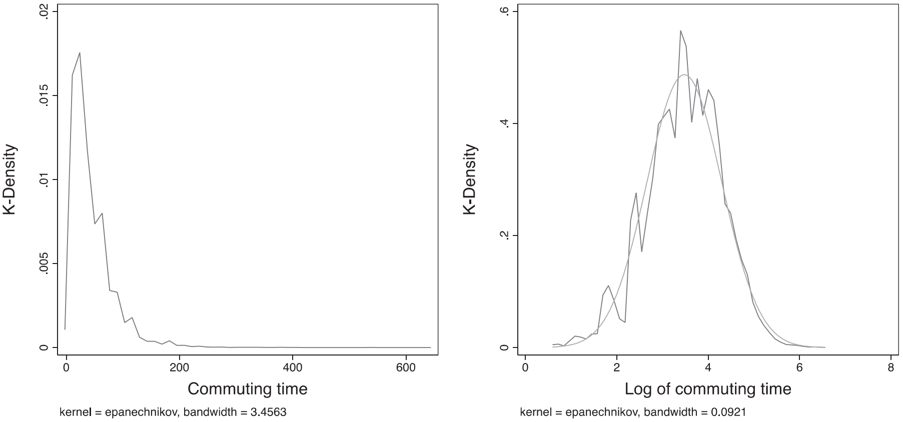

FIGURE 1 K-densities of commutes and log-commutes
Notes: The sample (ATUS 2003–2014) is restricted to private sector employees who work the diary day and commute by private vehicle. Commuting times are measured in minutes.
Source: Author's elaboration.

Regarding the potential explanatory variables used to develop our analysis, the ATUS allows us to define several socio-demographic characteristics, including gender, age, the level of education (coded as primary education for less than twelve years of education, secondary education for twelve years of education, and university education for more than twelve years of education), being a naturalized US citizen, race, live-in couple, couple's working status, the presence of children, family size, and household total income. Economic and labour characteristics are also considered, including hourly wages, agreed weekly working hours, being a supervised worker, and the industry and occupation of the worker.[5] Information about the type of housing unit (house/apartment/flat, hotel, mobile home, or other housing), and the type of dwelling (family home, rented room, or other type of dwelling) is also available.

Geographic characteristics are also considered in the analysis.[6] The ATUS includes information on whether households reside in the central city within a metropolitan area, on the fringe of a metropolitan area (or just in a metropolitan area if no distinction is made), or in a non-metropolitan area. The ATUS also includes information on the metropolitan statistical area (MSA) where the individuals are located, and on the population size of the MSA, coded as follows: (i) Non-metropolitan; (ii) 100,000–249,999; (iii) 250,000–499,999; (iv) 500,000–999,999; (v) 1,000,000–2,499,999; (vi) 2,500,000–4,999,999; and (vii) 5,000,000+. The population size of the MSA of residence may be important, as prior research has found that workers in larger cities have longer commutes (Gordon, Kumar, & Richardson, 1989; Kahn, 2000; Mieszkowski & Mills, 1993). Finally, in order to check whether time-use variables are related to each other, we use the time devoted to the labour market (excluding commuting), the time devoted to leisure, the time devoted to leisure at work (defined as loafing, Burda, Genadek, & Hamermesh, 2016), and the time devoted to personal care (Gimenez-Nadal & Molina, 2016; Rosales-Salas & Jara-Díaz, 2017).

[5] The ATUS codes industries and occupations into 14 and 11 categories, respectively. For industry, the following categories are defined: agriculture, forestry, fishing, and hunting; mining; construction; manufacturing; trade; transportation and utilities; information; financial activities; professional and business services; educational and health services; leisure and hospitality; other services; public administration; and armed forces. For occupation, the following categories are defined: management, business, and financial; professional and related; service; sales; office and administrative support; farming, fishing, and forestry; construction; installation, maintenance, and repair; production; transportation and materials moving; and armed forces.

[6] Commuting has been identified using cross-metropolitan variations in commuting time. Within this framework, individual housing attributes have been proven to be significantly associated with commutes (Cutler & Gleaser, 1997; Ross & Zenou, 2008).

Table 1 shows summary statistics of these potential explanatory variables. The average age is 41.2 for males and 41.6 for females, with this difference being small but significant at the 99% level. A small proportion of males (9.5%) and females (6.8%) have a primary education level, while there are more men with secondary education (30.3% of the males vs. 29.0% of the females), and most of the sample have gone to university, especially among women (64.2% of the women, vs. 60.3% of the men). Most of the sample is composed of Americans, naturalized citizens, and whites. Around 23% of the males are second-generation immigrants, against 19% of the females. The majority of the sample live in couple, with 67.2% of the males and 54.9% of the females cohabiting in a married or unmarried couple, and most of the sample have children (54% of males and 54.9% of females). The average family size is around three individuals, and male respondents report higher total family income than that reported by females ($62,787/year vs. $56,144/year). In terms of labour conditions, 55.8% and 70.7% of the male and female workers are supervised, or monitored, by their respective employers (we define supervised and unsupervised workers from the classification of Gimenez-Nadal et al., 2018b, according to industry), and the average number of weekly working hours of males is meaningfully higher than that of females, 43.8 vs. 37.5 hours per week, respectively. In terms of wages, the average wage rate for males is $20.60 per hour, vs. $16.30 per hour for women (nominal wage rates collected in the ATUS have been transformed to real hourly wages by dividing them by the price deflator from the Federal Reserve Bank of St. Louis). Male individuals devote each day, on average, 106 minutes to leisure during free time, 509 minutes to paid work, 30 minutes to leisure while at work, and 105 minutes to personal care activities; females devote 97 minutes to leisure, 464 minutes to paid work, 27 minutes to leisure while at work, and 114 minutes to personal care. All these differences are significant at standard levels.

Regarding the variables of place of residence, 21.3% of males and 22% of females reside in the centre of a metropolitan area (with the difference not being statistically significant at standard levels), 63.5% of males and 62.5% of females reside in the fringe of a metropolitan area, and 15.2% and 15.5% in non-metropolitan areas, respectively. Most of the sample (72.7% of the males and 70.7% of the females) also reside in a family home, against 26.3% of males and 28.2% of females who reside in rented rooms. The proportion of individuals residing in other types of dwelling is almost negligible. Furthermore, 95.6% of males and 95.4% of females live in a house, apartment, or flat, in contrast to other types of residences, such as hotel, mobile home, and other housing. There are no gender differences in the population size of the metropolitan area of residence, with the average size being about 250,000–999,999 inhabitants.

4 | RESULTS

The strategy proposed in Section 2 is applied to the log of commuting time. Results are shown in Figure 2, which is interpreted as follows: in the X-axis, we represent the number of explanatory variables included in the model (i.e., the step of the process) and in the Y-axis, the mae associated with the optimal variable of each step. Because including wages and time-use features (work time, leisure, shirking, and personal care) could lead to endogeneity issues, we repeat the analysis without including these variables.[7]

We find that, when taking into account all the variables (left panel of Figure 2), the optimal model, in terms of model predictive power, includes: (i) MSA size; (ii) work time; (iii) live in a family house; (iv) have a construction occupation; (v) hourly wages; (vi) living in couple; (vii) living in the fringe of a MSA; and (viii) living in the MSA of Detroit. All of these variables are statistically significant at the 99% level, indicating that individuals who work longer hours, have higher wage rates, live in couple (vs. single workers), live in the fringe of a metropolitan area (in comparison to those who live in urban cores or in non-metropolitan areas), and live in a family home (vs. rented room or other type of dwelling), all devote more time to commuting. Additionally, the population size of the metropolitan area of residence is also positively correlated with the time devoted to commuting.

[7] A plot of each of the steps, that is, a plot of the process step by step, showing in a figure per step all the potential variables included in each step, vs. their associated m.a.e., is available upon request.

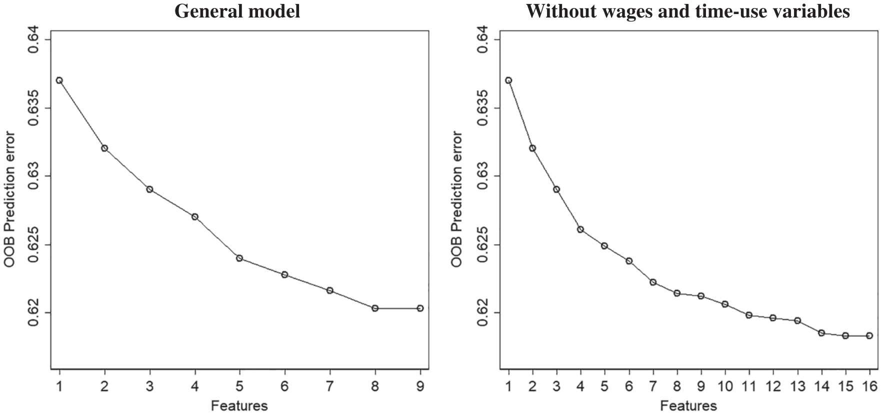

FIGURE 2 Output: Prediction errors of the models

Notes: The sample (ATUS 2003–2014) is restricted to private sector employees who work the diary day and commute by private vehicle. The Y-axis represents the absolute error of prediction associated with the optimal model estimated in each iteration, represented in the X-axis. Dependent variable is the log of commuting time. Time-use variables are measured in minutes. Wages are measured in $ per hour. In the first (second) panel, including (excluding) time-use features and wages, we observe a downward trend until the 8th (15th) iteration, with the 9th (16th) model being worse than the previous one, in terms of its predictive capabilities.

Source: Author's elaboration.

We compare the model used in the algorithmic approach of Section 2 with a model derived from a classical approach, that is, a standard forward stepwise selection process, which adds explanatory variables to a regression model in terms of their significance (i.e., their associated p-value). This model is estimated in column (2) of Table 2. We also compare the model used by the algorithmic approach with the LASSO (Tibshirani, 1996). Results from the LASSO can be found in column (3) of Table 2. A description of these two alternative methodologies is in the Appendix.

The first difference between the model derived from the algorithmic approach of Section 2 and the forward stepwise selection model is the larger number of variables included in the latter model. It must be said that all eight explanatory variables included in the model of column (1) are included by the forward stepwise selection process in the model of column (2), and that these variables maintain their statistical significance and the sign of the estimated parameter. In terms of the accuracy of the two models, we estimate the R^2, the adjusted R^2, the AIC, the BIC, the root mean squared error, and the mean absolute error over test sets of the models.[8]

As expected, the R^2, adjusted R^2, AIC and BIC indicate that the model estimated in column (2) is considerably better, in terms of goodness of fit, than the model estimated in column (1), i.e., an R^2 of 0.106 vs. 0.066, and an adjusted R^2 of 0.104 vs. 0.066. However, as has been previously pointed out, these measures tend to outperform the accuracy of models, especially when they include a large number of explanatory variables. The same happens with the predictive power of the models measured through training sets, that is, the root (mean squared error) MSE, indicating that the forward stepwise selection procedure provides a better model, in terms of predictions, than

[8] We use these measures to compare results because they are standard and well-known in the social sciences. However, we acknowledge that new approaches, such as DALEX (Biecek, 2018), are designed to deal with "model-agnostic" explanations, and help to understand and compare different predictive models and their interpretability. We also acknowledge that the accuracy of the estimated models is low. However, these are the best models that can be estimated using the ATUS data, in terms of accuracy. Prior analyses have obtained similar or lower R^2 with a considerably larger set of explanatory variables: < 0.04 in Ross and Zenou (2008); < 0.1 in Gimenez-Nadal et al. (2018a); and between 0.08 and 0.17 in Van Ommeren and Van der Straaten (2008).

TABLE 2 Regression model estimates

Variables	(1) algorithm	(2) forward stepwise selection	(3) LASSO	(4) algorithm	(5) forward stepwise selection	(6) LASSO
Wages & time-use char.						
Work time	0.001*** (0.000)	0.000*** (0.000)	-	-	-	-
Leisure time	-	−0.001*** (0.000)	−0.001	-	-	-
Personal care time	-	−0.000*** (0.000)	-	-	-	-
Loafing time	-	-	-	-	-	-
Hourly wage	0.003*** (0.000)	0.001*** (0.000)	0.001	-	-	-
Individual characteristics						
Age	-	0.012*** (0.002)	0.008	-	0.020*** (0.003)	0.012
Age squared	-	-	-	-	−0.010*** (0.004)	−0.003
Being male	-	0.120*** (0.011)	0.125	0.163*** (0.010)	0.131*** (0.012)	0.131
Secondary education	-	-	−0.023	-	-	−0.019
University education	-	-	-	-	-	-
Naturalized citizen	-	-	−0.030	−0.008 (0.023)	-	−0.027
Being white	-	−0.066*** (0.015)	−0.065	−0.060*** (0.014)	−0.070*** (0.015)	−0.068
Being American	-	−0.078*** (0.024)	−0.053	−0.104*** (0.019)	−0.075*** (0.024)	−0.054
Being Asian	-	−0.091*** (0.029)	−0.079	−0.045 (0.029)	−0.104*** (0.030)	−0.092
Father born in US	-	-	−0.023	-	-	−0.017
Mother born in US	-	−0.045** (0.022)	−0.025	-	−0.041* (0.022)	−0.025
Living in couple	0.092*** (0.010)	0.137*** (0.016)	0.138	0.068*** (0.010)	0.135*** (0.016)	0.135
Couple work	-	−0.098*** (0.014)	−0.097	-	−0.099*** (0.014)	−0.097
Having children	-	−0.085*** (0.013)	−0.087	-	−0.078*** (0.014)	−0.076
Family size	-	0.014*** (0.005)	0.013	-	0.013** (0.005)	0.012
Geographical & housing char.						
Metropolitan Centre	-	-	0.016	-	-	0.015
Fringe metrop.	0.079*** (0.010)	0.079*** (0.010)	0.089	0.089*** (0.010)	0.084*** (0.010)	0.093

(Continues)

TABLE 2 (Continued)

Variables	(1) algorithm	(2) forward stepwise selection	(3) LASSO	(4) algorithm	(5) forward stepwise selection	(6) LASSO
Dwelling: Family home	0.095*** (0.011)	0.080*** (0.012)	0.128	0.113*** (0.011)	0.079*** (0.012)	0.139
Dwelling: Rented room	-	-	0.049	-	-	0.060
House/appartment/flat	-	-	−0.052	-	-	−0.064
Hotel	-	-	0.237	-	-	0.194
Mobile home	-	0.083*** (0.025)	0.034	-	0.086*** (0.025)	0.021
MSA size	0.050*** (0.002)	0.034*** (0.002)	0.032	0.046*** (0.002)	0.033*** (0.002)	0.032
Labor characteristics						
Being a supervised worker	-	-	0.067	-	0.061*** (0.016)	0.062
Agreed working hours	-	0.001*** (0.000)	0.001	0.003*** (0.000)	0.002*** (0.000)	0.002
Years working	-	−0.010*** (0.002)	−0.004	-	−0.010*** (0.002)	−0.005
Years working squared	-	-	-	-	-	−0.001
Family total income	-	-	-	-	-	-
Constant	2.775*** (0.021)	3.052*** (0.067)	3.034	3.066*** (0.027)	2.570*** (0.073)	2.657
Observations	27,439	27,439	27,439	27,439	27,439	27,439
R-squared	0.066	0.106	0.107	0.078	0.096	0.096
Adjusted R-squared	0.0658	0.104	0.104	0.078	0.094	0.096
AIC	65,037.33	63,948.31	63,970.23	64,691.2	64,258.13	64,288.69
BIC	65,110.52	64,441.81	64,701.78	64,822.72	64,775.97	65,003.20
Root MSE	0.791	0.775	0.777	0.786	0.779	0.781
MAE over test sets	0.620	0.626	0.613	0.618	0.630	0.617

Notes: Robust standard errors in parentheses (columns (1), (2), (4), and (5)). The sample comes from the ATUS 2003–2014 and is restricted to private sector employees who work the diary day and commute by private vehicle. Dependent variable is the log of commuting time. Time-use variables are measured in minutes. Wages are measured in $ per hour. Occupation, industry and MSA dummy variables are not shown. Column (1) includes: construction and extraction industry and Detroit-Ann Arbor-Flint MSA region of residence. Column (2) includes: mining; construction; trade; business; education and health; leisure and hospitality; service industries; management, business and financial services; construction and extraction; maintenance; and production occupations; and MSA region of residence. Column (4) includes: construction; leisure and hospitality; trade industries; construction and extraction; management business and finance occupation. Column (5) includes: mining; construction; trade; transport; business; education and health; leisure and hospitality; services industries; management, business and finance; office and administration; construction and extraction; maintenance occupations; and MSA region of residence. Columns (3) and (6) include 16 dummy variables about industry and occupation dummies, and 40 dummy variables for MSAs.

***, ** and *significant at 1%, 5% and 10% respectively in Columns (1), (2), (4), and (5).

the algorithmic procedure proposed in Section 2. However, when we compare the models in terms of the mean absolute prediction error over test sets, we find that the model estimated in column (1) predicts slightly more reliably than the model of column (2), with errors of 0.620 and 0.626, respectively. This result is derived from the process used to estimate the former model, whose explanatory variables have been included precisely to minimize this error. It is important to note that both prediction errors are similar, but the model in column (1) reaches it with a much smaller set of explanatory variables than the model in column (2). Consequently, the forward stepwise selection model indicates that the information collected in the variables included in column (2), and not included in column (1), does not provide useful information to the model in terms of prediction, that is, they may suppose the inclusion of noise. Hence, we could conclude that the model in column (2) presents a problem of overfitting, as it reaches a similar (even slightly worse) prediction error to the model in column (1) with a significantly larger set of regressors. Further, given that statistical inferences should not be made from inaccurate models (Breiman, 2001), results and conclusions derived from the column (2) model are not necessarily reliable.

We now compare these models with the LASSO, estimated in column (3) of Table 2.[9] We find that the LASSO produces the model with the largest sets of explanatory variables among the procedures studied. This indicates that, for the objective of variable selection and finding non-overfitted models, in the particular case of commuting time, the LASSO is not as accurate as our empirical approach. When we compare the classical measures of goodness of fit (R^2, adjusted R^2, AIC, BIC, and root MSE), we observe that the LASSO overperforms the model derived from our methodological approach. Further, these measures are quite similar to those estimated in the forward stepwise selection model. On the other hand, when we compare the mean absolute prediction errors over test sets, we observe that the LASSO slightly outperforms our proposed methodology, and also outperforms the forward stepwise selection model. This result may be expected, as one of the main advantages of the LASSO is that it outperforms OLS when regression models have high variances (i.e., large sets of explanatory variables). Nevertheless, the LASSO in column (3) produces a gain in model accuracy of 0.065% with respect to the model in column (1), but including 81 additional explanatory variables. Consequently, the model derived from our empirical approach performs somewhat similarly to the LASSO model, with many fewer explanatory variables. This may reduce the potential problems of overfitting.

Given that hourly wages and other uses of time in the diary day may lead to endogeneity problems, we now estimate the same three models excluding these variables. The results of applying our algorithmic approach are shown in the right panel of Figure 2. According to it, the optimal model is composed of: (i) MSA sizes; (ii) being male; (iii) living in a family house; (iv) having a construction occupation; (v) agreed weekly working hours; (vi) living in the fringe of an MSA; (vii) having a management occupation; (viii) being a naturalized citizen; (ix) being white; (x) working in the leisure industry; (xi) being American; (xii) working in the construction industry; (xiii) living in couple; (xiv) working in the trade industry; and (xv) being Asian. Note that the explanatory variables that appeared in this model (excluding wages and work times, which have been eliminated) are often present in the previous model, and the statistical significance and the sign of the parameters associated with these variables remain unchanged, as shown in column (4) of Table 2. Furthermore, we find that male workers devote, on average, 12% more time to commuting than do female workers. White and American workers devote less time to commuting than do their counterparts, while being a naturalized citizen and Asian are not significantly related to commuting time, even when their presence in the model is necessary to provide an optimal performance. Finally, the agreed number of weekly work hours is positively related to commuting. All these variables are often present in the commuting models of the literature,

[9] It is important to note that the output of functions performing OLS (e.g., lm(.) in R, or regress in Stata) usually provide more useful outputs than those performing the LASSO (glmnet(.) and lassoregress, respectively), including standard errors and p-values. Thus, we only show estimated coefficients and measures of goodness of fit regarding LASSO models in Table 2. It is also important to note that the LASSO does not provide standard by-covariate regression tables, and then results provided (which are the output of the command "lassoregress" in Stata) may suffer from "table-output bias."

indicating that, in spite of the probable excess of explanatory variables of such models, the most important features are usually taken into account by authors.

Column (5) of Table 2 shows estimates of the corresponding forward stepwise selection model, in which time-use features and wage rates have been eliminated from the potential regressors. The main difference between the models of columns (4) and (5) is again the considerably larger number of explanatory variables in the latter, indicating that the model in column (4) may again be overfitted. All the variables of column (4) are included in the model of column (5) by the forward stepwise selection process, except the variable that identifies naturalized US citizens. Given that this variable is not significant, according to column (4), its absence in column (5) is not surprising, as the standard forward stepwise selection process incorporates variables according to their statistical significance.

The R^2, adjusted R^2, AIC, BIC and root MSE of the models of column (4) indicate that this model is preferable over the model of column (3), analogously to what happened in columns (1) and (2). However, in terms of the predictive power of the models measured through test sets, we find mean absolute prediction errors of 0.618 and 0.630 for the models of columns (3) and (4), respectively, indicating that the model derived from the algorithmic procedure is again slightly more accurate than the model derived from the standard forward stepwise selection approach, using many fewer explanatory variables. As in column (2), in contrast to column (1), this difference indicates that the information collected in the explanatory variables present in column (4) but not in column (3) may suppose the inclusion of noise and, then, the model in column (5) presents a problem of overfitting.

Finally, column (6) of Table 2 shows estimates of the LASSO model where hourly wages and time-uses are excluded, similarly to columns (4) and (5). As in the previous case, the LASSO again produces a model with a large set of explanatory variables, and the classical measures of goodness of fit are also similar to those estimated in the forward stepwise selection model, outperforming the algorithmic approach. In terms of the mean absolute prediction error over test sets, the LASSO slightly outperforms our proposed methodology but, as in the previous case, we find an mae of 0.617, which supposes an improvement of 0.016% over the model in column (4), including 73 additional explanatory variables. Consequently, results are in line with the previous scenario, and the proposed methodology attains a similar predictive power to the LASSO, with many fewer explanatory variables, suggesting a reduction in the risk of overfitting. We could conclude that, according to these results, the LASSO does not appear to be a useful tool for variable selection in commuting time research.

An important question that can emerge from our proposed algorithmic approach is whether this bootstrapping technique is able to accommodate multicollinearity; that is, whether explanatory variables selected by the empirical approach are linearly correlated or not. According to the empirical model, if two (or, without loss of generality, more than two) potential regressors are strongly correlated, then they explain the dependent variable similarly. In the case of one of these regressors being selected by the empirical approach, then the information that the other can provide to explain the dependent variable would already be provided by the former. In this situation, we would not expect that the predictive power of the model including both regressors significantly outperforms the predictive power of the model including just one of them. If this is the case, only one of these correlated regressors would be selected by our empirical approach.

To empirically study whether the models used in our algorithmic approach may have problems of multicollinearity, we use the variance inflation factor (VIF). The models estimated in columns (1) and (4) of Table 2 show a VIF of 1.05 and 1.41, respectively. This indicates that the empirical approach proposed does not suffer from multicollinearity problems. For instance, the maximum VIFs associated with a variable is 1.11 in the case of column (1). In column (4), the maximum VIFs are 2.29 and 2.11, associated with being American and being a naturalized citizen, respectively, indicating that there are no issues of multicollinearity. On the other hand, when we study the VIF of the forward stepwise selection model, we find a mean VIF of 2.21 associated with column (2) of Table 2, where years and years-working are strongly correlated (VIFs of 25.2 and 24.25, respectively), and being American and having an American mother are moderately correlated (VIFs of 3.67 and 3.7, respectively). In the case of the model of column (5) the mean VIF is 3.79, indicating a slightly larger problem of multicollinearity, which is mainly driven by the same variables as in the previous case. Finally, the estimated VIFs in the LASSO models (columns (3) and (6) of Table 2) are 3.61 and 10.08, respectively. In the first case, there is a high correlation between age and years working, between the types of dwelling, between certain occupations and being a supervised worker, and among being American, having an American mother, and having an American father. In the second case, there is also a high collinearity between age squared and years working squared. Hence, we find that the algorithmic approach proposed in Section 2 performs better in terms of multicollinearity than the forward stepwise selection model and the LASSO model.

To sum up, results show that the measurement of the goodness of fit of a model using its predictive power over test sets can identify the inclusion of noisy information to models, given that noise increases when useless information is added to the model, in contrast to the optimal set of explanatory variables. These results are counter to the classical goodness of fit measures, which cannot properly identify the addition of useless regressors, and tend to give too much credit to overfitted models. Results also show that the algorithmic process proposed in Section 2 for the selection of explanatory variables in regression models provides simpler (in the sense of fewer explanatory variables) but more accurate models than the standard forward stepwise selection process (and then can help to propose non-overfitted models). Furthermore, we do not find strong correlations among the potential regressors used throughout the analysis, and thus multicollinearity does not appear to be a big problem when studying commuting time, when our algorithmic approach is applied. Finally, we find that commutes can be optimally identified by a reduced amount of regressors: the time devoted to work, income, the type of dwelling and cohabitation, certain socio-demographic variables (e.g., gender, nationality, race, and citizenship status), and the metropolitan status and population size of the area of residence. Given that these variables are usually taken into account by researchers in the modelling of commuting time, and that our results show robust coefficient estimations, we conclude that results from prior research concerning commuting time may not be qualitatively biased, even when overfitted models are estimated in the literature. Consequently, overfitting may not suppose a main problem in commuting time research. Despite that, overfitting should be taken into account, as it may distort estimates and produce confounding results. However, the goodness of fit of the optimal models is far from good, with both prediction errors and classical measures indicating poor model performance. This sheds light on the complexity of commuting patterns, and the importance of non-controllable factors, such as traffic congestion.

5 | CONCLUSIONS

In this paper, we analyse the time devoted to commuting by US workers, with a focus on model accuracy to avoid overfitting, using the American Time Use Survey for the years 2003 to 2014. We apply a forward stepwise selection algorithmic technique, based on bootstrapped absolute prediction errors over test sets, in order to avoid overfitted models and to find the most accurate predictors of US commuting time (Gimenez-Nadal et al., 2019). Results show the importance of factors such as the time devoted to work, income, the type of housing unit and cohabitation, the residential location within cities, and the population size of the area of residence. However, given that time-use variables and wages may be endogenously related to commuting, we repeat the process excluding these features, and determine the significance of gender, work schedules, being white, being American, being a naturalized US citizen, and working in certain industries and having certain occupations.

According to our sample and empirical approach, these variables are sufficient to accurately model commuting time. Comparing this strategy with other techniques of variable selection (forward stepwise selection, and the LASSO), we find that the addition of more explanatory variables could lead to overfitted models and, consequently, perhaps to confounding estimates. Nevertheless, most of the explanatory variables are often present in commuting models that have appeared in the literature, and then overfitting may not always suppose a big problem in commuting time research (despite that confounding estimates should be a source of bias). A secondary benefit of the empirical approach is that, as it is designed to find more accurate models, it does not depend strongly on model assumptions. This may be important in the study of commuting time, as model assumptions are rarely studied in the applied research on commuting and, if they are not fulfilled, results and conclusions may be biased. Finally, it is important to note that, despite having found the best models in terms of predictions of commuting time, they show relatively poor performance. This highlights the complexity of models of commuting time, and that they depend on non-controllable and/or stochastic factors.

The paper has certain limitations. First, the technique proposed is only valid for cross-sectional data. In addition, reverse causality issues cannot be taken into account, and unobserved heterogeneity may have a strong effect on commuting time. Because the data used is cross-sectional, these problems cannot be solved. The proposed algorithmic approach is based on bootstrap techniques that require relatively high computation costs. In each step of the process, we need to estimate a bootstrap model for each of the potential explanatory variables of the data. Nevertheless, if the model to be estimated does not require higher computation costs (i.e., the case of the OLS models analysed in this paper), this limitation is not of major importance. Finally, the data set also imposes certain limitations, as commuting is a phenomenon that depends on stochastic and unobservable factors, such as traffic congestion and the autocorrelation of individuals facing the same levels of congestion in the same metropolitan areas. The ATUS does not allow us to take these issues into account.

ACKNOWLEDGEMENT

Funding from the Government of Aragón (Spain) [project S32_17R “Economía de la población, mercado de trabajo y economía industrial”] is gratefully acknowledged.

ORCID

José Ignacio Gimenez-Nadal https://orcid.org/0000-0002-1610-5451
José Alberto Molina https://orcid.org/0000-0002-9437-4606
Jorge Velilla https://orcid.org/0000-0002-0553-6360

REFERENCES

Aguiar, M., & Hurst, E. (2007). Measuring trends in leisure: The allocation of time over five decades. *The Quarterly Journal of Economics*, *122*, 969–1007. https://doi.org/10.1162/qjec.122.3.969

Bianchi, S. M., Milkie, M. A., Sayer, L. C., & Robinson, J. P. (2000). Is anyone doing the housework? Trends in the gender division of household labor. *Social Forces*, *79*, 191–228. https://doi.org/10.1093/sf/79.1.191

Biecek, P. (2018). DALEX: Explainers for Complex Predictive Models. arXiv Preprint 1806.08915.

Bonke, J. (2005). Paid work and unpaid work: Diary information versus questionnaire information. *Social Indicators Research*, *70*, 349–368. https://doi.org/10.1007/s11205-004-1547-6

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*, 199–231. https://doi.org/10.1214/ss/1009213726

Burda, M., Genadek, K. R., & Hamermesh, D. S. (2016). Not working at work: Loafing, unemployment and labor productivity. NBER Working Paper 21923.

Crane, R. (2007). Is there a quiet revolution in women's travel? Revisiting the gap in commuting. *Journal of the American Planning Association*, *73*, 298–316. https://doi.org/10.1080/01944360708977979

Cropper, M. L., & Gordon, P. (1991). Wasteful commuting: A re-examination. *Journal of Urban Economics*, *29*, 2–13. https://doi.org/10.1016/0094-1190(91)90022-Y

Cutler, J. M., & Gleaser, E. (1997). Are ghettos good or bad? *The Quarterly Journal of Economics*, *112*, 827–872. https://doi.org/10.1162/003355397555361

Dargay, J. M., & Clark, S. (2012). The determinants of long distance travel in Great Britain. *Transportation Research Part A: Policy and Practice*, *46*, 576–587.

Dargay, J. M., & Van Ommeren, J. N. (2005). The effect of income on commuting time using panel data. Paper presented at the 45th Conference of the European Regional Science Association at the Vrije Universiteit Amsterdam, Amsterdam.

Deding, M., Filges, T., & Van Ommeren, J. (2009). Spatial mobility and commuting: The case of two-earner households. *Journal of Regional Science*, *49*, 113–147. https://doi.org/10.1111/j.1467-9787.2008.00595.x

Fingleton, B. (2006). A cross-sectional analysis of residential property prices: The effects of income, commuting, schooling, the housing stock and spatial interaction in the English regions. *Papers in Regional Science*, *85*, 339–361. https://doi.org/10.1111/j.1435-5957.2006.00089.x

Frey, B. S., & Stutzer, A. (2004). Stress that doesn't pay: The commuting paradox. IZA Discussion Paper 1278.

Friedman, M. (1953). *The methodology of positive economics*. Chicago, IL: University of Chicago Press.

Gimenez-Nadal, J. I., Lafuente, M., Molina, J. A., & Velilla, J. (2019). Resampling and bootstrap algorithms to assess the relevance of variables: Applications to cross-section entrepreneurship data. *Empirical Economics*, *56*, 233–267.

Gimenez-Nadal, J. I., & Molina, J. A. (2016). Commuting time and household responsibilities: Evidence using propensity score matching. *Journal of Regional Science*, *56*, 332–359. https://doi.org/10.1111/jors.12243

Gimenez-Nadal, J. I., Molina, J. A., & Velilla, J. (2018a). The commuting behavior of workers in the United States: Differences between the employed and the self-employed. *Journal of Transport Geography*, *66*, 19–29. https://doi.org/10.1016/j.jtrangeo.2017.10.011

Gimenez-Nadal, J. I., Molina, J. A., & Velilla, J. (2018b). Spatial distribution of us employment in an urban wage-efficiency setting. *Journal of Regional Science*, *58*, 141–158. https://doi.org/10.1111/jors.12351

Gimenez-Nadal, J. I., & Sevilla, A. (2012). Trends in time allocation: A cross-country analysis. *European Economic Review*, *56*, 1338–1359. https://doi.org/10.1016/j.euroecorev.2012.02.011

Gordon, P., Kumar, A., & Richardson, H. W. (1989). The influence of metropolitan spatial structure on commuting time. *Journal of Urban Economics*, *26*, 138–151. https://doi.org/10.1016/0094-1190(89)90013-2

Hanson, S., & Hanson, P. (1993). The geography of everyday life. In T. Garling, & R. G. Golledge (Eds.), *Behaviour and environment: Psychological and geographical approaches* (pp. 249–269). Amsterdam: Elsevier.

Hennessy, D. A., & Wiesenthal, D. L. (1999). Traffic congestion, driver stress, and driver aggression. *Aggressive Behavior*, *25*, 409–423. https://doi.org/10.1002/(SICI)1098-2337(1999)25:6<409::AID-AB2>3.0.CO;2-0

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer Texts in Statistics. New York: Springer.

Jara-Díaz, S., Bhat, C., & Tudela, A. (2015). Emerging data and methodological considerations in time-use analysis. *Transportation Research Part A: Policy and Practice*, *76*, 1–3.

Jara-Díaz, S. R., & Rosales-Salas, J. (2015). Understanding time use: Daily or weekly data? *Transportation Research Part A: Policy and Practice*, *76*, 175–195.

Johansson, B., Klaesson, J., & Olsson, M. (2002). Time distances and labor market integration. *Papers in Regional Science*, *81*, 305–327. https://doi.org/10.1007/s101100200000

Kahn, M. E. (2000). The environmental impact of suburbanization. *Journal of Policy Analysis and Management*, *19*, 569–586. https://doi.org/10.1002/1520-6688(200023)19:4<569::AID-PAM3>3.0.CO;2-P

Kahneman, D., & Krueger, A. B. (2006). Developments in the measurement of subjective well-being. *The Journal of Economic Perspectives*, *20*, 3–24. https://doi.org/10.1257/089533006776526030

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, *306*, 1776–1780. https://doi.org/10.1126/science.1103572

Kimbrough, G. (2016). What drives gender differences in commuting behavior: Evidence from the American Time Use Survey (No. 16–4). University of North Carolina at Greensboro, Department of Economics.

Kirby, D. K., & LeSage, J. P. (2009). Changes in commuting to work times over the 1990 to 2000 period. *Regional Science and Urban Economics*, *39*, 460–471. https://doi.org/10.1016/j.regsciurbeco.2009.01.006

Kitamura, R., Fujii, S., & Pas, E. I. (1997). Time-use data, analysis and modeling: Toward the next generation of transportation planning methodologies. *Transport Policy*, *4*, 225–235. https://doi.org/10.1016/S0967-070X(97)00018-8

Leigh, J. P. (1986). Are compensating wages paid for time spent commuting? *Applied Economics*, *18*, 1203–1214. https://doi.org/10.1080/00036848600000073

Liu, W., Zhang, F., & Yang, H. (2017). Modeling and managing morning commute with both household and individual travels. *Transportation Research Part B: Methodological*, *103*, 227–247. https://doi.org/10.1016/j.trb.2016.12.002

Ma, K. R., & Banister, D. (2006). Excess commuting: A critical review. *Transport Reviews*, *26*, 749–767. https://doi.org/10.1080/01441640600782609

Manning, A. (2003). The real thin theory: Monopsony in modern labor markets. *Labour Economics*, *10*, 749–767.

Mckenzie, B., & Rapino, M. (2009). Commuting in the United States: 2009. U.S. Department of Commerce, Economics and Statistics Administration, U.S. Census Bureau.

McQuaid, R. W. (2009). A model of the travel to work limits of parents. *Research in Transportation Economics*, *25*, 19–28. https://doi.org/10.1016/j.retrec.2009.08.001

McQuaid, R. W., & Chen, T. (2012). Commuting times: The role of gender, children and part-time work. *Research in Transportation Economics*, *34*, 66–73. https://doi.org/10.1016/j.retrec.2011.12.001

Mieszkowski, P., & Mills, E. S. (1993). The causes of metropolitan suburbanization. *The Journal of Economic Perspectives*, *7*, 135–147. https://doi.org/10.1257/jep.7.3.135

Mulalic, I., Van Ommeren, J. N., & Pilegaard, N. (2014). Wages and commuting: Quasi-natural experiments' evidence from firms that relocate. *Economic Journal*, *124*, 1086–1105. https://doi.org/10.1111/ecoj.12074

Novaco, R. W., & Gonzalez, O. I. (2009). Commuting and well-being. In Y. Amichai-Hamburger (Ed.), *Technology and Psychological Well-being* (pp. 174–205). Cambridge: Cambridge University Press.

Osth, J., & Lindgren, U. (2012). Do changes in gdp influence commuting distances? A study of Swedish commuting patterns between 1990 and 2006. *Tijdschrift voor Economische en Sociale Geografie*, *103*, 443–456. https://doi.org/10.1111/j.1467-9663.2011.00697.x

Rodriguez, D. (2004). Spatial choices and excess commuting: A case study of bank tellers in Bogota, Colombia. *Journal of Transport Geography*, *12*, 49–61. https://doi.org/10.1016/S0966-6923(03)00025-5

Rosales-Salas, J., & Jara-Díaz, S. R. (2017). A time allocation model considering external providers. *Transportation Research Part B: Methodological*, *100*, 175–195. https://doi.org/10.1016/j.trb.2017.01.019

Ross, S. L., & Zenou, Y. (2008). Are shirking and leisure substitutable? An empirical test of efficiency wages based on urban economic theory. *Regional Science and Urban Economics*, *38*, 498–517. https://doi.org/10.1016/j.regsciurbeco.2008.05.009

Rouwendal, J., & Nijkamp, P. (2004). Living in two worlds: A review of home-to-work decisions. *Growth and Change*, *35*, 287–303. https://doi.org/10.1111/j.1468-2257.2004.00250.x

Rupert, P., Stancanelli, E., & Wasmer, E. (2009). Commuting, wages and bargaining power. *Annals of Economics and Statistics*, *95/96*, 201–220. https://doi.org/10.2307/27917410

Sandow, E. (2008). Commuting behavior in sparsely populated areas: Evidence from northern Sweden. *Journal of Transport Geography*, *16*, 14–27. https://doi.org/10.1016/j.jtrangeo.2007.04.004

Sandow, E., & Westin, K. (2010). People's preferences for commuting in sparsely populated areas: The case of Sweden. *Journal of Transport and Land Use*, *2*, 87–107.

Schaeffer, M., & Street, S. (1988). Effects of control on the stress reactions of commuters. *Journal of Applied Social Psychology*, *18*, 944–957. https://doi.org/10.1111/j.1559-1816.1988.tb01185.x

Small, K. A., & Song, S. (1992). 'Wasteful' commuting: A resolution. *Journal of Political Economy*, *100*, 888–898. https://doi.org/10.1086/261844

Stone, A. A., & Schneider, S. (2016). Commuting episodes in the United States: Their correlates with experiential wellbeing from the American time use survey. *Transportation Research Part F: Traffic Psychology and Behaviour*, *42*, 117–124. https://doi.org/10.1016/j.trf.2016.07.004

Susilo, Y. O., & Maat, K. (2007). The influence of built environment to the trends in commuting journeys in the Netherlands. *Transportation*, *34*, 589–609. https://doi.org/10.1007/s11116-007-9129-5

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, Methodological*, *58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Ommeren, J., Rietveld, P., & Nijkamp, P. (1999). Job moving, residential moving and commuting: A search perspective. *Journal of Urban Economics*, *46*, 230–253. https://doi.org/10.1006/juec.1998.2120

Van Ommeren, J. N., & Van Straaten, J. W. (2008). The effect of search imperfections on commuting behaviour: Evidence from employed and self-employed workers. *Regional Science and Urban Economics*, *38*, 127–147. https://doi.org/10.1016/j.regsciurbeco.2008.01.008

Waldfogel, J. (2007). Parental work arrangements and child development. *Canadian Public Policy/Analyse de Politiques*, *33*, 251–272. https://doi.org/10.3138/cpp.33.2.251

Walks, A. (2014). *The urban political economy and ecology of automobility: Driving cities, driving inequality, driving politics*. New York: Routledge.

Wener, R. E., Evans, G. W., Phillips, D., & Nadler, N. (2003). Running for the 7:45: The effects of public transit improvements on commuter stress. *Transportation*, *30*, 203–220. https://doi.org/10.1023/A:1022516221808

White, M. J. (1999). Urban areas with decentralized employment: Theory and empirical work. *Handbook of Regional and Urban Economics*, *3*, 1375–1412. https://doi.org/10.1016/S1574-0080(99)80005-4

Yee-Kan, M. (2008). Measuring housework participation: The gap between "stylised" questionnaire estimates and diary-based estimates. *Social Indicators Research*, *86*, 381–400. https://doi.org/10.1007/s11205-007-9184-5

Zax, J. S. (1991). Compensation for commutes in labor and housing markets. *Journal of Urban Economics*, *30*, 192–207. https://doi.org/10.1016/0094-1190(91)90036-7

How to cite this article: Gimenez-Nadal JI, Molina JA, Velilla J. Modelling commuting time in the US: Bootstrapping techniques to avoid overfitting. *Pap Reg Sci*. 2019;98:1667–1684. https://doi.org/10.1111/pirs.12424

APPENDIX

FORWARD STEPWISE SELECTION AND THE LASSO

Stepwise selection is a classical strategy for variable selection, which in addition is computationally efficient. In the background of variable selection, forward stepwise selection initially takes a model with only the constant term, and iteratively adds explanatory variables, one at a time, attending to their importance, according to p-values (James, Witten, Hastie, & Tibshirani, 2013). In the first step of the procedure, the model estimates a simple regression model with each of the potential explanatory variables, and then selects the explanatory variable with the lowest p-value (other alternatives are based on the highest R^2). Then, in the next step, the procedure adds to that model each of the remaining explanatory variables, and again keeps the variable with the lowest p-value. The process continues iteratively, and ends when the lowest p-value is higher than a certain bound (e.g., 0.10, in this particular analysis).

In recent decades, other techniques for variable selection have appeared in the literature, such as ridge regression and, in particular, the LASSO (Tibshirani, 1996). Ridge regression is an estimation procedure of linear regressions, similar to ordinary least squares (OLS), where the main difference is that the latter includes a "penalty" term $\boldsymbol{\lambda}$ (usually self-computed via cross-validation), that is defined from a tuning coefficient that shrinks estimated coefficients towards 0. In this way, OLS regressions are equivalent to ridge regressions if this penalty term is null. Ridge regression has certain advantages over OLS estimates, such as performing better when models have high variance (James et al., 2013), which is highly likely when models become complex; that is, they have a large number of explanatory variables. Ridge regression also shows advantages over OLS when searching for best subset selection. However, ridge regression has one main disadvantage, in that it requires setting the number of regressors in the model in advance; that is, selecting the variables to be included in the model.

The LASSO (Tibshirani, 1996) is a recent statistical tool that is often used as an alternative to ridge regression, which overcomes its main limitation for variable selection, and maintains its advantages over OLS. Although LASSO coefficients are estimated in a similar way as in ridge regression, also including a shrinkage penalty, the LASSO allows coefficients to be exactly zero. Then, the lasso performs variable selection better than ridge regression. LASSO coefficients are estimated to minimize the following expression:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\left(\beta_j x_{ij}\right)\right) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{A1}$$

where y represents the dependent variable, x_j represents the j-th explanatory variable, the first sum is equivalent to OLS, the second sum is the penalty term, and p represents the number of regressors. The selection of an appropriate value of the tuning coefficient $\boldsymbol{\lambda}$ is important, as it determines the effect of the shrinkage. The usual procedure to determine $\boldsymbol{\lambda}$ is to estimate on a grid of values of $\boldsymbol{\lambda}$ and use (cross-validation) prediction errors. A detailed review and discussion of the LASSO, ridge regression, and its comparison to OLS can be found in Chapter 6 of James et al. (2013).

DOI: 10.1111/pirs.12424

Resumen. La investigación sobre los desplazamientos diarios al trabajo ha surgido en las últimas décadas, pero no se ha analizado si las técnicas empíricas utilizadas son las adecuadas. Por lo tanto, los resultados de la investigación previa podrían basarse en modelos que no son precisos, lo que puede llevar a conclusiones erróneas. El estudio aplica un enfoque algorítmico basado en *bootstrap*, la selección de variables y la media absoluta de los errores de predicción, el cuál se ha diseñado para evitar el sobreajuste. El estudió empleó la Encuesta Americana de Uso del Tiempo y se encontró que los modelos con un conjunto reducido de variables explicativas tienen una precisión similar a los modelos econométricos estándar. Nuestros resultados esclarecen la importancia de determinar si los modelos pueden estar sobreajustados.

抄録:最近の数十年の間で通勤に関する研究が行われるようになったが、研究に用いられる実証的な方法が適切か否かについては、これまで分析されていない。すなわち、既存研究の結果は不正確な研究モデルに基づいている可能性があり、誤った結論に至る可能性がある。今回、ブートストラップ法、変数選択、過剰適合にならないように設定した平均絶対誤差率を基にしたアルゴリズムによるアプローチを採用した。American Time Use Survey（米国の生活時間調査）を使用したところ、説明変数が少ないモデルでも標準的な計量経済学的モデルと同等の精度があることがわかった。今回の結果から、モデルが過剰適合になるかを決定することの重要性が明らかになった。