Geographically weighted regression for spatial network data: an application to traffic volumes estimation

La regressione geografica ponderata per dati su network spaziali: un'applicazione alla stima dei volumi di traffico

Andrea Gilardi, Riccardo Borgoni and Jorge Mateu

Abstract Estimating traffic volumes on road networks represents a critical issue in various areas of research such as transport studies and road safety analyses. In these cases, the traffic figures are usually recorded via sparse manual counts or expensive automatic tools (e.g. cameras or inductive loops). However, given the increasing availability of mobile sensors (e.g. smartphones and GPS sat-nav), in the last years several methods were developed to extract traffic information from geo-referenced mobile devices. This paper proposes a geographically weighted regression (GWR) approach to combine fixed counts and GPS data to estimate traffic flows, re-adapting the appropriate statistical methods to the spatial network context. The suggested methodology is exemplified using data collected in the City of Leeds (UK).

Abstract La stima dei volumi di traffico rappresenta un problema rilevante in diversi ambiti come la mobilità urbana e le analisi sulla sicurezza stradale. I dati sul traffico vengono solitamente ottenuti da conteggi manuali o costosi strumenti automatici (e.g. telecamere o spire induttivi). Tuttavia, data la sempre maggiore disponibilità di sensori mobili (come smartphone e dispositivi GPS), negli ultimi anni sono stati sviluppati diversi approcci per ricavare stime di traffico da devices portatili. In questo articolo si propone un modello di regressione geografica ponderata (GWR) per la stima dei volumi di traffico che unisce dati GPS a conteggi manuali, riadattando la metodologia alla rete stradale. Il lavoro viene testato analizzando i conteggi stradali registrati nella città di Leeds (UK).

Key words: Geographically weighted regression, Linear networks, Network analysis, Traffic volumes estimation

Andrea Gilardi, Riccardo Borgoni
Department of Economics, Management and Statistics, University of Milano - Bicocca, e-mail: andrea.gilardi@unimib.it; e-mail: riccardo.borgoni@unimib.it

Jorge Mateu
Department of Mathematics, Universitat Jaume I , e-mail: mateu@uji.es

1 Introduction

Estimating traffic volumes on urban networks represents a critical issue in several areas of research such as transport studies [3], road safety analyses [6], and investigations on street networks efficiencies [5]. In these cases, the traffic flows can be used to quantify transportation demand, simulate driving and commuting behaviours, or approximate road risk exposure.

The traditional ways to compute traffic figures involve manual counts with ad-hoc cameras or automatic counts with road-fixed sensors (e.g. inductive loops and spirals). Unfortunately, both techniques have several limitations linked to their limited spatial coverage, high economical costs of installation and maintenance, and error proneness. For these reasons, in the last years several authors explored different approaches to extract traffic information from geo-referenced mobile sensors (e.g. smartphones and sat-navs), creating a complementary way to estimate the road counts. The mobile sensors have several benefits, such as extremely detailed spatial resolution and (usually) extensive coverage. However, since not all vehicles driving in a road network are equipped with GPS devices, the figures inferred from the data may actually underestimate the real traffic flows.

Hence, in this paper we propose a geographically weighted regression (GWR) approach to combine the two data sources (i.e. classic road counts and GPS figures) into a unique traffic estimate. Moreover, considering that traffic flows measurement from fixed and mobile data represents a classical example of a phenomenon occurring in a spatial network, we re-adapt the suggested statistical technique to this particular spatial domain.

2 Geographical weighted regression for network data

GWR is a local form of spatial analysis that allows the estimation of relationships between a dependent variable and a set of predictors that vary over space [4]. More precisely, given a sample of n units in a region S observed at locations $\mathbf{s}_i,\ i = 1,\dots,n$ according to a given coordinate reference system, the GWR model writes as

$$y(\mathbf{s}_i) = \alpha(\mathbf{s}_i) + \mathbf{x}'(\mathbf{s}_i)\beta(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i), \quad i = 1,\dots,n, \tag{1}$$

where $y(\mathbf{s}_i)$ denotes the response variable, $\alpha(\mathbf{s}_i)$ is the intercept, $\mathbf{x}(\mathbf{s}_i)$ a q-column vector of explanatory covariates, $\beta(\mathbf{s}_i)$ are the corresponding spatially-varying coefficients, and $\varepsilon(\mathbf{s}_i)$ is a zero mean random error. Since the parameters depend upon the spatial locations, this approach permits one to map the variation in the regression coefficients, gaining understandings of the spatial patterns between the predictor and response variables.

Parameter estimation at a selected location $\mathbf{s}_j \in S$ can be carried out using locally weighted least squares

$$\hat{\beta}(\mathbf{s}_j) = \left[\mathbf{X}'(\mathbf{s})W(\mathbf{s}_j)\mathbf{X}(\mathbf{s})\right]^{-1}\mathbf{X}(\mathbf{s})'W(\mathbf{s}_j)y(\mathbf{s}_j), \tag{2}$$

where $W(\mathbf{s}_j) = diag(w_{1j},\ldots,w_{nj})$ is a local weighting square matrix, w_{ij} is the weight associated to unit i when the regression is estimated at location $\mathbf{s}_j$, and $\mathbf{X}(s)$ represents the design matrix. The weights are defined in terms of a kernel function K that decays gradually with d_{ij}, i.e. the distance between the ith observation and the point $\mathbf{s}_j$. In particular, a Gaussian kernel function is adopted in this paper

$$K(d_{ij}) = \exp\{-d_{ij}^2/2h\}, \tag{3}$$

where the bandwidth parameter h determines the spatial range of the kernel. In the case study presented in the next section, the value of h is selected using cross-validation by minimising the mean square error of traffic flows predictions.

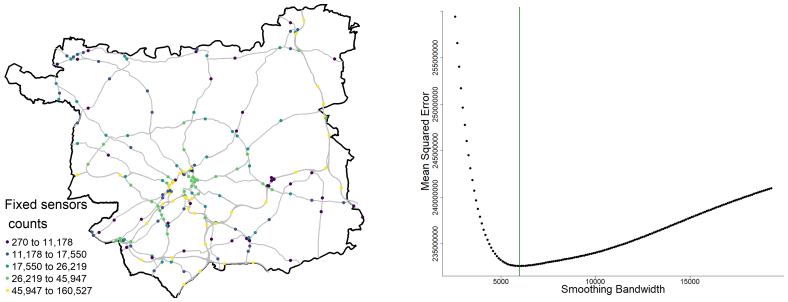
Usually, the inputs in Equation (3) are Euclidean distances in a planar setting, e.g. $d_{ij} = ||\mathbf{s}_j - \mathbf{s}_i||$. However, in our context, the sample units are observations recorded at a set of n road segments represented by their centroid locations $\mathbf{s}_i, i = 1,\ldots,n$. Hence, the distance d_{ij} should be calculated preserving the graph structure of the road network. In the rest of the paper we refer to the shortest path distance in order to take into account the spatial domain of the data. More precisely, indicating by $L = (V,E)$ the one-dimension graph object generated by the street network (where V and E denote the sets of vertices and edges, respectively), a path ρ_{ij} connecting any two generic locations $\mathbf{s}_i$ and $\mathbf{s}_j$ on the network is defined as a finite sequence $\{\mathbf{p}_m\}_{m=1}^{M}$ of adjacencent vertices in V such that the edges with endpoints $[\mathbf{s}_i, \mathbf{p}_1]$ and $[\mathbf{p}_\mathbf{M}, \mathbf{s}_j]$ belong to E. The length of ρ_{ij} can be computed as

$$||\mathbf{s}_i - \mathbf{p}_1|| + \sum_{m=1}^{M-1} ||\mathbf{p}_{m+1} - \mathbf{p}_m|| + ||\mathbf{p}_M - \mathbf{s}_j||,$$

and we define d_{ij} as the minimum length of all paths connecting $\mathbf{s}_i$ and $\mathbf{s}_j$ [1, 2].

3 Estimation of traffic flows in Leeds: data and results

The case study considered in this section is based on fixed and mobile daily traffic volumes recorded in the road network of Leeds (UK) from January to December 2019. The network and the GPS counts, which represent the spatial domain and the covariate used in our model (see Equation (1)), were obtained from TomTom Move service (`https://move.tomtom.com/`). The spatial network is composed by 8959 geo-referenced segments that are associated to traffic volumes estimated using mobile devices connected to cars and anonymous GPS-equipped smart-



(a) (b)

Fig. 1 Leeds road network and locations of fixed cameras used to detect traffic counts by the Department for Transport (a); MSE curve for bandwidth cross validation (b).

phones. Hence, the TomTom data have a reasonable spatial coverage, although they are known to underestimate the real flows.

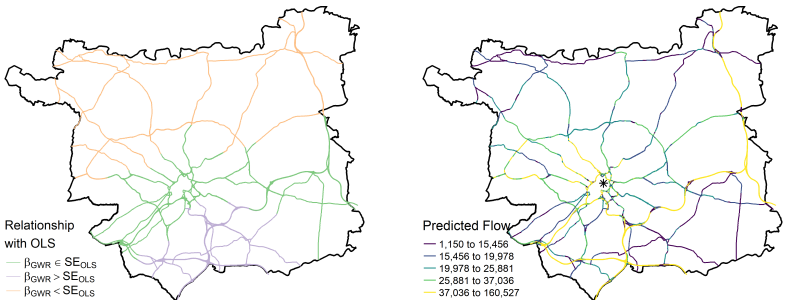
The regular counts, which are derived from fixed traffic cameras, were downloaded from the section Road Bulk Downloads of the platform Road Traffic Statistics developed by the UK Department for Transport (`https://roadtraffic.dft.gov.uk/downloads`). The 197 available count points locations were projected onto the road network, and, for each fixed camera, we assigned the corresponding traffic estimate to the overlapping road segment. The GWR was implemented assuming the actual frequencies observed at each count point as the response variable and the traffic flows measured by the mobile devices as predictor. Figure 1(a) displays the road network and the count points locations, which are distributed in several parts of the municipality.

As already mentioned, the smoothing parameter h in Equation (3) was estimated using leave-one-out cross-validation by minimising the mean squared error of traffic flows predictions. More precisely, we selected a series of bandwidth values in the range of the observed shortest path distances among all road cameras, and we calculated $\sum_{i=1}^{197}[y_i - \hat{y}_{\neq i}(h)]^2$ for all possible values of h. The quantity y_i denotes the observed road count, while $\hat{y}_{\neq i}(h)$ is the predicted flow obtained using bandwidth h and all observations but the ith one. The MSE estimates are reported in Figure 1(b) which suggests an optimal bandwidth as large as 5500 approximately.

Considering the smoothing parameter associated to the minimum MSE, we estimated the geographically weighted regression and predicted the traffic counts for

Table 1 Comparison between values detected by GPS sensors (first row), fixed cameras (second row) and predicted counts (third row) according to the model described in Equation (1).

	Minimum	1st Quartile	Mean	3rd Quartile	Maximum
GPS counts	0	865	1452	2272	12420
Fixed sensors counts	1150	12320	20999	38177	160527
Predictions	2820	16636	22965	32830	147694



(a) (b)

Fig. 2 Predicted daily traffic flows according to the model described in Equation (1) (a); Comparison with OLS estimates (b).

all segments in the network. As we can see from the equations reported above, the GWR is a local approach, which implies that each prediction requires the estimation of a different set of parameters. We explored this aspect more precisely by developing a comparison between the estimator detailed in Equation (2) and a classical OLS. The results are reported in Figure 2(a), where a segment is coloured in violet if the corresponding GWR estimate lays above the 95% confidence interval (CI) of the overall OLS estimate, in orange if the GWR estimate lays below the CI, and green otherwise. A clear spatial pattern has been found in the estimated coefficients. This points out that the relationship between point and GPS traffic measurements is not stationary in space and suggests the adoption of the local approach.

We report in Figure 2(b) a choropleth map of the estimated traffic counts at the road segment level. The figure clearly highlights several roads corresponding to a motorway (i.e. the yellow segments connecting the south area with the north/north-east) and the most important arterial thoroughfares reaching the city centre (i.e. the black star in the middle of the map). Roads in the north-west suburbs are found to be exposed to lower traffic flows as compared to the rest of the city.

As already mentioned, the TomTom figures underestimate the real flows, while the fixed cameras are too sparse to provide useful traffic estimates. The GWR approach, integrating the two data sources, combines their benefits. Table 1 details a convenient summary of the road counts employed in this paper, highlighting the merits of the GWR estimates. More precisely, the first and second rows summarise the GPS and camera data, respectively, and clearly point out that the mobile counts underestimate the real flows. The last row reports a summary of the predicted counts according to the GWR introduced in the previous paragraphs. We can observe that the predicted flows have a similar scale than real traffic data from fixed cameras while preserving a global coverage of the entire network. To conclude, we calculated the pseudo R^2 of the GWR finding values that ranged from 0.67 to 0.99 with a median value equal to 0.91. These quantities indicate a good performance for the estimated model.

4 Conclusions

This study demonstrates that GWR is a powerful tool to predict traffic flows at the road network level, combining data detected with fixed devices and GPS sensors. The classical approach was adjusted to take into account the particular spatial domain, substituting the Euclidean distances with the more appropriate shortest path distances. Our case study focused on daily traffic flows in the road network of Leeds from January to December 2019, and we showed that the suggested methodology allows a realistic estimation of traffic counts in all segments of a street network, combining the main benefits of the two data sources. In fact, the results detailed in Section 3 prove that the proposed model is suitable for the problem at hand.

We plan to extend the analysis presented in this work in several directions. First, the spatio-temporal dynamics of traffic flows could be explored, developing a traffic counts estimate for different hours of the day or different days of the week. Furthermore, we will enhance the geographically weighted regression including a few external covariates (e.g. road types, speed limit and road curvature) that could improve the model's fit.

References

1. Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G. and Davies, T.M.: Analysing point patterns on networks—A review. Spatial Statistics **42**, p. 100435 (2021)
2. Barthélemy, M.: Spatial networks. Physics reports **499(1-3)**, pp. 1-101 (2011)
3. Caceres, N., Romero, L.M., Benitez, F.G. and del Castillo, J.M.: Traffic flow estimation models using cellular phone data. IEEE Transactions on Intelligent Transportation Systems **13(3)**, pp. 1430-1441 (2012)
4. Fotheringham, A.S., Brunsdon, C. and Charlton, M.: Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons (2003)
5. Zadeh, A.S.M. and Rajabi, M.A.: Analyzing the effect of the street network configuration on the efficiency of an urban transportation system. Cities **31**, pp. 285-297 (2013)
6. Zeng, J., Qian, Y., Wang, B., Wang, T. and Wei, X.: The impact of traffic crashes on urban network traffic flow. Sustainability **11(14)**, p. 3956 (2019)