SCIENTIFIC REPORTS
natureresearch

OPEN

Estimating Traffic Disruption Patterns with Volunteered Geographic Information

Chico Q. Camargo[1], Jonathan Bright[1], Graham McNeill[1], Sridhar Raman[2] & Scott A. Hale[1,3*]

Accurate understanding and forecasting of traffic is a key contemporary problem for policymakers. Road networks are increasingly congested, yet traffic data is often expensive to obtain, making informed policy-making harder. This paper explores the extent to which traffic disruption can be estimated using features from the volunteered geographic information site OpenStreetMap (OSM). We use OSM features as predictors for linear regressions of counts of traffic disruptions and traffic volume at 6,500 points in the road network within 112 regions of Oxfordshire, UK. We show that more than half the variation in traffic volume and disruptions can be explained with OSM features alone, and use cross-validation and recursive feature elimination to evaluate the predictive power and importance of different land use categories. Finally, we show that using OSM's granular point of interest data allows for better predictions than the broader categories typically used in studies of transportation and land use.

Understanding and forecasting traffic is an important task for urban policymakers. Road networks are by far the most heavily used part of transport infrastructure (for example, 64% of all trips in the UK were made by car in 2016[1]); yet compared to other transportation modes (such as rail and air) basic data about traffic flow on roads is largely lacking.

In the last decade, a variety of novel data sources have started to offer the possibility of filling this gap, such as data from GPS transponders on mobile phones[2] or data from social media[3], which are generating considerable academic interest. Here, we contribute to this growing literature on the use of new data sources to understand traffic by using volunteered geographic information from OpenStreetMap (OSM) to understand what types of land use are associated with traffic jams, as well as increased traffic volume.

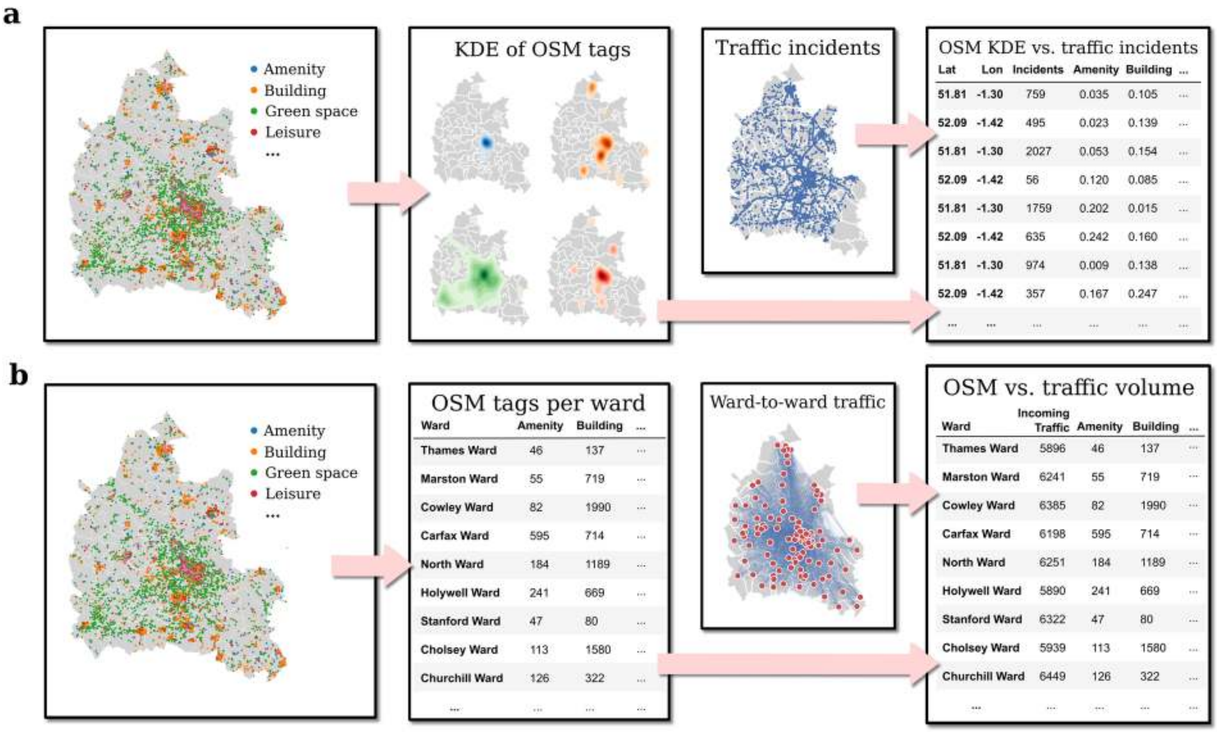
The connection between land use and transport is a classic subject in the literature, dating back to 19th century work by Ravenstein and Carey on human migration[4,5] to recent data-driven studies incorporating land use data into traffic volume prediction[6–10]. Despite the recent increase in the public availability of urban data, land use categories are typically classified at a highly aggregate level (e.g., defining areas as *residential*, *commercial*, or *industrial*) and data have typically been expensive to put together, meaning it is often only available for large cities such as London, New York or Paris, being less common in smaller or less dense locations[11–14]. It is this context that makes OSM a very useful tool, in that its data is highly granular, offering a classification of different types of commercial activity, public amenities and other forms of land use, but also in the fact that all this data is freely and openly available. The completeness and accuracy of OSM coverage has been assessed in previous studies[11,12,15–21], yielding positive but cautious results, particularly about road networks. It has also been used to successfully identify the types of trips which human mobility models struggle to predict accurately[10].

Results

We test the extent to which OSM data can offer a good estimation of the volume of overall traffic and the number of traffic disruptions, defined as any deviation from normal smooth traffic on a road network, by making use of a series of linear regression models. For the models of the traffic disruptions volume, observations are the geographic (latitude and longitude) points where traffic disruptions were observed in the network and the response variable is the number of traffic disruptions observed during the month of March 2017.

The data analysis pipelines for the two sets of linear models in this study are described in Fig. 1. As shown in the top panels (**a**), we first produce kernel density estimates (KDE) of every OSM category and meta-category. We then estimate the number of traffic disruptions at a given latitude and longitude using the KDEs of either the OSM

[1]Oxford Internet Institute, University of Oxford, Oxford, United Kingdom. [2]Oxford Brookes University, Oxford, United Kingdom. [3]Alan Turing Institute, London, United Kingdom. *email: scott.hale@oii.ox.ac.uk



OSM KDE vs. traffic incidents

Lat	Lon	Incidents	Amenity	Building	...
51.81	-1.30	759	0.035	0.105	...
52.09	-1.42	495	0.023	0.139	...
51.81	-1.30	2027	0.053	0.154	...
52.09	-1.42	56	0.120	0.085	...
51.81	-1.30	1759	0.202	0.015	...
52.09	-1.42	635	0.242	0.160	...
51.81	-1.30	974	0.009	0.138	...
52.09	-1.42	357	0.167	0.247	...
...	...	...	...	...	...

OSM tags per ward

Ward	Amenity	Building	...
Thames Ward	46	137	...
Marston Ward	55	719	...
Cowley Ward	82	1990	...
Carfax Ward	595	714	...
North Ward	184	1189	...
Holywell Ward	241	669	...
Stanford Ward	47	80	...
Cholsey Ward	113	1580	...
Churchill Ward	126	322	...
...	...	...	...

OSM vs. traffic volume

Ward	Incoming Traffic	Amenity	Building	...
Thames Ward	5896	46	137	...
Marston Ward	6241	55	719	...
Cowley Ward	6385	82	1990	...
Carfax Ward	6198	595	714	...
North Ward	6251	184	1189	...
Holywell Ward	5890	241	669	...
Stanford Ward	6322	47	80	...
Cholsey Ward	5939	113	1580	...
Churchill Ward	6449	126	322	...
...	...	...	...	...

Figure 1. Schematic pipeline of the linear model for the two sets of linear models in this study. As shown in the top panels (**a**), we first produce kernel density estimates (KDE) of every OpenStreetMap (OSM) category and meta-category, which we then compare with the number of traffic disruptions at a given latitude and longitude. The bottom panels (**b**) show we also aggregate the OSM data points into a total count per ward, which we then compare with the traffic volume going into every ward in Oxfordshire.

meta-categories or of the OSM categories at each point. To produce the KDEs, we made use of a Gaussian kernel searched over a range of bandwidth parameters before adopting a bandwidth of 0.001, which captures the range of spatial variation of all OSM points of interest. The specific value of the bandwidth parameter did not qualitatively affect our results. These KDEs allow us to estimate the density of any type of OSM feature at all of the points where traffic disruptions were reported.

As shown in the bottom panels (**b**), we also perform a second set of linear regressions where we aggregate the OSM data points into a total count for every one of the 112 electoral wards in the county of Oxfordshire, UK. We then estimate the volume of traffic going into every ward using either counts of the OSM meta-categories or all OSM categories for each ward.

Estimating traffic disruptions. The first linear model to estimate traffic disruptions only makes use of the meta-categories of OSM features (see Table 1a). These meta-categories represent traditional classifications of land use types. The model only weakly fits the traffic disruptions data, resulting in an adjusted R^2 of 0.11, which is a goodness-of-fit metric that takes into account the different number of independent variables and is a common metric for model comparison in computational social science[22–24]. Individual coefficients show that commercial areas are the ones most associated with high traffic, while industrial areas are the least so. We also tested different versions of the model only estimating distributions on weekdays and weekends, as the nature of traffic disruptions on these days could be different, but the overall fit to the log-transformed data was similar.

The second model has more granular land-use data by making use of all OSM categories that were observed at least a hundred times in Oxfordshire, resulting in KDEs for 40 different types of point (from pubs, schools and restaurants to graveyards, postboxes and gardens). This model fits the log-transformed data considerably better than the meta categorization model as captured by the adjusted R^2. This granular model results in an adjusted $R^2 = 0.55$. This large increase in adjusted R^2 is not simply the result of more input/independent variables; adjusted R^2 accounts for the number of independent variables and will decrease when variables are added that do not affect the dependent variables. The model coefficients of largest absolute value are represented in Table 1b, and their corresponding p-values are indicated as well. It is important to note that the OSM category *residential* is not equivalent to the meta-category *residential*, as the latter includes more OSM categories. We discuss this point in more detail in the next sections.

The second, granular model gives estimates of how things we might expect to explain local traffic jams vary with actual traffic disruptions. For example, one would expect places of worship and schools to both have a relatively high number of traffic disruptions, but the coefficients in this model indicate a large difference between the coefficient corresponding to the relationship between the number of points of interest tagged as schools ($c = 0.042$) and the log-transformed number of traffic disruptions and the corresponding coefficient for places of

Variable	Estimate
(**a**)**Meta-categories only**	
Residential	− 0.09**
Industrial	− 0.18**
Recreational	− 0.10*
Institutional	0.14*
Green space	0.26***
Commercial	0.32***
Observations	6529
Adjusted R^2	0.11
(**b**) **Granular model**	
Residential	0.61***
Farmland	0.56***
School	0.042**
Place of worship	0.009**
…	
Apartments	− 0.09**
Observations	6529
Adjusted R^2	0.55

Table 1. Granular land-use categories from OpenStreetMap allow for more detailed understandings of traffic disruptions. Compared with the traditional land-use categories shown in (**a**) that produce an adjusted $R^2 = 0.11$, the granular classifications used in (**b**) increase the adjusted R^2 to 0.55. Only a small subset of the 40 predictor variables are shown for (**b**), with all other coefficients shown in Table S3 in the Supplementary Information. Respectively, *, ** and *** indicate $p < 0.05$, $p < 0.01$ and $p < 0.001$.

worship ($c = 0.009$). The analysis, however, is only correlational: OSM points of interest tagged as farmland and parking have high positive coefficients, which suggests that the high number of traffic disruptions around such points might be due to traffic network features such as narrow roads rather than the effects of these OSM features directly.

Estimating traffic volume.

We also test the effectiveness of OSM data in estimating the traffic volume in Oxfordshire. For this variable, rather than using KDEs to estimate the density of each OSM feature at a given road, we aggregate the number of points of interest tagged with each meta-category and category, producing two sets of independent variables for every ward: one corresponding to the total number of points tagged with each one of the 6 OSM meta-categories, and one corresponding to the points in every ward in the 40 categories. We then produce two corresponding linear regression models using the log-transformed total traffic flowing into a ward as the dependent variable.

Not surprisingly, some OSM categories are also highly correlated, in the sense that they often appear in the same wards. Figure 2 lists all OSM categories, and shows these correlations in detail. It shows a heatmap displaying the Pearson correlation between the distribution of OSM categories over wards, giving higher values to pairs of OSM categories that often appear in the same wards (e.g, *forest* and *meadow*), and lower values to pairs of wards that rarely co-occur (e.g., *farmyard* and *fast food*). The figure also shows the result of performing hierarchical clustering on the OSM categories according to their correlation. There is a cluster formed by *farm*, *farmland*, *farmyard*, *forest*, *meadow*, *graveyard* and *reservoir*, which separates these rural categories from more urban categories as *university* or *retail*.

The linear regression models built with the traffic volume data show the same qualitative trend as the ones built with traffic disruption data. The first model, with the 6 meta-categories, results in an adjusted R^2 of 0.26. Its coefficients indicate that OSM points tagged as *commercial* are associated with heavier incoming traffic, while points tagged as *recreational* are negatively associated with it. Coefficients corresponding to meta-categories are presented in Table S1 in the Supplementary Information.

The finer-grained model, featuring the 40 OSM categories shown in Fig. 2, naturally shows a more nuanced scenario. Not only does it provide a better fit to the data, with an adjusted R^2 of 0.45, but it also provides more detail into the meta-categories used in the simpler linear models. Categories such as *telephone* and *university* show strong associations with higher levels of incoming traffic, whereas categories such as *forest*, *meadow* and *allotments* show weaker associations.

For both the incoming traffic volume per ward and the number of traffic disruptions, the jump from 6 meta-categories to 40 OSM categories implied a change from a linear model with a poor fit to a model with a better fit, indicated by the changes in their adjusted R^2. It is natural to then ask if all 40 OSM categories are necessary for the new model to work, or if an equally good fit could be obtained by selecting a different number of meta-categories, or a subset of those 40 OSM categories, excluding correlated categories. This is discussed in the next subsection.

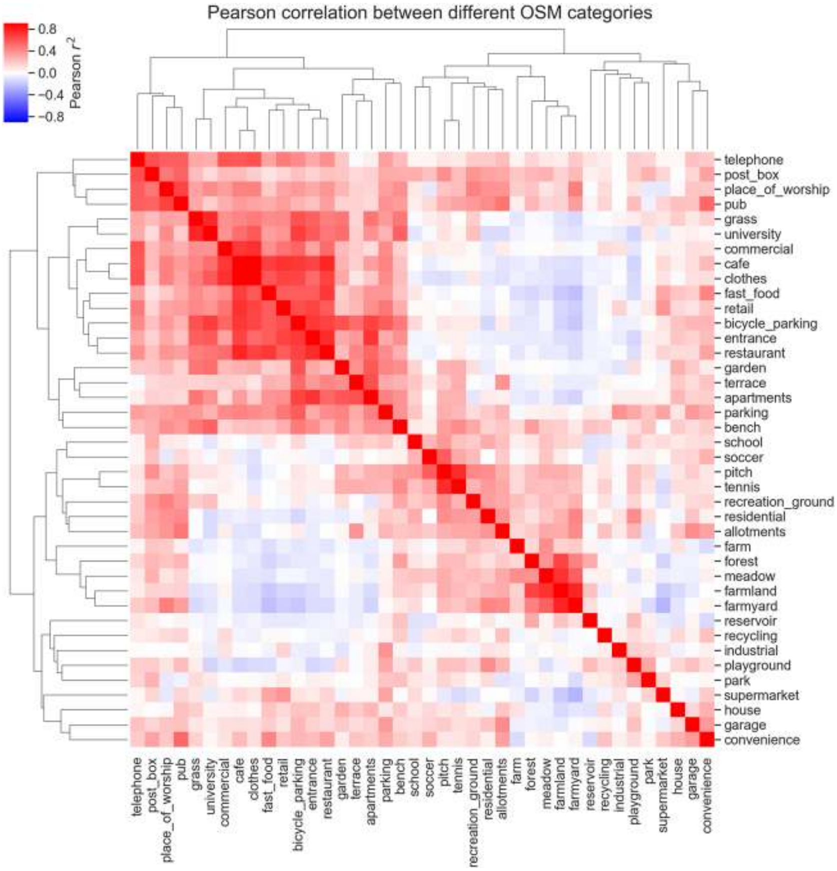


Figure 2. Clustermap showing the Pearson correlation of the distribution of different OSM categories over all Oxfordshire wards. The heatmap shows the correlation between the number of points of interest tagged as every OSM category in this study. The trees show how OSM categories cluster according to their correlation. For example, OSM categories such as *farm*, *farmland*, *farmyard* form a cluster, indicating that they often appear in the same wards, while not being as correlated to categories such as *cafe* and *fast food*.

Feature selection. We address the explanatory power of each variable in these linear models using feature ranking with recursive feature elimination, aided by cross-validated selection of the best number of features, as implemented in the *scikit-learn* Python library[25]. For both dependent variables, i.e., the incoming traffic volume per ward and the volume of traffic disruptions on a point in the road network, we perform 1000 rounds of k-fold cross-validation with $k = 10$, scoring models for their adjusted R^2. For every cross-validation round, the 6 or 40 independent variables are then ranked according to their importance, which in this case is the magnitude of their corresponding coefficients in the linear models. Selected features are assigned rank 1, with the next-best variable being assigned rank 2, and so on until the last variable.

As multiple cross-validation rounds might result in different rankings of their predictor variables, we combine all rankings by calculating the stability of every variable, as well as its mean rank. Stability selection[26] is a method which provides a useful balance between feature selection and data interpretation, by evaluating how often a given feature is included among the most important (i.e., rank 1) for a model. Strong or important features should achieve scores close to 1, indicating that most of the 1000 cross-validation rounds ranked them as one of the best features for prediction. Any weaker but still relevant features should still have non-zero scores, as they ought to be selected as best features at least occasionally. Finally, irrelevant features should return near-zero scores, indicating that they are very unlikely to feature among the selected variables.

For the volume of traffic disruptions, both the mean rank and the stability analysis reveal the same pattern, as shown in Tables 2 and 3. As there is no specific threshold separating stable from unstable features, these tables show the all six meta-categories in Table 2 and the 10 variables with the lowest rankings and highest stability scores in Table 3. The meta-category *residential* features at the top, with both mean rank and stability equal to 1, indicating a variable that featured as important in all of the 1000 cross-validation rounds. It is then followed by the meta-category of *recreational*, which still features as important, with all other meta-categories featuring with a lower rank, and a stability less than 0.6. The corresponding granular OSM categories show the categories *farmland*, *residential*, *parking*, *forest*, and *farmyard* at the top, with mean rank and stability of 1.000, indicating that they were considered important variables in all 1000 cross-validation rounds. These categories are followed by *farm*, *meadow*, and *industrial*, with stability of 0.999 and respective mean ranks of 1.001, 1.002 and 1.003.

	ranking	stability
(a) **Meta-categories only, traffic disruptions**		
residential	1.000	1.000
recreational	1.311	0.689
commercial	1.758	0.553
industrial	2.216	0.542
green space	2.794	0.422
institutional	3.379	0.415
(b) **Meta-categories only, traffic volume**		
commercial	1.000	1.000
recreational	1.734	0.266
institutional	2.676	0.058
residential	3.636	0.040
green space	4.606	0.030
industrial	5.602	0.004

Table 2. Average ranking and stability of different meta-categories in predicting the number of traffic disruptions and the incoming volume for every Oxfordshire ward.

	ranking	stability
(a) **Granular model, traffic disruptions**		
farmland	1.000	1.000
residential	1.000	1.000
parking	1.000	1.000
forest	1.000	1.000
farmyard	1.000	1.000
farm	1.001	0.999
meadow	1.002	0.999
industrial	1.003	0.999
reservoir	1.010	0.993
soccer	1.020	0.990
(b) **Granular model, traffic volume**		
fast-food	1.000	1.000
post box	1.028	0.972
cafe	1.080	0.948
bench	1.211	0.869
soccer	1.409	0.802
commercial	1.648	0.761
telephone	1.916	0.732
parking	2.200	0.716
convenience	2.508	0.692
farm	2.855	0.653

Table 3. Average ranking and stability of different OSM categories in predicting the number of traffic disruptions and the incoming volume for every Oxfordshire ward. Only the top 10 variables according to ranking are shown.

Tables 2 and 3 also show the mean rank and stability results for the total incoming traffic volume. Reported results are for trips on weekday mornings, but qualitatively similar results are obtained when using the full collection of trips in the dataset as shown in Table S2 in the Supplementary Information. The meta-category *commercial* features at the top, with both mean rank and stability equal to 1, indicating a variable that featured as important in all of the 1000 cross-validation rounds. It is then followed by the meta-category of *recreational*, which still features as important, with all other meta-categories featuring with a lower rank, and a stability less than or equal to 10%. The corresponding granular OSM categories show *fast-food* at the top, with a mean rank and stability of 1. The categories *post box* and *cafe* feature next. OSM categories such as *farm* and *farmyard* feature with lower mean ranks, and stability under 0.7. One must bear in mind that the OSM categories *residential* and *commercial* are not equivalent to the meta-categories *residential* and *commercial*. This point is discussed in more detail in the next section.

Discussion

The analysis presented in this paper shows how fine-grained land use categories can be used to estimate traffic volume and traffic disruption patterns. In particular, we have shown that the fine-grained features available on OpenStreetMap can greatly increase the explanatory power of linear models. Of course, some variation still remains unexplained, and it is likely that more dynamic features (such as weather patterns or working week fluctuations) would need to be taken into account to account for this. Besides, since OpenStreetMap is continuously updated, this work can only provide a cross-sectional snapshot of the data. Nevertheless we have shown that static features can offer important explanatory power, and this too is useful to provide a perspective on how things might change as the features themselves change. We have also shown the importance of different land use categories by using recursive feature elimination, and have used cross-validation to examine the predictive power of different models.

One useful application of these data and methods is to offer estimated answers to questions such as "what impact will placing another cafe at a given point have on traffic jams at that location?" For example, according to our fine-grained traffic models, the impact of a new school on the number of traffic disruptions in its area should be comparable to the impact of a new retail store or fast food restaurant. The linear model coefficients associated with the presence of these amenities are all approximately $c_i = 0.05$, meaning that an increase by 1 in these variables (number of schools, retail stores, and restaurants) implies an increase of 5% in the log-transformed number of traffic disruptions, i.e., an increase in 12% in the monthly number of traffic disruptions at the location. These same categories—*school*, *retail*, and *fast food*—also have a positive correlation with the monthly volume of traffic going into a ward, even if with different coefficients. Respectively, the three categories have coefficients of 0.0010, 0.0021, and 0.0028, implying respective increases in 0.2%, 0.5%, and 0.7% in the total (non-log transformed) traffic flowing into areas.

It is important to remember the limitations of OpenStreetMap land use categories. For example, the OSM categories *residential* and *commercial* are not equivalent to the meta-categories *residential* and *commercial*, and the OSM dataset includes tags such as *farmland* and *farmyard* along with *farm*, which was deprecated and substituted by the two other farm categories in 2017[27]. Categories and meta-categories might differ in the quality of the annotation, and in how informative they are to the traffic predictions. The cross-validation and recursive feature elimination performed here are first steps in tackling this issue. The rank and stability analysis provide additional evidence that higher numbers of traffic disruptions are observed in residential and rural areas, indicated by meta-categories such as *residential* and OSM categories such as *farmland*, *forest* and *farmyard*. This result matches the distribution of OSM categories over all wards, as indicated in Fig. 2, which shows that OSM tags such as *house*, *farmland*, *residential*, and *farmyard* are often seen in the same wards, while rarely co-occurring with OSM categories such as *commercial* or *cafe*. The latter two OSM categories do not feature as important predictors for the number of traffic disruptions, but they do feature as important predictors for traffic volume, where they show the highest rank and stability, which is also observed for the meta-category *commercial*.

Our study also suggests promising avenues for future research. One of these would be to take advantage of the constantly evolving nature of OpenStreetMap to track the emergence of new physical features, and relate these to changes in traffic conditions, thus extending the correlations we have highlighted in this paper into a causal setting. Another would be to combine these with other sources of observational data, such as licensing applications, planning permission, and building regulations, to see if these can build on the baseline model we have constructed. Finally, it would be worthwhile extending our study to other countries and contexts. One limitation of our study is that it focuses solely on one administrative region in the UK: it would be worthwhile to explore if the value of OSM's granular point of interest data is generalizable. A larger dataset covering multiple locations and with fully held out test data would also allow the exploration of non-linear models with less danger of greatly overfitting the data. As our ability to understand and explain traffic patterns improves so will the ability of policymakers to effectively design urban transport systems that serve the needs of their citizens.

Materials and Methods

Our geographical focus is the English county of Oxfordshire, a geographical area of just over 2, 605 km^2 and which contains around 680, 000 inhabitants. For our OpenStreetMap (OSM) data, we downloaded *points of interest* from the OSM database which provide indications of the way land is used. Points of interest were downloaded in November 2017. One of the authors then assigned each point of interest to six meta-categories of land use: *residential*, *industrial*, *commercial*, *recreational*, *institutional* and *green space* (our assignment of each category is available as supplemental data to this paper). These categories are standard across the transport and land-use literature (see, for example, the typologies present in[6,13,28]). We also preserved the more granular categorization which is already provided by OSM (and hence requires no manual annotation). For example, our meta-category of *commercial* contains categories such as *restaurant*, *pub* and *cafe*. Our classification of OSM categories into meta-categories is availabl from Zenodo as indicated in the Data Availability section. We chose to ignore OSM categories and meta-categories with less than a hundred points of interest in Oxfordshire, as well as categories indicating the location of the transport network itself, as these are obviously coterminous with our traffic disruption data.

We obtained the traffic disruption data from traffic disruption reports shared with us by the Oxfordshire County Council, which are sourced from a major traffic analytics company. These reports correspond to over 1.4 million traffic incidents from just over 6,500 points on the Oxfordshire traffic network (each point being approximately a 10 m × 10 m square). The number of traffic disruptions counts at each point ranged from 1 to 64,313, and with an average of 219 traffic disruption counts per point, a standard deviation of 1382 counts, and a median of 21 traffic disruption counts per point. It is important to note that many traffic disruptions such as the ones studied in this paper do not result in casualties or police reports, meaning that data on car accidents only reflects a fraction of the incident estimates presented here.

For the traffic volume data, we used anonymised and aggregated GPS mobile phone data provided by a major smartphone operating system. Similar data sets have been validated and successfully used in urban mobility studies in San Francisco[29] and Amsterdam[30]. The data set contains estimated trip volumes for origin-destination pairs of wards in Oxfordshire between January and February 2017 in hourly increments. We took a subset of the data, only using trips inferred by the company to be made by vehicle (and not walking or cycling), and trips on weekdays made between 7 am and 12 pm (noon), which we aggregated into a total traffic going into every Oxfordshire ward over the two-month period. Using the whole day and/or including weekend trips yielded qualitatively similar results. Finally, we obtained shapefiles for the border of all Oxfordshire wards from the Digimap mapping data service[31]. Datasets were manipulated using dataframes from the Python Pandas library[32].

Data availability

Data are available from Zenodo at https://zenodo.org/record/3383443.

Received: 1 August 2019; Accepted: 24 December 2019;
Published online: 27 January 2020

References

1. Department for Transport. Transport Statistics Great Britain https://bit.ly/2tsCsvq (2016).
2. Vlahogianni, E. I., Karlaftis, M. G. & Golias, J. C. Short-term traffic forecasting: Where we are and where weare going. *Transp. Res. Part C: Emerg. Technol.* **43**, 3–19 (2014).
3. McNeill, G., Bright, J. & Hale, S. A. Estimating local commuting patterns from geolocated twitter data. *EPJ Data Science* **6**, 24 (2017).
4. Ravenstein, E. G. The laws of migration. *J. statistical society Lond.* **48**, 167–235 (1885).
5. Carey, H. C. *Principles of social science* (JB Lippincott & Company, 1867).
6. Wegener, M. & Fürst, F. Land-use transport interaction: State of the art. https://doi.org/10.2139/ssrn.1434678 (2004).
7. Lenormand, M. *et al.* Comparing and modelling land use organization in cities. *Royal Soc. Open Sci.* **2**, 150449 (2015).
8. Louail, T. *et al.* Uncovering the spatial structure of mobility networks. *Nat. Commun.* **6**, https://doi.org/10.1038/ncomms7007 (2015).
9. Lee, M. & Holme, P. Relating land use and human intra-city mobility. *PloS one* **10**, e0140152 (2015).
10. Camargo, C. Q., Bright, J. & Hale, S. A. Diagnosing the performance of human mobility models at small spatial scales using volunteered geographic information. *arXiv preprint arXiv:1905.07964* (2019).
11. Zielstra, D. & Zipf, A. A comparative study of proprietary geodata and volunteered geographic information for germany. In *13th AGILE international conference on geographic information science*, vol. 2010 (2010).
12. Haklay, M. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environ. planning B: Plan. design* **37**, 682–703 (2010).
13. Liu, Y., Wang, F., Xiao, Y. & Gao, S. Urban land uses and traffic 'source-sink areas': Evidence from gps-enabled taxi data in shanghai. *Landsc. Urban Plan.* **106**, 73–87 (2012).
14. Thebault-Spieker, J. Hecht, B. & Terveen, L. Geographic biases are 'born, not made': Exploring contributors' spatiotemporal behavior in openstreetmap. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, 71–82 (ACM, 2018).
15. Girres, J.-F. & Touya, G. Quality assessment of the french openstreetmap dataset. *Transactions in GIS* **14**, 435–459 (2010).
16. Helbich, M., Amelunxen, C., Neis, P. & Zipf, A. Comparative spatial analysis of positional accuracy of openstreetmap and proprietary geodata. *Proc. GI-Forum*, 24–33 (2012).
17. Mashhadi, A., Quattrone, G. & Capra, L. The impact of society on volunteered geographic information: The case of openstreetmap. In *OpenStreetMap in GIScience*, 125–141 (Springer, 2015).
18. Arsanjani, J. J.Mooney, P.Zipf, A. & Schauss, A. Quality assessment of the contributed land use information from openstreetmap versus authoritative datasets. In *OpenStreetMap in GIScience*, 37–58 (Springer, 2015).
19. Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C. & Haklay, M. A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.* **31**, 139–167 (2017).
20. Bright, J., De Sabbata, S. & Lee, S. Geodemographic biases in crowdsourced knowledge websites: Do neighbours fill in the blanks? *GeoJournal* **83**, 427–440 (2018).
21. Bright, J., De Sabbata, S., Lee, S., Ganesh, B. & Humphreys, D. K. Openstreetmap data for alcohol research: Reliability assessment and quality indicators. *Heal. & Place* **50**, 130–136 (2018).
22. Choi, H. & Varian, H. Predicting the present with google trends. *Econ. Rec.* **88**, 2–9 (2012).
23. Wu, L. & Brynjolfsson, E. The future of prediction: How google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy*, 89–118 (University of Chicago Press, 2015).
24. Lin, A. Y., Cranshaw, J. & Counts, S. Forecasting us domestic migration using internet search queries. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May*, 13–17 (2019).
25. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
26. Meinshausen, N. & Bühlmann, P. Stability selection. *J. Royal Stat. Soc. Ser. B (Statistical Methodology)* **72**, 417–473 (2010).
27. OpenStreetMap contributors. Openstreetmap mapnik and cartocss update. https://github.com/gravitystorm/openstreetmap-carto/blob/master/changelog.md (2017).
28. Srinivasan, S., Provost, R. & Steiner, R. Modeling the land-use correlates of vehicle-trip lengths for assessing the transportation impacts of land developments. *J. Transp. Land Use* (2013).
29. Sana, B., Castiglione, J., Cooper, D. & Tischler, D. Using Google's Aggregated and Anonymized Trip Data to Support Freeway Corridor Management Planning in San Francisco, California. *Transp. Res. Rec. J. Transp. Res. Board* **2643**, 65–73, https://doi.org/10.3141/2643-08 (2017).
30. Knoop, V. L., van Erp, P. B. C., Leclercq, L. & Hoogendoorn, S.P. Empirical MFDs using Google traffic data. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3832–3839 https://doi.org/10.1109/ITSC.2018.8570005 (2018).
31. EDINA Digimap Ordnance Survey Service. OS MasterMap Topography Layer [Shape geospatial data], Scale 1, Tile: Oxfordshire, Ordnance Survey, Using: EDINA Digimap Ordnance Survey Service. https://digimap.edina.ac.uk/ (2018).
32. McKinney, W. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, vol. 445, 51–56 (Austin, TX, 2010).

Acknowledgements

The authors thank Llewelyn Morgan for facilitating access to data and supporting the project. A previous version of this paper was presented at CARMA 2018: 2nd International Conference on Advanced Research Methods and Analytics. The authors would like to thank those involved in the conference for their feedback. This project was supported by funding from InnovateUK (grant 52277-393176), NERC (grant NE/N00728X/1), the Lloyd's Register Foundation and The Alan Turing Institute Data–Centric Engineering Programme (EPSRC grant EP/N510129/1).

Author contributions

S.A.H. and J.B. secured the funding and coordinated the project. J.B. and G.M. designed the study. C.Q.C., J.B. and S.R. collected the data. C.Q.C. and J.B. carried out the analysis, and C.Q.C. performed the cross-validation. J.B. wrote the first draft and C.Q.C., G.M. and S.A.H. edited it. All authors gave final approval for publication.

Competing interests

The authors declare that they have no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-57882-2.

Correspondence and requests for materials should be addressed to S.A.H.

Reprints and permissions information is available at www.nature.com/reprints.