Environmental Research 240 (2024) 117457

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Mobile monitoring of air pollutants; performance evaluation of a mixed-model land use regression framework in relation to the number of drive days.

Jules Kerckhoffs [a,*], Gerard Hoek [a], Roel Vermeulen [a,b]

[a] Institute for Risk Assessment Sciences, Utrecht University, Utrecht, the Netherlands
[b] Julius Centre for Health Sciences and Primary Care, University Medical Centre, University of Utrecht, the Netherlands

A B S T R A C T

We used black carbon data from a mobile monitoring campaign in Oakland, USA measuring street segments up to 40 times and compared a data-only, LUR model and mixed-model approach with a long-term average, represented by the average concentration based on 40 drive days on that street segment. The mixed model outperformed the data-only and LUR model estimates, with 80% explained variance after 5 drive days and 90% after 14 drive days. The data-only approach needed 8 and 15 to achieve an explained variance of 80% and 90%, respectively, The LUR model never achieved an explained variance higher than 70%. The mixed model is a scalable approach, as it can be used before all street segments in a domain are measured by developing a LUR model and adds information with increasing repeats per street segment.

1. Introduction

The goal of most air pollution exposure studies for epidemiological research is to create spatial maps with exposure predictions at the finest spatial scale possible. Over the past decade, mobile monitoring has shown that it can offer this high granularity. A disadvantage of using mobile monitoring data directly for exposure assessment, however, is that it needs a substantial number of repeated air pollution measurements per segment to obtain stable estimates, which is often not a realistic approach. Frequent repeats are necessary because of the inherent uncertainty of mobile measurements due to the short-term nature of the measurements (often seconds per street segment). That is why most studies to date have used empirical modelling to stabilize exposure estimates. Several studies already showed that robust LUR models can be developed with only limited amount of data, both in coverage of the spatial domain (not all streets need to be driven) and the number of repeats per street segment (Kerckhoffs et al., 2017; Messier et al., 2018a; Hatzopoulou et al., 2017). However, by using empirical models that transfer knowledge between similar settings, hyperlocal information (e.g., unknown sources) may be lost.

In previous papers, we demonstrated that a mixed-model approach can achieve robust spatially-explicit concentration estimates via a land use regression (LUR) model (fixed-effects), while allowing street segments to deviate from the LUR prediction based on between-segment variation of the measurements as a random effect (Kerckhoffs et al., 2022a, 2022b). All street segments were used to develop a LUR model, and all individual measurements can alter the prediction of the fixed-effect part of a particular street segment based on the measurements of that street segment (random effect). This could be the case for streets where the LUR model is unable to predict accurately, due to local sources, particular traffic situations or missing street configuration variables. When we compared the model with external NO_2 measurements in Amsterdam, the mixed model prediction based on mobile measurements correlated 0.85 with the external long-term measurements, compared to a correlation of 0.74 and 0.75 for the data-only approach and LUR model, respectively.

However, in our previous work, we only had limited street segments with many drive days available to study the scalability of the mixed model approach and the trade-off between the fixed and random effects components with increasing number of drive days. In another study by Messier et al. (2018b), described in more detail below, 20.000 street segments were measured over 40 times. It was found that only 4 to 8 drive days per street segment are needed for a data-only model to outperform a LUR model, both tested on the average concentrations of those 40 repeats. Here, we use the same dataset to study the influence of drive-days on the performance of the mixed model approach and compare these to the performance of a LUR or data-only model. To do this, we select subsets of drive days per street segment and compare the

* Corresponding author.
E-mail address: j.kerckhoffs@uu.nl (J. Kerckhoffs).

https://doi.org/10.1016/j.envres.2023.117457
Received 21 July 2023; Received in revised form 29 September 2023; Accepted 18 October 2023
Available online 19 October 2023
0013-9351/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

average of the mobile measurements (measurement data only), the LUR prediction (fixed model only) and the mixed-model prediction (fixed plus random effects model) with the average concentration based on 40 drive days on that street segment.

2. Methods

2.1. Data

We used black carbon (BC) data from the Air View mobile monitoring campaign in Oakland (Messier et al., 2018a; Apte et al., 2017), in which two Google Street View cars collected air pollution measurements for over 1 year in Oakland, CA. About 21.000 unique street segments were sampled with lab-grade instruments (Teledyne API Model AE-633), with 1 s resolution. All 1-Hz measurements were assigned to the nearest 30-m street segment and averaged per drive day. Different from the data processing in Apte et al. (2017), but similar to the data processing in our previous work, we computed a mean value per street segment (Mean of Means) because this better reflects actual average concentrations, opposed to the median value (Median of Means) used in Apte et al. (2017) Mean concentrations were on average 1.3 times higher as median values and correlated well to the median ($r = 0.88$). More details of the full measurement campaign can be found in Apte et al. (2017).

Regarding the geographic variables used in the LUR development, we used the same set that was used in Messier et al. (2018a). These variables included street classifications, binary local truck routes, local zoning classifications, normalized difference vegetative index, percent landcover, road length, population density in buffers of 50–2500 m, and continuous point source variables such as National Priority Listing sites, airports, and ports.

2.2. Model development

We compare three different methods to predict long-term average concentrations of BC on every street segment, (i) Data-Only, (ii) LUR model and (iii) Mixed models, with different number of driving days. Like the study by Messier et al. (2018a), we assume that the average of 40 repeated short-term measurements on different driving days during a year represents a robust long-term average concentration of a street segment. Then, for each approach, we select a subsample of driving days (bootstrapped 100 times per number of drive days) and compute the average value for the data-only approach and develop a model for the LUR model and mixed model approach. These subsamples are then compared to the full set of 40 driving days per street segment.

All LUR models are developed with a stepwise forward linear regression model, in which the variable explaining the most variation in BC measurements enters the model first. The model continues to add variables adding the most explained variance (R^2) step by step until no variable can increase model performance by at least 1%. Mixed models use the same variables as the LUR as the fixed-effect part of the model and use the street segments as the random effect. This means that a mixed model with only one driving day cannot be computed and is the same as the LUR model. More details on the mixed model approach can be found in Kerckhoffs et al. (2022a).

2.3. Performance evaluation

We evaluate two different strategies to subsample the data. One randomly selects street segments from the entire study period, while the other strategy selects driving days in a chronological manner (e.g., twenty consecutive days). This means that the first comparison is a generalisable approach to indicate how many days of driving are needed to achieve a stable estimate of an average concentration per street segment. The second strategy is a more practical scenario but is more difficult to generalize as it depends on the size of the spatial domain sampled.

For the first comparison, we select 40 drive days for street segments with more than 40 drive days. This means that a 'subset' of 40 drive days yields a perfect score. Next, we randomly select drive days from the full dataset, starting with one drive day and continue adding drive days in steps of one up to ten drive days and subsequently in steps of five drive days to the maximum of 40 drive days. We repeated this sampling scheme 100 times and recorded the R^2 and RMSE for each subset. In every repeat a new subset of 40 drive days was selected.

For the second approach, we used a more realistic strategy, because it is impossible to randomly drive on street segments throughout the city. Here, every subset was based on a random start date and drive days were added in chronologically. This means that after 10 driving days, some street segments are measured ten times, while other street segments might have been measured only once. Like the other analysis, we first selected street segments with at least 40 drives days. To allow for start dates towards the end of the campaign and still allow a dataset of 150 consecutive drive days for example, driving days from the start of the campaign were added after the last drive day of the original campaign as if the whole driving campaign was 360 days. Subsets started with ten drive days and were subsequently increased with 10 drive days until 180 drive days. Also, this sampling scheme was repeated 100 times.

3. Results and discussion

3.1. Data-only mapping

Predicting long-term air pollution concentrations only based on mobile measurements can be very labour intensive, especially on a large scale. In our analyses, the data-only approach predicts less than 30% of variation in the long-term average concentration if the measurement consists of one drive-day per street segment. We found that at least 15 drive days are needed per street segment to generate a robust ($R^2 > 0.9$) long-term average BC concentration (Fig. 1), like what was found for BC measurements in the study by Messier et al. (2018a). For a R^2 of 0.8, eight drive days were needed. However, when compared to a LUR model, a data-only approach outperforms a LUR model at about 5 drive days on average. Regarding RMSE, a data-only approach always outperforms a LUR model (Fig. 1).

A similar pattern is shown in Fig. 2. We found that about 80 consecutive driving days (with an average of 4 h per day on street segments with at least 40 drive days)) are required to establish a robust long-term estimate on each street segment ($R^2 > 0.9$). To explain 80% of the variance, an average of 35 (CI: 20–40) consecutive driving days are needed. We note that this analysis is more difficult to generalize, as the number of days needed for a robust model depends on the size of the domain that needs to be sampled. In this case, an area of about 30 km^2 and about 21.000 street segments. Since not all street segments are covered in the first driving days, we restricted our analyses to the street segments covered within those days. Alternatively, missing street segments can be imputed with the average concentration of all street segments, which is shown by the dotted line in Fig. 2.

3.2. LUR models

Regarding R^2 values, LUR model performance is hardly influenced by the number of drive days per street segment. With one drive-day on every street, the LUR model explained 63% of the variation, with limited improvement in performance with increasing number of drive-days (70% for 40 drive-days). It is important to note that the test set (average of 40 drive day measurements) in this analysis is kept the same for all correlations.

Training model R^2 values were a lot lower with less drive days, as shown in figure S1. Mobile measurements consist of a few seconds per street segment and are therefore very variable, making the mobile measurements itself difficult to predict with a LUR model. However, for its use in epidemiological studies, we are often more interested in

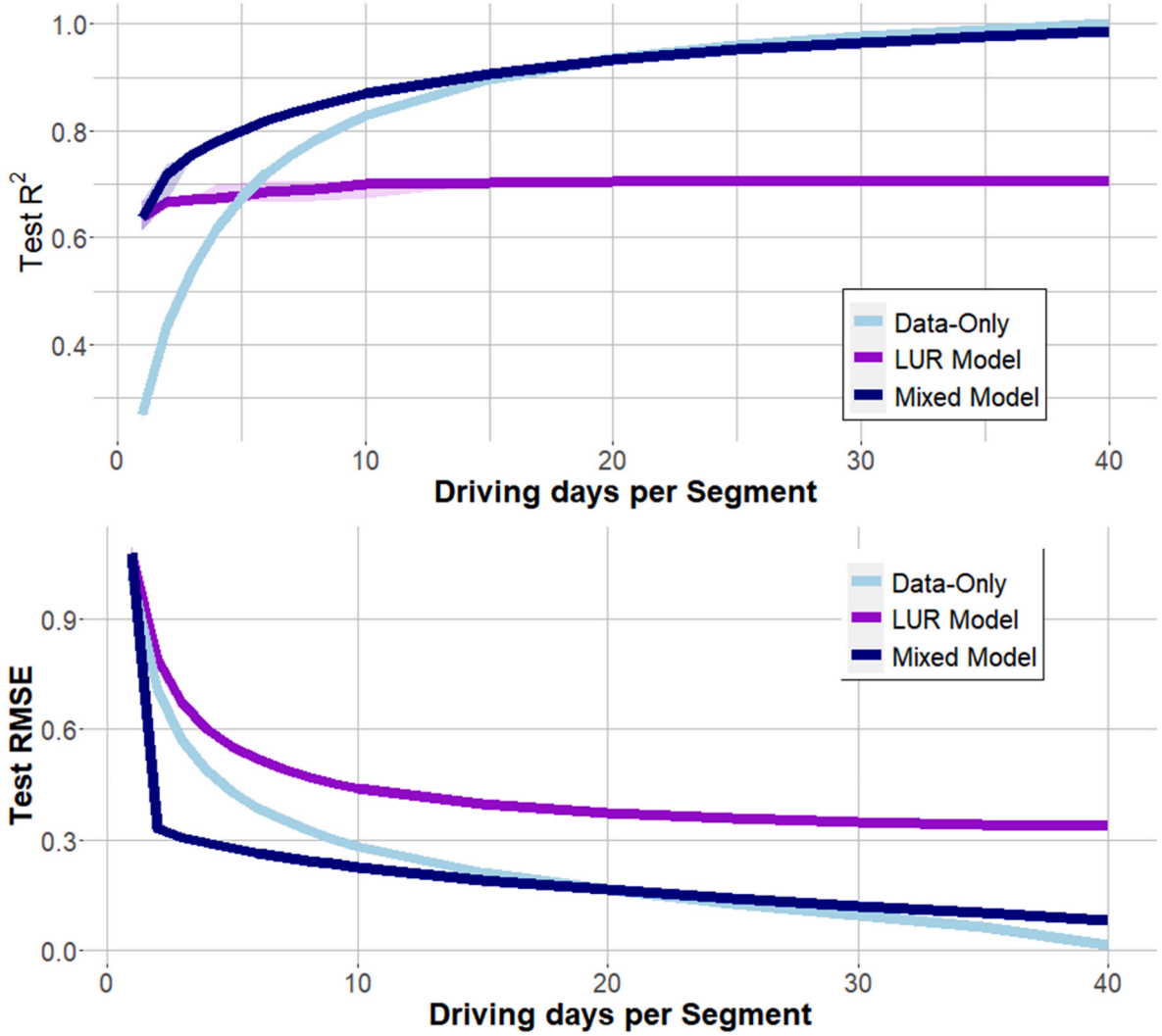

Fig. 1. Model performance for the explained variance (top) and RMSE (bottom) when the average BC concentration based on a subset of drive days per street segment is compared to an average BC concentration based on 40 drive days per street segment.

predicting long-term average concentrations than short-term mobile measurements. Increasing the accuracy of individual measurements, for example by having more repeats, increases the training model R^2 of LUR models (figure S1). This was also found by Hatzopoulou et al. (2017). They compared LUR models based on 3 repeats per street segment with LUR models based on 16 repeats and found that a model based on 16 repeats achieved a better adjusted model R^2. However, external model performance depends more on the accuracy of the test data than the accuracy of the training data. Multiple studies have shown that it is possible to predict robust long-term concentrations based on mobile monitoring data (Kerckhoffs et al., 2017; J et al., 2019; Shairsingh et al., 2018; Hasenfratz et al., 2015; Sabaliauskas et al., 2015; Weichenthal et al., 2016). Models with low training R^2 can still achieve high R^2 values when tested on higher-quality data. Figure S2 shows that BC predicted concentrations based on a model developed with one drive day per street segment correlates very well (average $r = 0.95$) with predicted concentrations based on the full model with 40 drive days per street segment.

Several studies found that is not even necessary to measure all street segments for a stable LUR model (Kerckhoffs et al., 2017; Messier et al., 2018a; Hatzopoulou et al., 2017). The total number of street segments in certain domains can be reduced significantly if the coverage and variation in road network and other predictors is preserved. Hatzopoulou et al. (2017) decreased the number of road segments from 611 to only 100 road segments in steps of 50 and R^2 values remained stable up until 200 road segments. Even LUR models based on 100 segments predicted on average 73% of the variation (opposed to 74% for the full dataset), albeit with a wider confidence interval (55–85% opposed to 70–78% for the full dataset).

Regarding the RMSE, the LUR model only stabilizes after 10 drive days per street segment. The LUR model is able explain the variation between street segments with high and low concentrations very well but seems to miss some nuance in absolute values within those categories in the first 10 drive days.

For the analysis with consecutive driving days, we show that the LUR model becomes stable after about 35 consecutive driving days (Fig. 2 and Figure S3). A LUR model based on 35 consecutive days of driving results in a similar LUR model based on the full measurement campaign. Even a LUR model based on only 10 days of driving correlates on average very well (average $r = 0.90$) to the LUR model based on the same street segments with measurements from the full campaign, though that depends if all types of street segments and all types of land use in the spatial domain are covered within those 10 driving days (Figure S4).

A combination of the amount of drive days per street segment and percent coverage of the domain was analysed with similar data as this paper by Messier et al. (2018a) Model performance was only slightly lower with 1 drive day per street segment and with only 10% of the

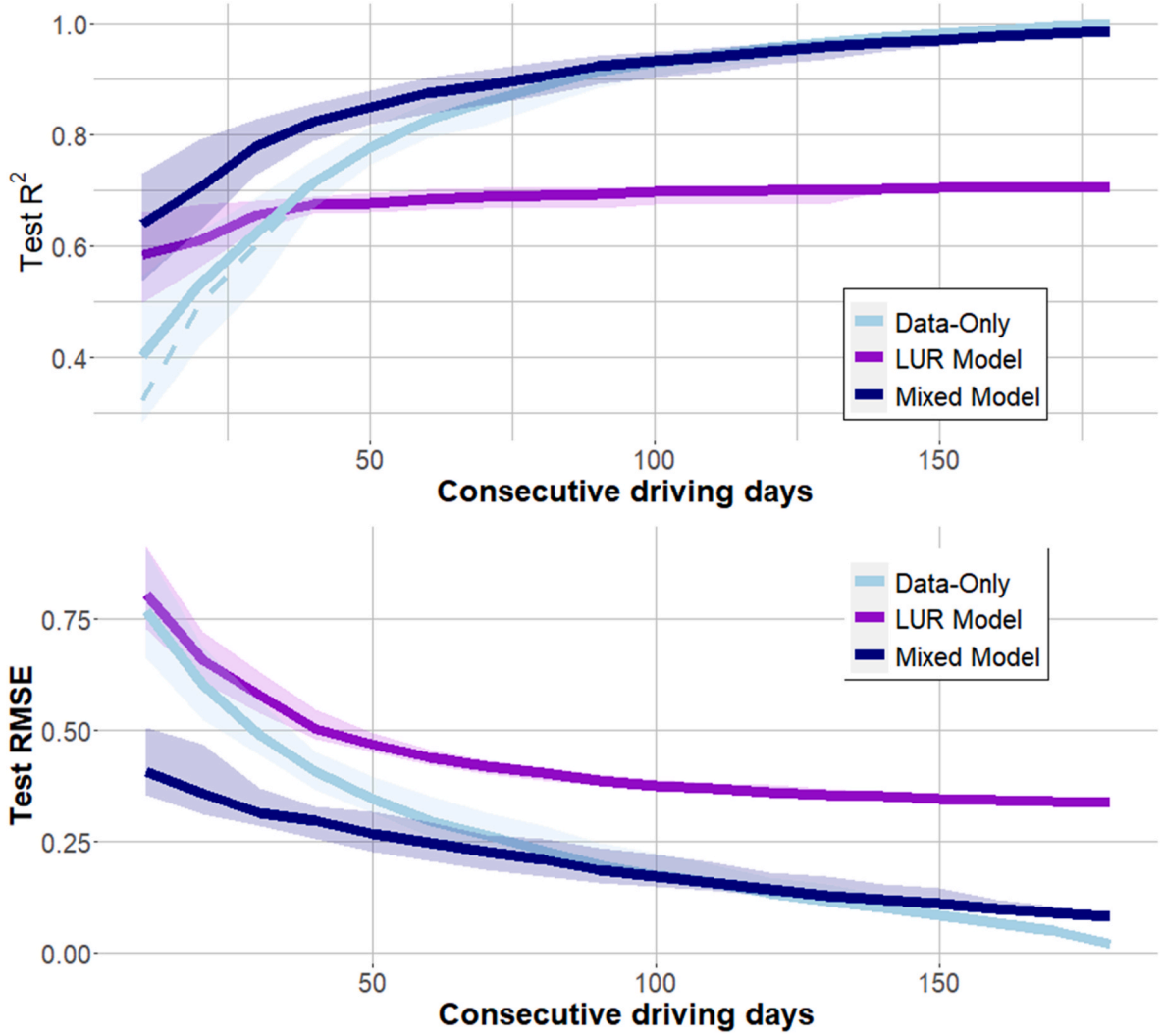

Fig. 2. Model performance for the explained variance (top) and RMSE (bottom) when the average BC concentration based on a certain number of consecutive driving days per street segment are compared to the full average BC concentration based on 180 consecutive drive days.

streets covered than a LUR model with all drive days and 100% coverage.

3.3. Mixed models

The mixed model outperformed the data-only and LUR model estimates, with 80% explained variance after five drive-days and 90% after 14 drive-days (Fig. 1). This is because the mixed model can start from a LUR model with one drive day per street segment and can already add extra information on every street segment with two drive days per street segment. With about 15 repeats per street segment, the mixed model performance is nearly identical compared to a data-only approach. More importantly, the mixed model performs not less than the data-only approach after 15 drive days, so there is no harm using the mixed model approach either. This makes the mixed model a very scalable approach as it difficult to predict what the optimal number of drive days per street segments should be in other campaigns. For RMSE, the mixed model always outperforms the data-only and LUR model approach.

In our previous work, a mobile monitoring campaign in Amsterdam and Copenhagen, we only managed to have 7 repeats per street segment on average. However, also with less repeats than in this paper, we found that the mixed model was able to outperform data-only and LUR model approaches when compared to external long-term measurements ($r = 0.85$ vs 0.74 and 0.75). We found comparable results in this study. The mixed model based on 7 repeats correlated 0.83 to the long-term average of 40 repeats, while the data-only and LUR model with 7 repeats correlated 0.75 and 0.65, respectively.

For the consecutive days analyses, we also found that the mixed model outperforms the other two methods. Within the first few drive days, there is wide variance in performance since not all streets have been measured yet and the mixed model can only use the LUR model as estimate for the average concentration on each street segment. After about 100 days, the data only approach has enough repeats to match the performance of the mixed model.

4. Conclusions

The mixed model approach was able to explain variance in the street level average BC concentrations in Oakland with fewer repeated measurements per street segment than the data-only and LUR model approach. The mixed model approach needed on average five repeats per street segment to achieve 80% explained variance, whereas eight repeats were needed in the data-only approach. The LUR model never achieved an explained variance higher than 70%. We found that at least 15 repeats are needed to generate a robust ($R^2 > 0.9$) long-term average concentration per street segment based on measurements only. This would be possible in a small area, like a few streets or a neighbourhood but is a huge effort for a regular city. This means that when looking for

hotspots in concentration levels a large number of repeated measurements are needed, though this holds for mixed models as well because a LUR model can only find hotspots related to known sources.

So far, very few mobile monitoring campaigns could measure every street segment at least 15 times in a city. On the other hand, LUR models can generate robust long-term average concentration maps with only a limited number of mobile measurements (in coverage of the domain and repeats), but then hyperlocal variation in air pollution is lost. The mixed model is therefore a flexible and scalable solution because it can create a long-term average concentration map with limited effort without having to measure every street segment, while being able to update the map when more drives per street segment become available. If the goal of a mobile monitoring campaign is predicting the absolute values of long-term concentration levels closely, measuring with differing meteorology, at different times of the day, and in different seasons is important.

We found that the mixed-model already improves predictions compared to LUR models and data-only with two drive days (both in R^2 and RMSE). This means it can be used in a dedicated sampling campaign and in an opportunistic setting where commercial vehicles measure air pollution along their routes. Since mixed models cannot perform worse than the LUR model and data-only approach going into the model, using mixed models in mobile monitoring campaigns has no downside, suggestingthat this method can be used with other pollutants and in other locations.

CRediT authorship contribution statement

Jules Kerckhoffs: Conceptualization, Methodology, Formal analysis, Writing – original draft, Visualization. **Gerard Hoek:** Conceptualization, Writing – review & editing. **Roel Vermeulen:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2023.117457.

References

Apte, J.S., Messier, K.P., Gani, S., Brauer, M., Kirchstetter, T.W., Lunden, M.M., Marshall, J.D., Portier, C.J., Vermeulen, R.C.H., Hamburg, S.P., 2017. High-resolution air pollution mapping with Google street View cars: exploiting big data. Environ. Sci. Technol. 51 (12), 6999–7008. https://doi.org/10.1021/acs.est.7b00891.

Hasenfratz, D., Saukh, O., Walser, C., Hueglin, C., Fierz, M., Arn, T., Beutel, J., Thiele, L., 2015. Deriving high-resolution urban air pollution maps using mobile sensor nodes. Pervasive Mob. Comput. 16 (PB), 268–285. https://doi.org/10.1016/j.pmcj.2014.11.008.

Hatzopoulou, M., Valois, M.F., Levy, I., Mihele, C., Lu, G., Bagg, S., Minet, L., Brook, J., 2017. Robustness of land-use regression models developed from mobile air pollutant measurements. Environ. Sci. Technol. 51 (7), 3938–3947. https://doi.org/10.1021/acs.est.7b00366.

J, K., H, G., R, V., 2019. A nationwide land use regression model for ultrafine particles. Environmental Epidemiology 3, 195. https://doi.org/10.1097/01.ee9.0000607960.60423.0e.

Kerckhoffs, J., Hoek, G., Vlaanderen, J., van Nunen, E., Messier, K., Brunekreef, B., Gulliver, J., Vermeulen, R., 2017. Robustness of intra urban land-use regression models for ultrafine particles and black carbon based on mobile monitoring. Environ. Res. 159, 500–508. https://doi.org/10.1016/j.envres.2017.08.040.

Kerckhoffs, J., Khan, J., Hoek, G., Yuan, Z., Ellermann, T., Hertel, O., Ketzel, M., Jensen, S.S., Meliefste, K., Vermeulen, R., 2022a. Mixed-effects modeling framework for Amsterdam and Copenhagen for outdoor NO_2 concentrations using measurements sampled with Google street View cars. Environ. Sci. Technol. https://doi.org/10.1021/ACS.EST.1C05806 acs.est.1c05806.

Kerckhoffs, J., Khan, J., Hoek, G., Yuan, Z., Hertel, O., Ketzel, M., Jensen, S.S., Al Hasan, F., Meliefste, K., Vermeulen, R., 2022b. Hyperlocal variation of nitrogen dioxide, black carbon, and ultrafine particles measured with Google street View cars in Amsterdam and copenhagen. Environ. Int. 170, 107575 https://doi.org/10.1016/J.ENVINT.2022.107575.

Messier, K.P., Chambliss, S.E., Gani, S., Alvarez, R., Brauer, M., Choi, J.J., Hamburg, S.P., Kerckhoffs, J., Lafranchi, B., Lunden, M.M., et al., 2018a. Mapping air pollution with Google street View cars: efficient approaches with mobile monitoring and land use regression. Environ. Sci. Technol. 52 (21), 12563–12572. https://doi.org/10.1021/acs.est.8b03395.

Messier, K.P., Chambliss, S.E., Gani, S., Alvarez, R., Brauer, M., Choi, J.J., Hamburg, S.P., Kerckhoffs, J., Lafranchi, B., Lunden, M.M., et al., 2018b. Mapping air pollution with Google street View cars: efficient approaches with mobile monitoring and land use regression. Environ. Sci. Technol. 52 (21), 12563–12572. https://doi.org/10.1021/acs.est.8b03395.

Sabaliauskas, K., Jeong, C.H., Yao, X., Reali, C., Sun, T., Evans, G.J., 2015. Development of a land-use regression model for ultrafine particles in Toronto, Canada. Atmos. Environ. 110, 84–92. https://doi.org/10.1016/j.atmosenv.2015.02.018.

Shairsingh, K.K., Jeong, C.H., Wang, J.M., Evans, G.J., 2018. Characterizing the spatial variability of local and background concentration signals for air pollution at the neighbourhood scale. Atmos. Environ. 183, 57–68. https://doi.org/10.1016/j.atmosenv.2018.04.010.

Weichenthal, S., Ryswyk, K. Van, Goldstein, A., Bagg, S., Shekkarizfard, M., Hatzopoulou, M., 2016. A land use regression model for ambient ultrafine particles in Montreal, Canada: a comparison of linear regression and a machine learning approach. Environ. Res. 146, 65–72. https://doi.org/10.1016/j.envres.2015.12.016.