Modelling the spatio-temporal dynamic of traffic flows with gravity models and mobile phone data

Maurizio Carpita, Rodolfo Metulini

1. Introduction

The analysis of origin-destination traffic flows may be useful in many contexts of application and have been commonly studied through the *Gravity Model* (Tinbergen, 1962). The popularity of Tinbergen's log-linear specification of the Gravity Model is due to its good performance in modelling international trade flows and to the strong theoretical foundations provided in papers such as Anderson (1979) and Anderson & Van Wincoop (2003). At the macro-level, this model states that the volume of trade between any two countries is proportional to the product of their gross domestic products (GDP) and a distance deterrence function, where distance is broadly construed to include all factors that might create trade resistance. The Gravity Model equation can be straightforward translated to micro-level flow data, such as, for example, passenger flows, simply by substituting trade flows with the total number of passenger flows from two cities, a measure of dimension of the city of origin and of the city of destination (such as their population) instead of GDP, and the geographical (or network) distance among the two cities in place of trade resistance.

Using data on the flow of mobile phone signals of TIM (*Telecom Italia Mobile*) users among different *census areas* (ACE of ISTAT, the *Italian National Statistical Institute*), recorded on hourly basis for six months, in this preliminary study we model such a flows in the *Mandolossa* to predict flows' intensity during flood episodes in the context of smart cities emergency management plans. Traffic flows data can be integrated to mobile phones densities and used to develop dynamic exposure to flood risk maps, as proposed in Balistrocchi et al. (2020). From a prevention perspective, this could make the identification of preferential traffic flows possible, thus evidencing potential risks during inundation onsets or emergency situations.

Whereas, as explained above, for the classical Gravity Model a traditional *static* mass explanatory variable is represented by GDP or by residential population (Kepaptsoglou et al., 2010) also thanks to the availabiity of a time series of data, we propose to use a most accurate set of explanatory variables in order to better account for the *dynamic* over the time. First, we employ a time-varying mass variable represented by the density of TIM users by area and by time period, which has been estimated from mobile phone data using the method proposed by Metulini & Carpita (2021) and adopted by Balistrocchi et al. (2020) to derive crowding maps for flood exposure. Second, a proper set of time effects is included. We show that the joint use of these two novel sets of explanatory features allow us to obtain a better linear fit of the Gravity Model and a better traffic flow prediction for the flood risk evaluation.

2. The mobile phone flows and the other datasets

The TIM mobile phone flows used in this study has been provided by Olivetti (*www.olivetti.com/en/iot-big-data*) and FasterNet (*www.fasternet.it*), for the development of the MoSoRe

Maurizio Carpita, University of Brescia, Italy, maurizio.carpita@unibs.it, 0000-0001-7998-5102
Rodolfo Metulini, University of Salerno, Italy, rmetulini@unisa.it, 0000-0002-9575-5136

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Maurizio Carpita, Rodolfo Metulini, *Modelling the spatio-temporal dynamic of traffic flows with gravity models and mobile phone data*, pp. 99-104, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-461-8.19, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci (edited by), *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the on-site conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-461-8 (PDF), DOI 10.36253/978-88-5518-461-8

Project 2020-2022 co-founded by Lombardy Region (*bit.ly/2Xh2Nfr*), and has been used at the DMS StatLab of the University of Brescia (*dms-statlab.unibs.it*).

The original data flows are square origin-destination (OD) matrices of dimension $N \times N$, where $N = 235$ represents the number of *census areas* or ACE (*Aree di Censimento*, using the standard definition of ISTAT) in the Province of Brescia, available at each hour's interval for six months from September 2020 to February 2021, so the length of the time series is 24×181 ($T = 4,344$). Furthermore, ISTAT provided the shape files for SCE (*Sezioni di Censimento*), with additional information about the belonging to their ACE and its area (*www.istat.it/it/archivio/104317*).

We restrict our attention to a particular subset of OD matrices, as the core of the analysis regards the area of the *Mandolossa*, which has been identified with 4 ACE (Brescia Mandolossa, Cellatica, Gussago and Rodengo Saiano) intersecting with the identified flooding-risk area (return period of 10 years), as reported in the left chart of Figure 1. We choose other 38 neighboring ACE aggregated as represented in the map, which fulfil the criteria of having a minimum (considering the four ACE of the Mandolossa) outflow of 10 in both three sample days chosen randomly. The total flows counted between the 4 Mandolossa's ACE and the 38 selected neighboring ACE counts for about the 84% of the total outflows from the Mandolossa's ACE.

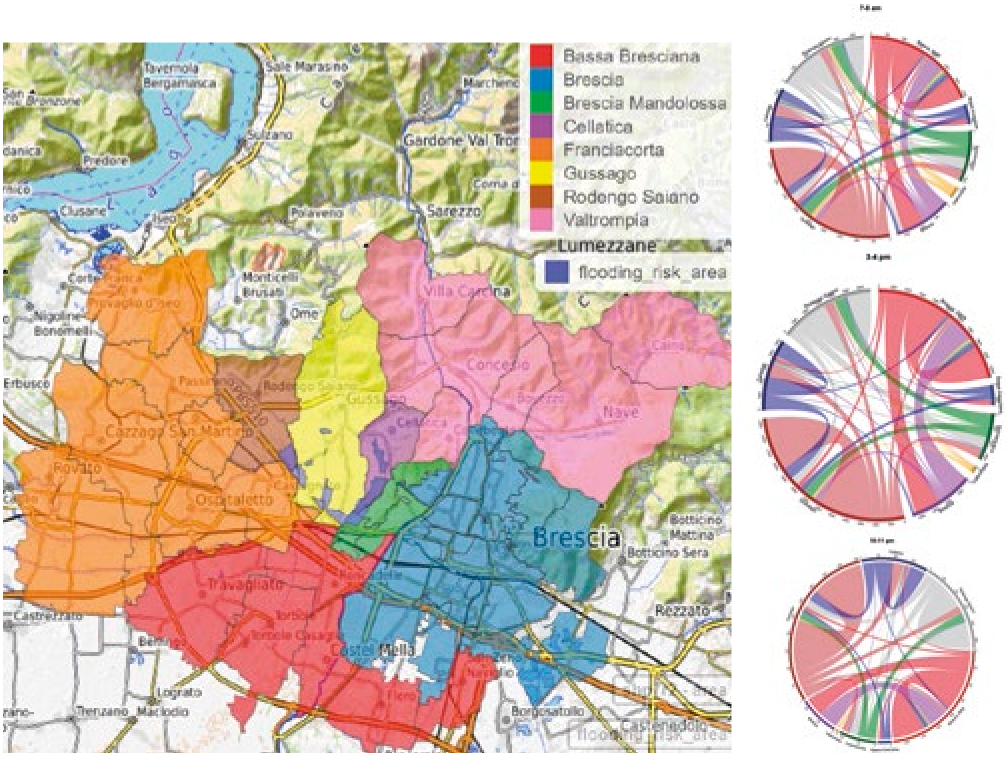

Figure 1: Map of flooding risk area, ACE in Mandolossa and neighboring (by macro-area) (left). Kriskograms of TIM flows between the eight macro-areas (right).

The three kriskograms in Figure 1 show flow between the 8 macro-areas of interest at three different hours (7-8 am, 3-4 pm, 9-10 pm): the diameter of the circles (proportional to the total flow) highlights that flows increase from morning to afternoon and decrease from afternoon to evening, and that for the four ACE of the Mandolossa flows are internal flows are very high. As show in Section 4, these evidences have suggested to introduce in the Gravity Model a parabolic effect for *hour* and a dummy effect for three internal flows.

About Gravity Model's variables, we collect ISTAT data on residential *population* in each SCE (and, by aggregation, each ACE) at January 1st, 2016, and on the *distance* in km between the centroids of the 4 Mandolossa's ACE and the other 38 neighboring ACE. Furthermore, to extend the classical Gravity Model we have used the *mobile phone density* of TIM users, computed for each hour and ACE of interest, which can be interpreted as the average number of mobile phones simultaneously connected to the TIM network in that area in that time interval (Carpita & Simonetto, 2014). These data are created by Metulini & Carpita (2021) and used in Balistrocchi et al. (2020) for the analysis of the Mandolossa in the period 2014-2016. As the mobile phone densities for 2020 and 2021 are not yet available, we have used as proxy the data in the same month, hour and day of the week of 2015 (from September to December) and 2016 (for January and February).

3. The Gravity Model and its extension

The classical Gravity Model states that *flows* from origin i to destination j (F_{ij}) are proportional to *masses* of both origin and destination (M_i and M_j) and inversely proportional to *distance* between them (D_{ij}), where G and γ are positive constants:

$$F_{ij} \propto G \cdot \frac{M_i \cdot M_j}{{D_{ij}}^{\gamma}} \tag{1}$$

Assuming masses as functions of *Populations* (P_i and P_j), the Gravity Model can be linearised using the logarithmic transformation of (1) and specified as a multiple linear regression model with a temporal dependence subscript t (in our case the hour), with random errors ϵ_{ijt} (LeSage & Pace, 2009):

$$\log(F_{ijt}) = \alpha + \beta_1 \cdot \log(P_i) + \beta_2 \cdot \log(P_j) - \gamma \cdot \log(D_{ij}) + \epsilon_{ijt} \tag{2}$$

Model (2) can be extended introducing as other explanatory variables the dynamic masses (dependent from t) *mobile phone densities* (MP_{it} and MP_{jt}), the fixed effect for *Internal flows* (IF_{ij}) and a vector of pure *Time effects* ($\mathbf{TE}_t$), with parameters α, β_1, β_2, γ, δ_1, δ_2, ω and $\boldsymbol{\lambda}$ that must be estimated:

$$\begin{aligned} \log(F_{ijt}) &= \alpha + \beta_1 \cdot \log(P_i) + \beta_2 \cdot \log(P_j) - \gamma \cdot \log(D_{ij}) + \\ & \quad \delta_1 \cdot \log(MP_{it}) + \delta_2 \cdot \log(MP_{jt}) + \omega \cdot IF_{ij} + \boldsymbol{\lambda}^T \mathbf{TE}_t + v_{ijt} \end{aligned} \tag{3}$$

It must be considered that this traditional log-linear specification of the Gravity Model along with Ordinary Least Squares (OLS) estimation method can be inappropriate when bilateral flows are frequently zero. Many studies estimate the log-linear model on samples of observations using the truncated OLS approach but, by disregarding pairs of observations that do not have a positive flows with each other can generate biased estimates (Helpman et al., 2008). Silva & Tenreyro (2006) have shown that log-linearisation of the Gravity Model leads to inconsistent estimates in the presence of heteroscedasticity in flows levels, and propose a Poisson specification along with the Poisson Pseudo Maximum Likelihood (PPML) estimator. However, when just interested on the flows between areas with positive flows, as in our explorative case study, it is possible to rely on OLS without any loss in estimation efficiency.

4. Application and preliminary results

The parameters of the classical Gravity Model (2) and its extension (3) presented in Section 3 have been estimated using the standard OLS method using data described in Section 2. For this preliminary study, a sample of flows of 6 hours (7,10,13,15,18,21) and 4 days of the week (Monday, Wednesday, Thursday and Saturday) for the six months from September 2020 to February 2021 has been extracted from the 4 Mandolossa's ACE and the 38 neighboring ACE. Then, this sample of 6,912 observations has been randomly partitioned in *training set* (6,000 observations) used for estimation and *test set* (912 observations) used to evaluate prediction performance.

To assess the goodness of fit of the four models considered in this preliminary analysis, Residual standard error and adjusted R^2 have been used, whereas the AIC (Akaike's information criterion) for the training set and the correlation between observed and predicted flows ($\mathrm{Cor}(Y,\hat{Y})$) for the test set have been used to assess prediction performance. The F tests of significance for the parameters of the considered (*full*) model and for the model included (*nested*) in the considered model are reported too.

Table 1 shows preliminary results for the four Gravity Models described in the previous section. `MOD1`, the classical Gravity Model in formula (1) with only *Population* and *Distance* as explanatory variables, has statistical significance (t and F tests have zero p-value), but rather low goodness of fit (adjusted R^2 is 34.5%) and prediction performance (for the test set, correlation between observed Y and estimated $\hat{Y}$ flows is 0.595); as expected, the estimated effects on the *Flows* are positive for *Population* and negative for *Distance*. `MOD2`, that includes *Mobile phone density* as explanatory variables, has statistical significance (F test reject the nested model `MOD1`), but doesn't improve substantially the fit (adjusted R^2 is 34.9%) and has the same prediction performance of `MOD1` (but AIC is a little lower and $\mathrm{Cor}(Y,\hat{Y})$ for the test set is 0.594). When the dummy for the three *Internal flows* is added to the model (see the end of Section 2), results noticeably improve: for `MOD3` the F test reject the nested model `MOD2`, the fit gets better (adjusted R^2 is 53.1%) and prediction performance increases (AIC decreases a lot and $\mathrm{Cor}(Y,\hat{Y})$ for the test set is 0.741); note that the presence in the model of *Internal flows* strongly reduce the effect of *Distance* (from -0.186 to -0.06) and slightly increase the effects of the two *Mobile phone density* on *Flows*. Finally, the introduction of the temporal effects as in `MOD4` further improves the results: the F test reject the nested model `MOD3`, adjusted R^2 is 62.7%, AIC decreases further and $\mathrm{Cor}(Y,\hat{Y})$ for the test set is 0.808. *Hour* has the expected significant and parabolic effect on *Flows* (increasing from morning to afternoon and decreasing from afternoon to evening), *Day of the week* has a significant and negative effect for Saturday and *Month* has significant and negative seasonal effects, i.e. flows are lower in Autumn and Winter with respect to September: this rather unexpected effect may have been caused by the limitations caused by the COVID19 pandemic that began in October 2020. Note that introducing the time effects doesn't change substantially the parameter estimates for the other regressors respect to `MOD3`.

5. Concluding remarks

Using data on the flow of mobile phone signals of TIM users among different ISTAT census areas the classical Gravity Model and some its extensions have been preliminarily adopted to study dynamic of such flows over the time in the *Mandolossa*, an area at the western outskirts of Brescia in northern Italy, with the final aim of predicting the traffic flow during flood episodes.

In addition to the usual population and distance regressors, the joint use as explanatory variables in the model of time-varying mass variable represented by the density of TIM users by area and by time period and a proper set of temporal effects allow us to obtain a better linear fitting with respect to the classical Gravity Model, and a better traffic flow prediction for the flood risk evaluation. These preliminary results are promising, but some in-depth analyses have yet to be carried out. As explained at the end of Section 3, it will be important to evaluate

Table 1: Preliminary results of four Gravity Models for the Mandolossa flows

Regressors	MOD1	MOD2	MOD3	MOD4
Population origin	1.023***	0.891***	0.730***	0.694***
Population destination	1.027***	0.851***	0.706***	0.671***
Distance in km	−0.186***	−0.186***	−0.060***	−0.060***
Mobile phone density origin		0.155**	0.231***	0.279***
Mobile phone density destination		0.196***	0.261***	0.301***
Internal flows			1.576***	1.578***
Hour				0.559***
Hour^2				−0.019***
Day of the week (reference: *Monday*)				
Wednesday				0.074*
Thursday				0.061$^{\cdot}$
Saturday				−0.269***
Month (reference: *September*)				
October				−0.088*
November				−0.273***
December				−0.350***
January				−0.291***
February				−0.250***
Constant	−18.992***	−19.156***	−18.306***	−21.804***
Residual standard error	1.138	1.134	0.963	0.858
Degrees of freedom	5,996	5,994	5,993	5,983
Adjusted R^2	0.345	0.349	0.531	0.627
F test full model	1,053***	644***	1,133***	632***
F test nested model		19.752***	2,329***	156***
AIC training set (6,000 obs.)	18,583	18,549	16,581	15,211
Cor(Y,$\hat{Y}$) test set (912 obs.)	0.595	0.594	0.741	0.808

Notes: For all the models, the variables flows, population, distance and mobile phone densities are in logarithms.
Parameter estimates have been obtained using the standard OLS method.
Significance codes for t and F tests: . $p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

the possibilities offered by the most appropriate estimation methods; moreover, the actual predictive capacity of the model for the purposes of the MoSoRe Project will have to be further investigated.

Finally, we are also evaluating to introduce in the Gravity Model other non-standard explanatory variables, related to to the number and the type of streets, the number of offices, restaurants or cinemas, which may be retrieved from `OpenStreetMap`, would allow to better characterize the areas of interest and further improve the model performance.

Future use of 5G and GPS technologies will facilitate the real-time assessments of the spatial distribution of people: with an early-warining system, alternative safe pathways could be identified and communicated to exposed people in order to facilitate their evacuation.

Acknowledgments

Research carried out at the DMS StatLab, University of Brescia (*dms-statlab.unibs.it*), co-funded by *MoSoRe@UniBS* (Infrastrutture e servizi per la Mobilita' Sostenibile e Resiliente) Project of Lombardy Region, Italy (CallHub ID 1180965; bit.ly/2Xh2Nfr). We acknowledge anonymous reviewer for the fruitful comments.

References

Anderson, J.E. (1979). A theoretical foundation for the gravity equation. *The American Economic Review,* **69**(1), pp. 106–116.

Anderson, J. E., Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review,* **93**(1), pp. 170–192.

Balistrocchi, M., Metulini, R, Carpita, M., Ranzi, R. (2020). Dynamic maps of human exposure to floods based on mobile phone data *Natural Hazards and Earth System Sciences,* **20**, pp. 3485–3500.

Carpita, M., & Simonetto, A. (2014) Big data to monitor big social events: Analysing the mobile phone signals in the Brescia smart city. *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation,* **5**(1), pp. 31–41.

Helpman, E., Melitz, M., Rubinstein, Y. (2008). Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics,* **123**(2), pp. 441–487.

Kepaptsoglou, K., Karlaftis, M.G., Tsamboulas, D. (2010). The gravity model specification for modeling international trade flows and free trade agreement effects: a 10-year review of empirical studies *The Open Economics Journal,* **3**(1), pp. 1–13.

LeSage J., Pace R.K. (2009). Introduction to Spatial Econometrics. *Chapman & Hall/CRC, New York (NY)*

Metulini, R, Carpita, M. (2021). A spatio-temporal indicator for city users based on mobile phone signals and administrative data *Social Indicators Research,* **156**(2-3), pp. 761–781.

Silva, J.S., Tereyro, S. (2006). The log of gravity. *The Review of Economics and Statistics,* **88**(4), pp. 641–658.

Tinbergen, J. (1962). Shaping the world economy; suggestions for an international economic policy *Twentieth Century Fund, New York (NY).*