Passenger Travel Demand Models: Factors Underlying Work & Study-Related Travel

Modelli interpretativi della domanda di mobilità: i fattori alla base degli spostamenti per motivi di lavoro e di studio

Giorgio Alleva
Dipartimento di Studi Geoeconomici, Linguistici, Statistici, Storici per l'analisi regionale
Università "La Sapienza" di Roma,
Via del Castro Laurenziano, 9, I0016 Rome, Italy
giorgio.alleva@uniroma1.it

Maria Felice Arezzo
Dipartimento di Studi Geoeconomici, Linguistici, Statistici, Storici per l'analisi regionale
Università "La Sapienza" di Roma,
Via del Castro Laurenziano, 9, I0016 Rome, Italy
mariafelice.arezzo@uniroma1.it

Riassunto: Il lavoro affronta la problematica dei modelli di generazione della domanda, utilizzati allo scopo di stimare gli spostamenti delle persone secondo l'area di origine. A partire da un'applicazione per la stima degli spostamenti per motivi di studio e di lavoro a livello provinciale, basata esclusivamente su variabili aggregate, vengono successivamente proposti modelli fondati sull'integrazione di data base contenenti sia informazioni individuali sia a livello territoriale aggregato. E' presentata un'applicazione condotta con alberi di regressione che ha consentito, a livello regionale, di stimare e di prevedere per un triennio il numero di spostamenti secondo diverse tipologie. Un'altra applicazione, tuttora in corso di realizzazione, riguarda l'integrazione dei dati censuari sul pendolarismo con archivi con informazioni a livello comunale.

Keywords: Transportation Demand Analysis, Trip Generation Models, Database Integration, Classification and Regression Trees

1. Issues, objectives and expected results

The issues dealt with in this paper can be set against the broader frame of reference referred to as transportation demand analysis. A crucial element during the evaluation of alternative infrastructure projects, and in defining the key features of a project (size, location , standard of service), is the determination of the problem such project is expected to solve. In fact, it is widely accepted that demand analysis plays a key role both in feasibility studies and in assessing progress in a program. The aim of demand analysis is, in fact, to identify and determine the existence of a substantial collective need which a single project or a project program intends to address. Thus, demand

analysis precedes feasibility studies and is essential for determining the actions to be taken and their measuring (Mazziotta, 2004).
On the other hand, the determination of the size and characteristics of demand also assumes an instrumental role, which is closely related to the financial and economic analysis of a project. In fact, such determination is necessary to set the prices and rates of the goods and services produced with the project, the amount of revenues or the benefits it generates and therefore its convenience. Demand analysis is also linked to risk and sensitivity analysis. In fact, to measure the strength of the standards of performance of a project it is fundamental to study its reaction to changes in the quantity and quality of demand. Even the interim and final evaluations of the effectiveness of a set of project activities require the existence of suitable demand analysis. As a matter of fact, the results or the effects of the activities performed in connection with a project must always be measured against the initial aims, i.e. the satisfaction of the needs (demand) of the recipients of the goods or services.
In the sector of transport studies (Cascetta, a 2000) a traditional classification of demand models concerns the various types of information that, in sequence or jointly, it is necessary to obtain in order to simulate the behavior of demand (traffic flows) with respect to infrastructure or transport services. Such elements of interest are:

- the estimation of demand originating from a certain point (node) or geographic area, the so-called demand generation model;
- the estimation of the breakdown of demand by origin and destination (estimation of O/D matrix);
- the estimation of demand by mode of transport (so called modal split models);
- the estimation of demand on infrastructural networks (assignment of traffic flows on the graph).

Typically we consider demand with reference to passengers and goods. Depending on the type of information to be obtained or the evaluation/decisions to be made, the analysis is performed on specific types of trip/transport or transported items (passengers or goods). As well as by origin and destination, studies of passengers trips are generally made on the basis of the reasons for traveling (business, study, tourism, etc), by type of passenger (age, education, occupational status, etc.), by radius of the trips (city, different city, other province, foreign country, etc), by frequency (daily, regular, occasional, etc.), by duration or by means of transport. The studies on goods transport can instead be differentiated not only in terms of transport modality but also by radius of the trips (typically the class of distance) by type of goods and by economic activity or other characteristics of the enterprise sending or receiving goods.
In this paper we present :a) different definitions of demand, b) different data-bases to build mobility demand models. In particular, the paper intends to identify the factors underlying transport demand and to explain how they help determine it in order to estimate its changes in time and space. It is important to emphasize that, considering the shortage of periodic information on how people travel with sufficient territorial detail, this paper combines the results of various surveys, not only to take account of individual and environmental variables - therefore related to the characteristics of the area in which the demand is originated - but also in order to use, jointly, variables found through different surveys. Problems, advantages and limits of possible approaches are thoroughly discussed, and a viable solution is recommended on the basis of models founded on the combined use of archives that make it possible to have information available both at the individual level and at various levels of territorial aggregation.

2. Generation models at the provincial level

The first application concerns the estimation, for each of the 103 Italian provinces, of the people that move daily to the usual work or study place in a municipality other than that in which they live. Being aware of the problems due either to residual spatial autocorrelation or to territorial aggregation level (modifiable area problem), our approach makes it possible to identify only those factors that explain people trips. Due to the aggregation level, individual components that explains the decision to travel are not captured by the model.
Nonetheless, this approach, traditionally utilized for trip generation models (Kanafani 1983), is very useful because it gives information that can be utilized for specifying the other models we present in the following notes and can therefore be considered as an exploratory phase to estimate trip generation for work and study reasons. The application made it possible to determine the existence of a single national model more than several models by geographic area and to ascertain its different explanatory capabilities.
The variables we included in the model are those traditionally indicated by the literature for demand generation models In particular, they concern economic and social development levels, service availability and transport means equipment. In detail, variables considered are: added value, activity rate, employment level in different business sectors, education level (Index of Non reaching compulsory school degree, for 15-52 years old people, criminality index, services to families and to firms (expressed as number of employees) Density of Population, stock of vehicles and vehicles matriculations.
All variables, including the dependent one, had been adjusted for the size effect: they were divided either by resident population or number of firms, or employees or provincial extension. All variables refer to year 2001, which is the population census year.
We fitted a stepwise regression which returned six different candidate models. Among them we pick one based either on statistical (goodness of fit, collinearity strength) or logical reasons (estimates signs, results' interpretability). Results are in the following table:

Variables	Unstandardized Coefficients		Sig.	Collinearity Statistics	
	B	Std. Error		Tolerance	VIF
(Constant)	-9,567	6,417	,139		
Activity rate	1,122	,101	,000	,629	1,589
Average Employed in Agriculture, Fishery and Forestry (for 100 residents)	-1,157	,255	,000	,699	1,430
Average Employed in Services (for 100 residents)	-,709	,098	,000	,492	2,031
Index of Non reaching compulsory school degree (15-52 years old people)	-,623	,165	,000	,431	2,322

a Dependent Variable: Resident population traveling outside their town of residence (for 100 residents)

Our results show that

- The activity rate has a positive influence on trip generation; areas with high rates of activity are also characterized by high propensity to travel;
- high levels of employment in agriculture and services are associated with low mobility.
- low education levels also impact in a negative way on the propensity to travel;

In the following table we report the model's goodness of fit index and the results of tests to verify the underlying regression model hypothesis:

R^2_{ADJ}	Durbin-Watson statistics	Condition Index (collinearity)	P-value for $Z_{Kolmogorov-Smirnov}$ statistics
0.729	2.065	51.007	0.818

The residual chart didn't show any particular pattern that might let us suspect heteroskedasticity. Moran's test on residuals didn't show spatial autocorrelation.

The fact that the explained variability is only about 73% of the total can be due to two reasons: first we had to omit interesting variables, such as household consumption, because they are not available at the provincial level of aggregation. Second we couldn't embody in the model those individual features that explain people's traveling choices.

To highlight geographic groups with a similar behavior in terms of the explanatory variables of the regression model, we run a hierarchical cluster analysis using average linkage method.

Three different groups appeared: one includes southern provinces and the islands which is typically set against the group composed of the northern provinces ; quite interesting is the third one which consists of big cities and strongly tourist-oriented provinces. As can be seen in the map, Italian traditional dualism between north and south is present also for the propensity to travel for work or study purposes.

3. Models built on database integration

3.1 Estimating the probability of traveling on the basis of the Multipurpose Istat Survey and Isfort Mobility Observatory

The second approach to transport demand estimation is based on trips classified according to typology, duration and means used. Relevant variables are:

- The *probability to make at least one trip,* either on a regular or an irregular basis and the *probability of trip absence;*
- The *average number of daily trips,* either on a regular or an irregular basis, and the transportation means utilized (Public,Private, etc));
- The *average trip time* either on a regular or an irregular basis.

The work was carried out by linking the yearly Istat Daily Life Survey (so called Multipurpose Survey) to the quarterly Isfort Mobility Survey[1].

Istat's survey makes it possible to analyze trips for study and work reasons while Isfort's survey makes it possible to determine the relationship between regular and irregular trips as well as trip absence[2].

[1] Istituto Superiore di Formazione sui Trasporti

The multipurpose survey gives little information on passengers transport, but gives a very reliable estimate of the number of people who travel *regularly* for work or study purposes. It is repeated every year, allowing for estimates forecasting.
The Isfort survey contains a great deal of information on passengers transport, but is not repeated over a long period of time and we can't therefore use it to understand estimates dynamics.
Models are built using mostly variables measured on individuals even if the main goal of the work is to understand the relationship between trip generation and a set of socio-economic characteristics on a territorial basis. This choice allows to account either for regional or individual features. On the other hand age, number of persons in the household, diplomas or degree, business sector, professional position, family income, transport means, are all variables that can be territorially represented only by synthesizing individual data.
Some territorial variables are also included; for instance, the type of town where the trip starts, population size and density, and capital province distance are variables that can be measured only at a territorial level.
Records linkage was executed through a really complicated preliminary work for homogenizing variables classification and therefore make them perfectly comparable.
The following part is un update of the results in Alleva, Arezzo, Falorsi, Falorsi, 2003.
In terms of methodology, we used a nonparametric model knows as CART (Classification and Regression Trees) (Breinman et al. 1984).
Let's suppose we have two samples, A and B, with size N_a and N_b respectively. On the i-th unit of sample A we measure the variables ($_a\mathbf{X_i}$, Y), with $_a\mathbf{X_i} = (X_{1i}, ... X_{ki,} X_{k+1i}, ..., X_{Mi})$.
For simplicity we assume that X variables can be of any kind, while Y has to be categorical with J possible values.
On sample B we observe the same independent variables as in A namely $_b\mathbf{X} = (X_{1i}, ... X_{ki})$. The goal is to associate to the i-th individual in B a value of the dependent variable Y_i, for i= 1, 2 .... N_b.
Sample A is randomly split into two sub sample L e T.
Using L, *only with regards to the variable in common with sample B,* we build L terminal nodes having size N_l, con l = 1, 2 ... L. These groups, by definition, are homogeneous with respect to variable Y. We then estimate the misclassification rate using T. In other words we use the learning sample L to build the classification rule d(**X**) and the training sample T to estimate the misclassification rate defined as follows:

$$\hat{R}^*(d) = \frac{\sum_{X \in \mathsf{T}} I\big(d(X) \neq j\big)}{N_{\mathsf{T}}}$$

where I is an indicator function assuming value 1 any time the estimated value differs from the true value of Y, while N_T is sample T size.

---

[2] Trip types take the following values: ***1,*** if the subject had a *regular trip* during the day when the measurement was done [in this case there are three conditions that must be satisfied (*i*) the subject had a trip to reach the usual work or study place; (ii) the overall amount of time for traveling in one day is greater than 15 minutes; (iii) subject used any transportation means but bike or feet]. ***2*** , if the subject had a trip during the day when the measurement was done that cannot be classified as regular; ***3*** if the subject during the day when the measurement was done had no trip at all.

For any node, the empirical probability distribution of Y is known:

$$\Pr[Y = i] = \frac{n_i}{N_l} \qquad i = 1,\ 2...J$$

where n_i is the number of individual such that [Y = i] and N_l is the node size.
As we said, the creation of homogeneous groups (terminal nodes) can be done only after a set of decision rules, depending on X, is defined. These rules allow us to assign any individual to one and only one group. Since we know the vector $_b\mathbf{X}$, we can assign any individual belonging to sample B to a terminal node. For this individual, the estimated value of the dependent variable Y is the mode of the node. In other words we know the probability distribution of Y in any terminal node. We choose the mode of Y as estimate of the unknown value for the i-th individual.
A CART algorithm is build to separate groups as much as possible. This means that inside group heterogeneity is minimized and therefore the emerging of a mode, especially if Y can take few values, is quite likely to happen.
Assuming that A and B are drawn from the same population, the misclassification rate for B is still:

$$\hat{R}^*(d)$$

Concerning the estimation of the demand for regional travel, the main goal is to analyze the relationship between trip generation and a set of socio-economic characteristics measured on a territorial basis. That's because, on one hand, we want to predict demand changes (predictive model) and, on the other, we want to identify which variables mostly influence such demand in order to monitor them over time (descriptive model).
The dataset we used to fit the descriptive model is composed of the 14.003 interviews made by the Isfort Mobility Observatory during the four quarters of 2000.
Consistent with factors generally considered in travel demand, the *explanatory variables* were recursively selected from the following list of variables:

Individual variables

- Sex;
- Age
- Municipality Code;
- Car owner
- Motorcycle owner;
- Motorbike owner up to 50 cc;
- Number of persons in the family;
- Number of driving licenses in the family;
- Relationship in the family;
- Marital status;
- Diplomas or degree (highest grade);
- Occupation;
- Professional Position;
- Business Sector;
- Household Income (monthly);

Municipality Variable

- Region;
- Macro Region;
- Population in the Municipality;
- Altimeter;
- Density of Population (Km2);
- Type of Municipality(=1 Province Capital, =0 Non Province Capital);
- Capital Province Distance (km) .

In the following table we report the initial setting and main results for any trials that led to the final descriptive model.

Table 1 - *Initial setting and main results for descriptive model*

Trial	Prior probability assigned to dependent variable values	Minimum size of parent and terminal nodes	Heterogeneity Index	Number of terminal nodes	Proportion of correctly classified individuals per movement type (regular, non regular, trip absence)	Relative cost
1	equal	60 e 25	ord. twoing	32	0.918; 0.517; 0.508	0.558
2	equal	60 e 25	twoing	56	0.920; 0.563; 0.500	0.550
3	equal	60 e 25	class-prob	15	-	0.740
4	equal	60 e 25	symm. Gini	16	0.911; 0.535; 0.439	0.555
5	equal	60 e 25	Gini	81	0.901; 0.499; 0.487	0.550
6	equal	70 e 30	twoing	28	0.919; 0.529; 0.520	0.667
7	0.26; 0.37; 0.30	70 e 30	twoing	24	0.858; 0.566; 0.527	0.652
8	0.28; 0.35; 0.37	70 e 30	twoing	19	0.872; 0.532; 0.476	0.345

It's important to underline that the independent variables indicated as the most important in the final model were always the same throughout all trials.
These variables are reported in the following table:

Table 2 - *Importance of independent variables in descriptive model*

Variable	Importance
Occupation	100.00
Professional Position	94.16
Business Sector	93.99
Age	48.40
Diplomas or degree (highest grade)	18.68
Car owner	18.46
Montly Household Income	15.67
Number of driving licences in the family	14.99
Number of persons in the family	5.94
Relationship in the family	5.12
	...

As can be seen no aggregate variables appear among the most important ones.
Quite interesting is also node composition, displayed in the following table, with regard to propensity to move.

Table 3 - *Composition of some nodes per trip type*

Node	Percentage of occurrences in the node		
	Regular	Non regular	Trip Absence
1	1.27	26.11	72.61
2	8.01	50.26	41.73
3	0.67	33.11	66.22
4	2.14	55.0.8	42.78
5	3.49	40.23	56.28
13	60.28	18.09	21.63
15	47.78	33.79	18.43
19	75.98	9.96	14.06

To better understand the nature of each node, we can take a look at its composition rule. For example, node 1 contains unemployed people (except soldiers and students) less than 65 years old and without vehicles.
The chosen predictive model has seven terminal nodes and was grown imposing the following conditions:

1. Both parent and terminal nodes must contain at least 70 individuals to guarantee reliability to classification;
2. Prior probabilities assigned to the values of the dependent variable are 0.28 for regular trips, 0.35 for non regular trips and 0.37 for absence of trips
3. Purity measure is twoing criterion
4. The original sample was randomly split into a learning sample of size 9406 and a test sample of size 3697 individuals

The following tables respectively reports predictive variables importance and the misclassification matrix associated with the model

Table 4 - *Importance of independent variables in predictive model*

Variable	Importance
Occupation	100.00
Professional position	93.95
Business Sector	93.77
Age	47.63
Diplomas or degree (highest grade)	16.78
Montly Household Income	15.69
Number of persons in the family	5.10
Marital status	3.44
Relationship in the family	2.65
Sex	1.43

The model's predictive performance is good and is in line with that indicated by specialized literature.
Once again, we report nodes compositions according to propensity to travel .

Table 5 - *Misclassification per trip type*

Observed	Predicted			Misclassification rate
	Regular	Non regular	Trip Absence	
Regular	91.15%	8.06%	0.79%	8.85%
Non regular	26.42%	49.64%	23.94%	50.36%
Trip Absence	27.55%	30.07%	42.38%	57.62%

2.2 The estimation of the probability to travel based on Population Census and data base at Municipality level

During June 2005, Istat released the results of the section of the Population Census of 2001 concerning the daily travels for work or study.
The availability every ten years of the information on the origin and destination of the population's trips of the method of transportation, time and duration, together with the individual characteristics of the traveler, represents such a massive amount of knowledge to study the propensity to travel of the population and in particular commuting.
This paragraph presents the guidelines of a study conducted on a sample of 244,000 individuals (resident at the time of the latest Population Census) selected at random from the list of the population aged at least 15, so as to represent a fixed share of 0.5% of this population for each of the 103 Italian provinces[3]. In particular, the main variable of interest is the probability to move daily from the municipality where the person lives to a different municipality.
According to the approach on which the previous models are based, the database is the union of individual data (from the Census) and data at municipality level (from different sources).
The individual data set comprises the following variables.

Individual Variables
Information about the person (that lives habitually in the dwelling)

- If the person goes daily to the usual work or study place
- From which dwelling the person goes to the habitual work or study place
- If the person come back daily in this dwelling from the usual work or study place
- Municipality Code where the person usually lives
- time used to go (only going) to the habitual work or study place
- Method of transportation (the one used for most of the travel distance)
- Sex
- Age
- Marital Status

[3] We thank dott. Aldo Orasi, Central Director of the "Direzione Censimento della popolazione,territorio e ambiente" of ISTAT. Only ISTAT's promptness to process the requested sample will permit the presentation of the results at the Messina Workshop.

- Diplomas or degree (highest grade);
- Post-graduate Degree or Doctorate
- Elementary School Attending, Secondary School Attending, ……
- Training Courses Attending, Technical or Professional courses
- Type of Course
- Occupation;
- Professional Position;
- Full time or Part-time work
- Temporary or Permanent Employment
- Business Sector;

Information about dwelling

- Municipality Code where the dwelling is
- Ownership/rent of the dwelling
- Total Room Number of the dwelling
- Total Surface of the dwelling (M2)
- *Number of the person of the family (that habitually live in the dwelling)*

The database at the municipality level comprises indicators of economic and social development, service availability and transport means equipment. Considering that the dataset was made available only recently, the preparation of the models is still under way and the first results will be presented at the Messina Workshop.

References

Alleva G., Falorsi P.D. Falorsi S., Arezzo M.F. (2003) Modelli interpretativi e previsivi della domanda di trasporto locale, *Working Paper del Dipartimento Studi Geoeconomici, Linguistici, Statistici e Storici per l'Analisi Regionale*, 2003

Breiman L., Friedman J., Olshen R., Stone C. (1984) *Classification and Regression Trees*, Wadsworth

Cascetta E. (2000) *Ingegneria dei trasporti*, Utet

Hastie T., Tibishirani, R., Friedman J. (2002) *The elements of statistical learning, Data minimg, inference and prediction*, Springer

Kanafani A. (1983) *Transportation Demand Analysis*, Mc Graw Hill Book Company

Mazziotta C. (2004) L'analisi della domanda negli studi di fattibilità, in *Investimenti pubblici e processo decisionale*, Formez, Strumenti n.18

Moura F.A.S., Holt D. (1999), Small area estimation using multilevel models, *Survey methodology*, 25**,** 73-80.

Purcell N.J., Kish L.(1980) Postcensal estimates for local areas, *International Statistical Review*, 48**,** 3-18.