Learning Global Variations in Outdoor $PM_{2.5}$ Concentrations with Satellite Images

Kris Y. Hong [1] **Pedro O. Pinheiro** [2] **Scott Weichenthal** [1]

Abstract

The Global Burden of Disease Study identifies outdoor fine particulate matter ($PM_{2.5}$) as the eighth leading risk factor for premature mortality globally. As such, understanding the global distribution of $PM_{2.5}$ is an essential precursor towards implementing pollution mitigation strategies and modelling global public health. In this paper, we present a convolutional neural network-based method for estimating outdoor $PM_{2.5}$ concentrations using satellite images centred on ground-level measurements. Our method achieves a root mean square error of 13.01 μg/m^3 on the test set, which is comparable to current state-of-the-art statistical models, but relies only on satellite images as input. The model offers a fast, cost-effective means of estimating global $PM_{2.5}$

1. Introduction

Exposure to ambient fine particulate matter ($PM_{2.5}$) is estimated to cause nearly three million premature deaths annually (Stanaway et al., 2018), leading to substantial loss of healthy life years and a global healthcare burden measured in billions of dollars each year (Landrigan et al., 2018). Better estimates of global $PM_{2.5}$ are needed to help inform pollution mitigation strategies and research.

Currently, ground-level $PM_{2.5}$ measurements are very costly to obtain. As such, exposures in locations without measurements are typically estimated using statistical methods (*e.g.* land use regression) that combine geographic information system (GIS) data with ground monitoring data to predict exposures in locations without measurements. While this approach generally works well (Weichenthal et al., 2016; Ryan & LeMasters, 2007), detailed GIS data are often available on a limited spatial scale and land use regression models are not generalizable across cities (Patton et al., 2015).

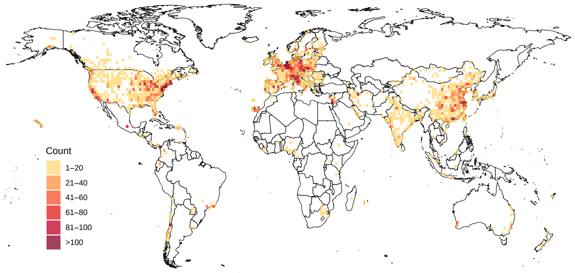

Figure 1. Locations of global monitoring sites for $PM_{2.5}$.

Alternatively, information on traffic, land use, the built environment, and other potential sources of exposure can be captured through satellite imagery. In this study, we explored the use of deep convolutional neural networks (LeCun et al., 1998) for estimating global variations in long-term average outdoor $PM_{2.5}$ concentrations using only satellite images.

2. Methods

2.1. Data Preparation

Long-Term Average Outdoor $PM_{2.5}$ Data. We extracted a global dataset of annual average ground-level $PM_{2.5}$ measurements and their corresponding latitude-longitude coordinates from the World Health Organization (WHO16). These data were collected primarily between 2010 and 2016 and included approximately 20,000 measurements from approximately 6,000 unique monitoring sites in 98 countries, shown in Figure 1.

Satellite Images. We downloaded satellite images centered on each ground-level $PM_{2.5}$ coordinate from Google Static Maps using the ggmap package in R (R Development Core Team, 2010; Kahle & Wickham, 2013). We downloaded four satellite images for each monitoring site, differing by integer zoom levels ranging from 13 (covering approximately 10×10km) to 16 (approximately 1.5×1.5km). All images have dimensions of $256 \times 256 \times 3$ to maintain a

[1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada [2] Element AI, Montreal, Canada. Correspondence to: Scott Weichenthal <scottandrew.weichenthal@mcgill.ca>.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

reasonable training time. The satellite images were captured in 2018.

Data Processing. All latitude-longitude coordinates for PM$_{2.5}$-image pairs were first geohashed to a precision of three (Niemeyer, 2008). The selected precision level of three corresponds to cells with areas less than 156×156km, with widths decreasing moving from the equator to the poles. The database was then randomly split into training (80%), validation (10%), and test sets (10%) such that the three sets were disjoint by geohash codes.

Multiple ground-level PM$_{2.5}$ measurements were available for some sites. This meant that multiple exposure values (*i.e.* year-to-year changes in annual average PM$_{2.5}$ concentration at the same location over time) could be assigned to the same satellite image. We approached this issue in two ways: 1) Models were developed averaging all available exposure data for each latitude-longitude pair; 2) Models were developed without averaging allowing individual images to have different exposure values based on changes in annual average PM$_{2.5}$ concentrations over time. Preliminary results favoured the second approach (*i.e.* allowing the same image to have different PM$_{2.5}$ concentrations over different years) and therefore we focused on this approach. As a result, global model evaluation was also based on single year annual average ground-level measurements.

2.2. Training and Evaluation

We designed two models to predict spatial variations in outdoor PM$_{2.5}$: one on a continuous scale using linear activations and another across deciles of exposure (ten balanced categories obtained by evenly splitting the database by the deciles of the exposure distribution) using softmax (Bridle, 1990). We trained all models with a fixed input size of $256 \times 256 \times 3$. We used dropout (with rates of 0.5) after the convolutional backbone, and after the densely connected network. All backbone models were initialized with pretrained ImageNet weights, and all models were trained using a batch size of 64 images (16 images per GPU) for up to 100 epochs. The learning rate was decreased by a factor of 0.1 if the validation accuracy did not improve for 10 epochs. We stopped the training if the validation accuracy did not improve for 20 epochs.

Final model selection was based on a systematic evaluation of several well-known architectures for the convolutional base including InceptionV3 (Szegedy et al., 2016), Xception (Chollet, 2017), and VGG16 (Simonyan & Zisserman, 2015). In addition, several optimizers were tested including RMSProp (Tieleman & Hinton, 2012) and Nadam (Dozat, 2016) with learning rates of 0.001 and 0.0001. A detailed leaderboard was maintained, tracking the performance of different combinations of model architectures and hyperparameters. The Xception architecture combined with the Nadam optimizer at a learning rate of 0.0001 performed best on the validation set, and these results are described in detail. For the final classification model, gradient-weighted class activation maps (Selvaraju et al., 2017) were used to examine specific portions of images used to make predictions.

Architecture	Zoom	Decile Class. Accuracy (%)	SD	RMSE
Xception	13	35.33	23.70	13.63
	14	33.06	23.70	14.18
	15	31.61	23.70	13.64
	16	31.61	23.70	14.31

Table 1. Model performance on the validation set across different zoom levels. The standard deviation of PM$_{2.5}$ values in the validation set are shown as a baseline for evaluating RMSE values.

Evaluation. For each task of predicting continuous/categorical PM$_{2.5}$, the model with the highest validation classification accuracy (for decile predictions) or the lowest validation root mean square error (RMSE) (for continuous predictions) was retained. For categorical models, we also report the “one-off accuracy” which reflects the proportion of the time the model predicts the correct class or one category away from the correct class.

As an additional model evaluation step, we compared continuous PM$_{2.5}$ estimates from our final global model (called IMAGE-PM$_{2.5}$) to those of the Data Integration Model for Air Quality (DIMAQ) used by the Global Burden of Disease study (Shaddick et al., 2018a;b). This comparison was conducted for approximately 9000 locations (113 countries) between 2010 and 2016 with 34,794 annual average measurements ranging from <1 μg/m^3 to 332 μg/m^3 (mean=20.04 μg/m3, SD=18.76 μg/m^3). In addition, we compared our global model estimates to mean DIMAQ estimates averaged over the entire 2010-2016 period. Finally, we calculated site-specific differences between our IMAGE-PM$_{2.5}$ estimates and mean DIMAQ estimates to evaluate potential geographic patterns in the magnitude of disagreement between the two models.

3. Results

The global database contained approximately 19,650 pollution-image pairs with annual mean PM$_{2.5}$ concentrations ranging from less than 1 μg/m^3 to 436 μg/m^3 with a mean value of 23.2 μg/m^3 (SD=22.9 μg/m^3).

Zoom level 13 satellite images performed best for both classification and regression (Table 1). Specifically, the final categorical model had a validation accuracy of 35.33%

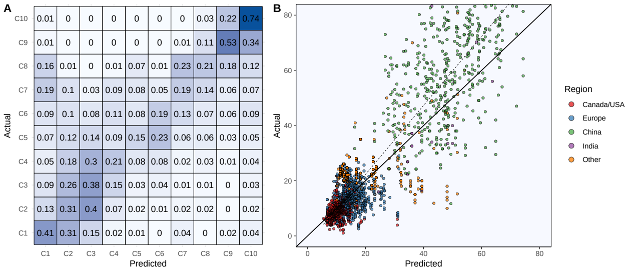

Figure 2. Measured versus predicted global $PM_{2.5}$ concentrations in the *test set* for 10-category classification (A) and regression (B).

across deciles (10% accuracy would be expected by random chance) (Table 1). The confusion matrix in Figure 2A illustrates model performance on the test set and indicates that predictions were best at lower and upper deciles with decreasing performance towards the inner classes. Overall, the global categorical model achieved a test accuracy of 33.69% and a one-off test accuracy of 65.71%.

For the global IMAGE-$PM_{2.5}$ continuous model, the lowest validation RMSE value was 13.63 μg/m^3 (Table 1). On the test dataset, the global model achieved an RMSE value of 13.01 μg/m^3 with an R2 value of 0.75 (Figure 2B); however, model predictions tended to underestimate measured values at higher concentrations as indicated by the dashed fit-line in Figure 2B.

We used gradient-weighted class activation maps (Selvaraju et al., 2017) to identify specific portions of images used for predictions. Figure 3 shows class-activation maps for five locations that were correctly classified across deciles of long-term $PM_{2.5}$ concentrations. From this figure it is clear that localized portions of each satellite image are being used to make predictions; however, the specific ground-level features that are playing the most important role remain unclear.

Continuous estimates of annual average $PM_{2.5}$ concentrations from IMAGE-$PM_{2.5}$ model were highly correlated (R2=0.79; slope = 1.019, 95% CI: 1.014, 1.025) with those predicted by the Data Integration Model for Air Quality (DIMAQ) used by the Global Burden of Disease (GBD) study. Agreement between the two models improved slightly when we compared IMAGE-$PM_{2.5}$ predictions to DIMAQ model estimates averaged over the entire seven-year period tested (2010-2016): R2=0.81; slope=1.022 (95% CI: 1.012, 1.025).

Figure 4 shows the global distribution of differences between long-term estimates of mean $PM_{2.5}$ concentrations (2010-2016) at the 9,000 sites compared in this analysis. Agreement was best in North America, Europe, and China. The largest differences were observed in regions where ground level $PM_{2.5}$ values (used in DIMAQ) were based predominantly (>70% of values) on PM_{10} data including India, Turkey, Romania, and Lithuania.

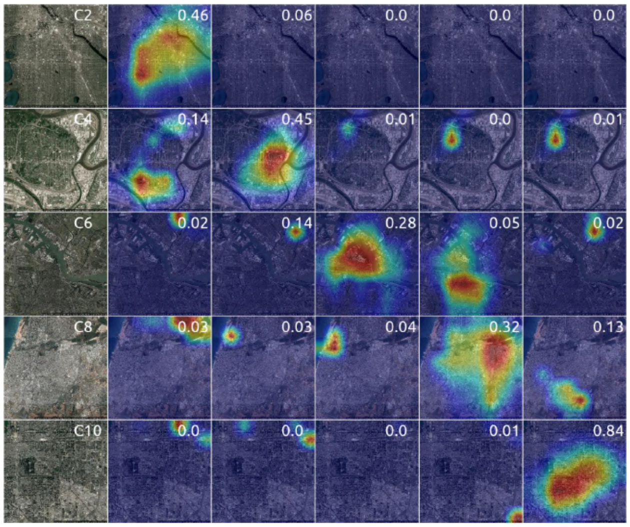

Figure 3. Gradient-weighted class activation maps (Grad-CAMs) for images correctly classified by the final global categorical model. The first column is the original input image. The second through sixth columns are the Grad-CAMs for classes 2, 4, 6, 8, and 10, respectively. Numerical values on the top-right indicate the predicted probability that the image belongs to the respective class. The cities are Minneapolis, US (C2); Kansas City, US (C4); Amsterdam, NL (C6); Tel Aviv, IL (C8); and Beijing, CN (C10).

4. Discussion

In this study we explored the use of convolutional networks as an alternative, cost-effective means of estimating global variations in long-term average outdoor $PM_{2.5}$ concentrations. In particular, we examined this approach across the global concentration range using ground monitoring data available from WHO16. To the best of our knowledge, this is the first study to explore the use of deep learning in estimating global variations in long-term average outdoor $PM_{2.5}$ concentrations and we noted several interesting findings.

First, the predictive performance of the IMAGE-$PM_{2.5}$ model presented in this study was similar to that of current state-of-the-art Bayesian hierarchical models employing combinations of remote sensing, chemical transport models, land use, and other information (Shaddick et al., 2018a;b). This is somewhat surprising given the wealth of source/emissions information included in state-of-the-art models. Specifically, Shaddick *et al.* (Shaddick et al., 2018a;b) reported a population-weighted RMSE value of 12.10 μg/m^3 (R2=0.91) for the DIMAQ model used in the Global Burden of Disease Study whereas the IMAGE-$PM_{2.5}$ in our investigation achieved an RMSE value of 13.01

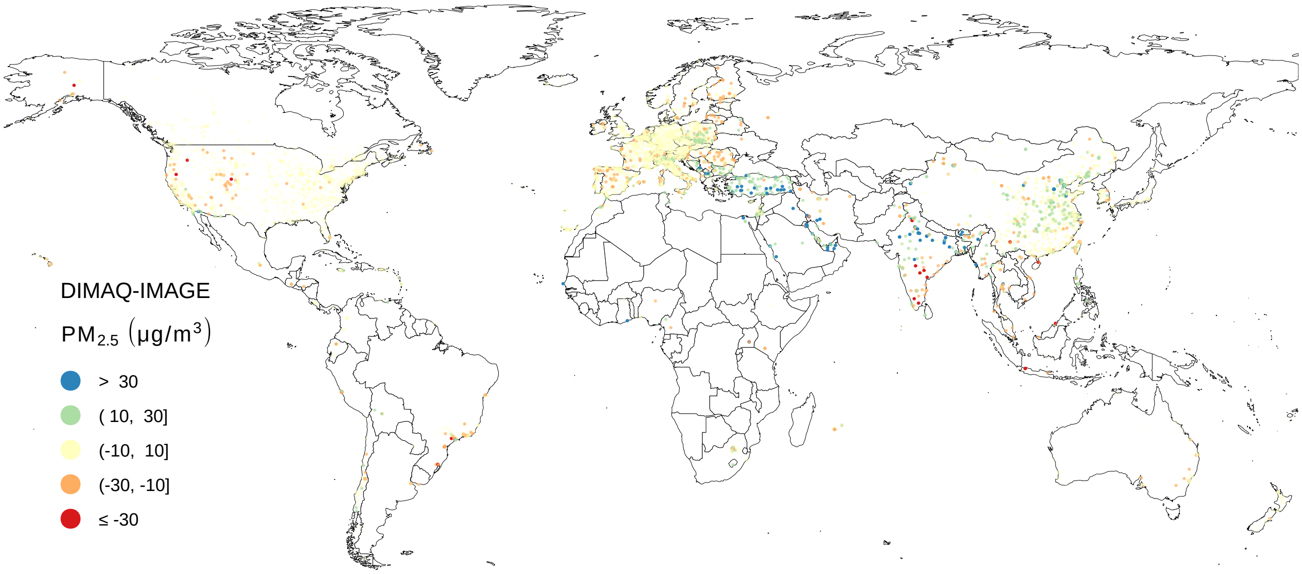

Figure 4. Differences in predicted long-term average $PM_{2.5}$ concentrations (2010-2016) using the IMAGE-$PM_{2.5}$ model and the DIMAQ model (Shaddick et al., 2018a;b).

μg/m^3 (R2=0.75) over a similar concentration range. In addition, our directi comparison of DIMAQ and IMAGE-$PM_{2.5}$ predictions indicated a strong correlation between model estimates with a slope close to 1. Interestingly, the largest discrepancies between the two models occurred in regions where ground level $PM_{2.5}$ data were derived from PM_{10} measurements. As the DIMAQ model incorporated $PM_{2.5}$ data derived from PM_{10} measurements and the IMAGE-$PM_{2.5}$ model did not, this difference may explain the larger discrepancies in these areas. Our IMAGE-$PM_{2.5}$ model may offer useful prior information for Bayesian hierarchical models such as DIMAQ when ground level measurements or emissions data are not available.

One of the clear disadvantages of deep learning models is the lack of transparency in how model predictions are generated. Deep convolutional neural networks are somewhat less opaque in that class activation maps can be used to investigate image characteristics/patterns used to make predictions. Our results suggest that model predictions of ground-level $PM_{2.5}$ concentrations were based on localized portions of satellite images and that both color and combinations of colors and geometric features were used in making predictions. However, it was not possible to identify specific aspects of the built environment that played an important role in generating model estimates. Interestingly, the zoom level of satellite images had an important impact on model performance and future studies should explore other image characteristics that could be optimized to reduce model errors. Likewise, as deep convolutional neural networks can have multiple inputs, it may be possible to incorporate additional ground-level information (*e.g.* sources, businesses, population density, etc.) within each image to capture more detailed data on local sources of $PM_{2.5}$ and thus improve model performance. A second limitation of our analysis was that the timing of satellite images did not overlap exactly with the timing of $PM_{2.5}$ measurements/estimates. This may have contributed error to our predictions in locations where major infrastructure changes were made between the time of $PM_{2.5}$ measurements and satellite imaging. However, variation in within-site $PM_{2.5}$ over the 2010-2016 period was generally small (SD<5 μg/m^3 for 80% of sites). We aim to include temporally matched satellite images in future iterations which may improve the model's performance. Moreover, our IMAGE-$PM_{2.5}$ model is also limited in that it does not contain a temporal component: predictions only change if the image changes. Therefore, the IMAGE-$PM_{2.5}$ model cannot be used to estimate short-term (*i.e.* year to year) changes in outdoor $PM_{2.5}$ concentrations and this limitation will be addressed in our ongoing work.

In summary, we developed a new method of estimating global variations in long-term average outdoor $PM_{2.5}$ concentrations using convolutional networks trained with a large dataset of satellite images and ground level measurements. Our new global IMAGE-$PM_{2.5}$ model relies on a single satellite image as input and can provide fast, cost-effective estimates of $PM_{2.5}$ concentrations with predictive performance comparable to modern Bayesian hierarchical models currently used by the Global Burden of Disease Project (Shaddick et al., 2018a;b). These findings represent an important advancement in our current understanding of how global variations in long-term average $PM_{2.5}$ concentrations can be modelled for global health applications. The

IMAGE-PM$_{2.5}$ model can be used as a stand-alone method of global exposure estimation or incorporated into more complex hierarchical model structures.

Acknowledgements

We graciously acknowledge support from an NSERC Discovery Grant, a CIHR Foundation Grant, the Microsoft Azure Compute Grant, and the the NVIDIA GPU Grant. We also thank the reviewers for their help in improving our paper.

References

Bridle, J. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. *Neurocomputing: Algorithms, Architectures and Applications*, 1990.

Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Dozat, T. Incorporating nesterov momentum into adam. 2016.

Kahle, D. and Wickham, H. ggmap: Spatial visualization with ggplot2. *The R Journal*, 2013. URL https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

Landrigan, P. J., Fuller, R., Acosta, N. J., Adeyi, O., et al. The lancet commission on pollution and health. *The Lancet*, 2018.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Niemeyer, G. Geohash. http://geohash.org, 2008. Accessed: 2019-03-24.

Patton, A. P., Zamore, W., Naumova, E. N., Levy, J. I., Brugge, D., and Durant, J. L. Transferability and generalizability of regression models of ultrafine particles in urban neighborhoods in the boston area. *Environmental science & technology*, 2015.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, 2010. URL http://www.r-project.org.

Ryan, P. H. and LeMasters, G. K. A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhalation toxicology*, 2007.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*, 2017.

Shaddick, G., Thomas, M. L., Amini, H., Broday, D., Cohen, A., Frostad, J., Green, A., Gumy, S., Liu, Y., Martin, R. V., et al. Data integration for the assessment of population exposure to ambient air pollution for global burden of disease assessment. *Environmental science & technology*, 2018a.

Shaddick, G., Thomas, M. L., Green, A., Brauer, M., Donkelaar, A., Burnett, R., Chang, H. H., Cohen, A., Dingenen, R. V., Dora, C., et al. Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2018b.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

Stanaway, J. D., Afshin, A., Gakidou, E., Lim, S. S., Abate, D., , et al. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 2018.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

Tieleman, T. and Hinton, G. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

Weichenthal, S., Van Ryswyk, K., Goldstein, A., Shekarrizfard, M., and Hatzopoulou, M. Characterizing the spatial distribution of ambient ultrafine particles in toronto, canada: A land use regression model. *Environmental pollution*, 2016.

WHO16. Who global urban ambient air pollution database (update 2016). https://whoairquality.shinyapps.io/AmbientAirQualityDatabase/. Accessed: 2019-03-24.