

Methods for the Analysis and Estimation of Household Waste¹

Modelli per l'analisi e la stima dei rifiuti urbani domestici

Mara Cammarrota

Istat, Via A. Ravà 150, 00142 Roma, Italy, email: cammarro@istat.it

Teresa Di Sarro

Istat, Via A. Ravà 150, 00142 Roma, Italy, email: tdisarro@istat.it

Giovanna Jona Lasinio

DSPSA, Università di Roma "La Sapienza", P.le A. Moro 5, 00185 Roma, Italy, email: giovanna.jonalasinio@uniroma1.it

Riassunto: Per rispondere agli obblighi richiesti dal Regolamento europeo n. 2150/2002 sulle statistiche sui rifiuti, un modello di stima per i rifiuti urbani prodotti dalle famiglie è stato proposto considerando variabili non domestiche (Cammarrota *et al.*, 2005). Scopo di questo lavoro è individuare una procedura per la validazione ed il miglioramento di tale modello. In particolare, una procedura di crossvalidation (*Leave One-Out Crossvalidation*, LOOCV) viene applicata sia sui dati relativi ai rifiuti urbani per tutte le province italiane sia sulle province appartenenti a tre cluster individuati con un metodo non gerarchico (*Partitioning Around Medoids*, PAM). I buoni risultati conseguiti confermano la validità del modello per la stima dei rifiuti domestici e per la ricostruzione di possibili dati mancanti sui rifiuti urbani.

Keywords: model validation, household waste, linear models

1. Introduction

The Regulation 2150/2002/EC on Waste Statistics requires Member States, among other things, data related to household waste. In Italy a problem occurs for these data as the Italian collection system of municipal waste is organised over the municipal territory with waste containers available to everyone with no mean to distinguish between different possible users. Then municipalities report the entire amount of waste collected including both household waste and waste coming from trade and services activities. Moreover, tourists who spent their holidays in given cities can put their waste in the same containers. In order to face this problem, a simple regression model was set to estimate *pure* household waste starting from selected variables correlated with municipal waste and has been proposed in Cammarrota *et al.* (2005). In this paper we focus on the validation and improvement of the proposed model. As a by product a way to estimate missing data is obtained.

¹ The present paper is financially supported by MIUR COFIN 2004 project Statistical methods for forecasting improvement in environmental sciences and the exploration and modelling of complex phenomena.

2. Model's improvements and validation

In Cammarota *et al.* (2005) the proposed model was:

$$y_i^* = a + \alpha_1 x_{1i} + \alpha_2 x_{2i} + e_i \quad i=1, \dots, n \quad (1)$$

where n is the number of sites, y_i^* is the amount of collected municipal waste in tonnes, x_{1i} is the number of trade and services employees per inhabitant with reference to activities belonging to sections from G to O of NACE Rev. 1, and x_{2i} is the tourism pressure ratio tourist overnights stays /residents $\cdot (365-k)$, where 365 is the number of days in a year and k is the estimated average number of days spent by residents outside the considered area (Cammarota *et al.*, 1999). However, after a deep data analysis, we decided to consider the amount of collected municipal waste per capita, as this transformation allows for more stable results. Then the actual model is given by:

$$y_i = a + \alpha_1 x_{1i} + \alpha_2 x_{2i} + e_i \quad i=1, \dots, n \quad (2)$$

where y_i is the amount of collected municipal waste per capita (in kilograms). The part of the municipal waste not explained by model can be ascribed to *pure* household waste (Cammarota *et al.*, 2005). Then $z_i = (e_i + a)$ is the estimation of the household waste, where e_i are residuals from model and a is the intercept of the same model.

In order to validate the proposed model and investigate further improvements we perform the following steps:

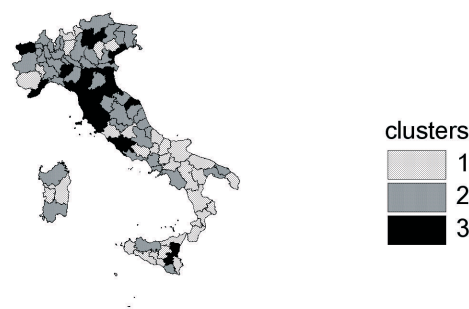
1. Leave One Out Crossvalidation (LOOCV): as household waste are not known we use the per capita amount of municipal waste as control value. More precisely we first estimate household waste for 102 out of 103 Italian provinces, then we average such values ($\bar{z}$) and we compute the per capita total waste amount of the removed one as: $\hat{y}_j = \bar{z} + \alpha_1 x_{1j} + \alpha_2 x_{2j}$. We repeat this step 103 times and compute MSE.
2. Clusters' based procedure: we build 3 clusters of Italian provinces according to model variables using a non-hierarchical clustering method, the Partitioning Around Medoids (PAM), chosen for its robustness with respect to outliers and extreme values.
3. Based on the 3 clusters and always through a crossvalidation type procedure, we verify which of the following approaches returns better results in terms of MSE: (a) one model for all provinces, we remove 3 provinces (one for each cluster), we compute the average household waste for each cluster ($\bar{z}_k, k=1,2,3$) and use these values to compute $\hat{y}_j$ for the left-out provinces; (b) one model for each cluster, we remove one province for each cluster, we compute $\bar{z}_k, k=1,2,3$ and use these values to estimate $\hat{y}_j$.

All analysis have been performed with reference to the year 2001 for the 103 Italian provinces using waste data published by the Italian Agency for Environmental Protection and Technical Services (APAT), whereas the other data come from Istat (Indagine trimestrale telefonica sui viaggi degli italiani, Statistiche del turismo, Censimento dell'industria e dei servizi).

From step 1 of the validation procedure we obtain a root MSE of 83.77 kg per inhabitant which corresponds to a minimum of 10.24% to a maximum of 28.94% error.

From step 2 and 3 we can see that a unique model at national level (approach (a)) does not perform better even if associated to averages computed on each cluster (root MSE equal to 102.33 kg per inhabitant). On the other hand using approach (b) we obtain a strong reduction of the root MSE (46.57 kg per inhabitant). The cluster based analysis deserves further attention. In Figure 1 Italian provinces are mapped according to the PAM classification procedure. The three clusters are composed respectively by 32, 46 and 25 provinces (Figure 1).

Figure 1: *Italian Provinces according to PAM classification (data for the year 2001)*



In Table 1 summary statistics for the clustering variables in each cluster are reported. It appears an ordering of the three clusters according to the chosen variables, cluster 1 includes provinces with the lowest variables levels and cluster 3 the ones with highest variables levels.

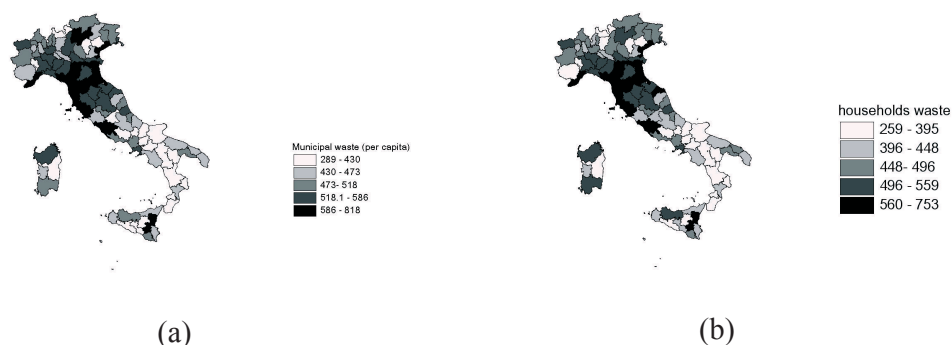
Table 1: *Summary statistics for Italian provinces' clusters*

	Minimum	1 st Quantile	Median	Mean	3 rd Quantile	Maximum
Cluster 1						
Municipal waste	289.4	388.5	416.5	404.9	436.4	454.8
Trade and services	0.06534	0.07897	0.08693	0.09638	0.10991	0.15302
Tourism	0.001321	0.00381	0.006774	0.008158	0.008942	0.034999
Cluster 2						
Municipal waste	458.6	474.1	496	503	529.2	567.2
Trade and services	0.07251	0.1207	0.13482	0.1424	0.1527	0.35994
Tourism	0.001894	0.004899	0.010144	0.018231	0.022388	0.148905
Cluster 3						
Municipal waste	572.3	605.9	644.8	654.1	679.6	817.9
Trade and services	0.09025	0.15147	0.16829	0.17267	0.18812	0.28824
Tourism	0.004276	0.017958	0.027033	0.041381	0.061474	0.16429

From approach (b) we obtain errors ranging from 8.95% to 14.06% for the first cluster, from 6% to 7.45% for the second one and from 8.55% to 11.94% for the third cluster. From these results we choose to estimate household waste per inhabitant by estimating model (2) for each cluster and using models' residuals and intercepts to obtain the desired values. In Figure 2 we report municipal waste (a) and estimated household waste (b) (both expressed in kg per capita) of Italian provinces mapped according to their quantiles (0, 0.2, 0.4, 0.6, 0.8, 1). From the maps it can be clearly seen as the

geographical pattern of households waste does not follow exactly the one of municipal waste.

Figure 2: *Municipal waste (a) and estimated household waste (b) (in kg per capita)*



3. Conclusions

Results obtained by crossvalidation and stratification show that the proposed model for the estimation of household waste gives coherent results. This approach can be useful for the reconstruction of missing data too. In particular, when municipal waste are missing in a given province, it will be possible to estimate them starting from the average value of belonging strata. One problem occurs for the yearly data availability of the regressors in model (2). Basic data to build the tourism pressure indicator are available every year from Istat sources. However the number of trade and services employees comes from Census 2001 of industry and services and no other sources provide yearly data for such variable. The Statistical archive of undertakings (Asia) created by Istat, covers activities from Sections C to O, excluding L, of NACE Rev.1. Looking forward to creation of Asia Institutions, that will cover Section L, a possible source of information for the number of trade and services employees could be the administrative source of waste represented by the Compulsory Declaration (MUD, Modello Unico di Dichiarazione Ambientale) (Cammarota *et al.*, 2005) and our current work is focused on this estimation problem.

References

- Cammarota M., Costantino, C., Fangstrom, I. (1999) Joint Final Report Of The Sectoral Infrastructure Project-Tourism, in: *Eurostat, Towards environmental pressures indicators for the EU: an examination of the sectors*, pp. 233-284, Luxembourg.
- Cammarota M., Jona Lasinio G., Di Sarro T. (2005) A proposal for the estimation of household waste, in *Atti del Convegno intermedio SIS 2005, Statistica e ambiente*, Messina, 21-23 settembre 2005, pp. 215-218.
- Regulation (EC) N. 2150/2002 of the European Parliament and of the Council of 25th November 2002 on waste statistics, Official Journal of the European Communities No. L 332/6 of 9.12.2002.