Combining street imagery and spatial information for measuring socioeconomic status

Esra Suel[1,2], **Marthe Boulleau**[2], **Majid Ezzati**[1], **Seth Flaxman**[2,3]
[1]School of Public Health, [2]Data Science Institute, [3]Department of Mathematics
Imperial College London
{esra.suel, marthe.boulleau14, majid.ezzati, s.flaxman} @imperial.ac.uk

Abstract

Emerging sources of large-scale data, such as remote sensing, street imagery, GPS trajectories, coupled with advances in deep learning methods have the potential for significantly advancing how fast, how frequently, and how locally we can measure urban features and population characteristics to inform and evaluate policies. One such example that attracted increasing attention from the research community is utilizing street level imagery for various measurement tasks in this broader context. We believe incorporating spatial information with Gaussian Processes (GPs) can give us better performance when using street images. To test this hypothesis, we empirically investigated multiple approaches for combining spatial and street image information using neural networks and GPs for predicting income, crowding, and education levels in London, UK. Results demonstrated using GPs only with spatial information (without any inputs from images) gives us a good baseline. Complementary value of street images were demonstrated for the socioeconomic status measures we investigated. Further, our results showed superior performance of GP regression of residuals compared to other methods including feeding spatial information as input directly to neural networks.

1 Introduction and Related Work

Urbanization and social inequalities are two of the major policy themes of our time, intersecting in large cities where rich and poor live side by side. Reducing inequalities is at the forefront of the global sustainable development agenda as well as a policy objective in many cities [GLA, 2017, 2018]. However, datasets for informing these policies and measuring their actual impacts are currently from disjointed, and inefficient surveillance systems. Measuring socioeconomic status (SES) at high spatial and temporal resolution, for instance, is crucial yet poses a significant challenge even in developed parts of the world. Emerging sources of large-scale data, such as remote sensing, imagery, and GPS trajectories, have the potential for significantly advancing how fast, how frequently and how locally we can measure urban features and population characteristics.

Researchers both from domain sciences and computer science are increasingly interested in tackling problems associated with applying advanced learning techniques focusing on automatic feature extraction to measurement and data collection tasks. Relevant applications of machine learning with imagery include: poverty detection [Jean et al., 2016, Steele et al., 2017, Xie et al., 2015] and harvest size and crop yield [Lobell, 2013, You et al., 2017] from satellite data, and income [Gebru et al., 2017], perceived safety [Naik et al., 2014, 2017], and greenness and openness [Seiferling et al., 2017, Richards and Edwards, 2017] from Google Street View (GSV) images.

Here, building on our previous work, we use London to evaluate the feasibility of using street level images for measuring multiple indicators of socioeconomic status including income, crowding, and

Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

education. We believe incorporating spatial information with Gaussian Processes (GPs) can give us better performance; to test this hypothesis we compare measurement performances of multiple methods for combining street images and spatial information using neural networks (NN) and GPs. Two recent studies by You et al. [2017] and Jean et al. [2018] have proposed using GPs in the context of making predictions from satellite imagery. In our work, we also investigate the use of GPs combined with NN, but on a problem with different types of images (i.e. street level images) and multiple SES indicators as outputs. Our focus here is to investigate the performance of multiple approaches, some with NN only and some with a combination of NN and GPs.

2 Data and Methods

We first obtained the Office for National Statistics (ONS) Postcode Directory for the UK[1] and selected the 181,150 postcodes assigned to the 33 local authority districts of the Greater London administrative area. Unique images were available from GSV for 119,681 postcode locations in London. Four images for each location were extracted by specifying the camera direction (i.e., 0°, 90°, 180°, 270°) relative to the GSV vehicle to cover a 360° view. Hence, we used a total of 478,724 images corresponding to 119,681 locations in London.

Output data (labels) was obtained from the UK Census 2011[2] and Greater London Authority household income estimates 2015[3]. In the UK, Census results and other neighborhood statistics are reported using output areas designed specifically for statistical purposes. The output labels used here are detailed below and were available for lower super output areas (LSOA; average population of 1,614 with a total of 4,833 LSOAs in London). For income, we used the mean annual household income estimates by Greater London Authority, reported for each LSOA. For the regression tasks, the income values that were used are represented in 10000 GBPs taking values between 3 and 19. For education, we used the percentage of population with low educational attainment levels (i.e. people who do not have at least a Level 2 education where the five categories for highest attained qualification were: no qualification, Level 1, Level 2, Level 3, and Level 4 and above). For crowding, we used the percentage of households classified as being overcrowded (i.e., having at least one fewer room than required) by ONS. ONS derives the number of rooms required using a formula based on ages of the household members and their relationships to each other. For education and crowding, percentages of deprived populations for each LSOA are used, hence the values are between 0 and 1.

Images correspond to photos taken by GSV vehicles at specific locations represented by coordinates and postcodes. Each postcode location is assigned to a single LSOA where ground truth data is available; individuals GSV images were matched with output labels using this information.

We took a transfer learning approach where we used VGG16 network [Simonyan and Zisserman, 2014] pre-trained with ImageNet [Russakovsky et al., 2015] as a fixed feature extractor to convert RGB images to 4096 dimensional codes. On top of the convolutional neural network (CNN) layers with pre-trained weights, we trained a fully connected NN to perform regression from four GSV images extracted for each location. In some of our experiments longitude and latitude coordinates were also fed as inputs to the trained part of the network with a slightly different network architecture explained further in the next section. We compute LSOA level predictions as the average of postcode level predictions assigned to that LSOA. To evaluate performance, we used 60% of LSOAs for training and made predictions for the held-out 40% test set; 5-fold cross validation was carried out for hyper parameter tuning and NN architecture selection using the training set only. In a set of our experiments, GP regression was used as an additional separate step for incorporating spatial information (i.e longitude and latitude coordinates of LSOA centroids) on top of CNN outputs from images, as detailed in the next section.

3 Experiments and Results

We compared seven different approaches in terms of prediction performance. In the first approach, CNN-I, four images for each location were fed into the VGG16 network and 4x4096 codes were extracted from layer $h5$. A fully connected NN was trained using the architecture presented in Table 1.

[1] `https://ons.maps.arcgis.com/home/item.html?id=1e4a246b91c34178a55aab047413f29b`

[2] `https://www.ons.gov.uk/census/2011census`

[3] `https://data.london.gov.uk/dataset/household-income-estimates-small-areas`

The main principle for the fully connected part is that the network used all four images from each location jointly in the four channels shown; the information coming from different channels are then aggregated and fed into the final layers to yield a single continuous output. No location information is used in this experiment.

The second approach, CNN-IC, used spatial information (i.e. latitude/longitude) in addition to GSV images, using the architecture shown in Table 1. The same architecture from CNN-I was used for the initial layers, the 1D final layer output was then concatenated with longitude and latitude coordinates associated with each image location yielding a 3D input to the following layers of the network.

NN-C, the third approach was used as baseline comparison where only coordinate information is used. The same architecture as in CNN-IC was used where the 1D input from CNN-I architecture was replaced with random noise as in Table 1, allowing the use of exactly the same architecture.

CNN-I			
I-1	I-2	I-3	I-4
VGG	VGG	VGG	VGG
4096	4096	4096	4096
512	512	512	512
256	256	256	256
128	128	128	128
$\oplus$			
64			
1			

CNN-IC			
I-1	I-2	I-3	I-4
VGG	VGG	VGG	VGG
4096	4096	4096	4096
512	512	512	512
256	256	256	256
128	128	128	128
$\oplus$			
64			
1			
$\oplus$	1	lat	lng
3			
10			
10			
5			
1			

NN-C			
$x \sim \mathcal{N}(0,1)$			
$\oplus$	1	lat	lng
3			
10			
10			
5			
1			

Table 1: Network architectures used for NN only approaches; GP approaches use CNN-I predictions as explained in more detail below. Light gray cells represent input features, black cells represent output values, and the darker gray cells represent feature extraction with pre-trained weights using VGG16. I-1, I-2, I-3, and I-4 correspond to four street level images extracted for each location using different camera directions 0°, 90°, 180°, 270° capturing a 360° view.

For the next set of approaches, several alternatives for using GPs to incorporate spatial information were investigated. The main idea is to get LSOA level predictions using CNN-I, and then incorporate spatial information using a GP based on centroid coordinates of LSOAs. In all the approaches, we used Matern-3/2 kernel for the GPs based on initial five-fold cross-validation experiments on the training set. Throughout the experiments, we used the GPy [GPy, 2012] package to fit the GP models, where all the parameters of the Matern-3/2 kernel and the noise parameters were determined using maximum-likelihood (II) estimation, optimizing the log-marginal likelihood.

The first GP-based approach, GP on Coordinates - GPC, does not use the image information at all and simply fits a GP regressor on the coordinates to predict labels from the training set. The second approach, GP for Residuals on coordinates - CNN-I-GPR, fits a GP regressor to the residuals of the CNN-I predictions. The third approach, GP on coordinates and Predictions with Product Kernel - CNN-I-GPPK, feeds predictions of CNN-I as additional input to the GP regressor that uses a product kernel with two terms, the first using spatial coordinates and second CNN-I predictions. We note that using a linear model for the second kernel yields a similar method as CNN-I-GPR hence we explore Matern-3/2 kernel instead. We also empirically evaluated an alternative approach, where the CNN-I predictions were concatenated with the spatial coordinates instead of using a product kernel, the performances were lower. We believe the cause is the differences in numerical ranges; different spatial lengths would be required when concatenated in the same kernel for good performance. In the last approach, we refer to as CNN-I-GPC, we use GP based on coordinates as a spatial smoothing model and fit it to ground truth labels on the training set and CNN-I predictions on the test set. Observation noise is assumed minimal on the ground truth set while a higher value is assigned to the

predictions on the test set. This latter higher value of noise used on predictions was estimated based on five-fold cross validation experiments on the training set.

	Income			Crowding			Education		
	r	NRMSE	MAPE	r	NRMSE	MAPE	r	NRMSE	MAPE
CNN-I	.84	.057	12	.81	.009	65	.78	.009	27
CNN-IC	.85	.049	12	.82	.007	46	.82	.006	22
NN-C	.83	.052	12	.77	.009	55	.72	.010	26
GPC	.90	.031	8	.83	.007	40	.82	.007	21
CNN-I-GPR	.93	.023	7	.88	.005	31	.88	.005	17
CNN-I-GPPK	.92	.026	7	.89	.005	28	.89	.004	16
CNN-I-GPC	.90	.032	8	.82	.008	38	.82	.007	20

Table 2: Prediction performances of investigated approaches on the test set. r: Pearson's correlation coefficient, NRMSE: normalized root mean squared error, MAPE: mean absolute percentage error.

Results are shown in Table 2. In line with previous work, prediction performances for all output labels were high, demonstrating the potential for using GSV images and corresponding location information for improving measurements in cities. There are differences in performance across different labels, that potentially relate to differences in available information contained in imagery and space. We note the difference in how accuracy metrics relate to interpretable units. For income, MAPE will relate to errors in absolute values of income in GBPs. For crowding and education, it relates to errors in the percentage of the population that live in overcrowded households and with lower educational attainment. Hence we expect the former to have lower values for MAPE, and higher values for NRMSE.

In this study, our focus was on performance comparison of different approaches for combining imagery and spatial information; we observed several interesting points. First, all GP-based approaches for combining imagery and spatial information (i.e. CNN-I-GPR, CNN-I-GPPK, and CNN-I-GPC) outperformed the NN based approach i.e. CNN-IC. Strikingly, GPC that uses only the coordinates achieved very similar performance, even higher in some cases, compared to CNN-I and CNN-IC. It gives us a good baseline to which image only predictions should always be compared. Using NN only on the coordinates i.e. NN-C did not show similar success. These results demonstrated the richness of spatial information and success of GPs in leveraging it to improve performances of learning based measurement tasks. It also suggests that a stricter evaluation, leaving out larger neighborhoods, is necessary, as the current train/test split might favor GP regression with test locations located among train locations.

Second, combined use of imagery and coordinates improved performance compared to baseline where only images or coordinates were used. Specifically, CNN-IC outperformed CNN-I and CNN-C for all indicators. Similarly, CNN-I-GPR and CNN-I-GPPK outperform GPC. Such difference suggests complementary information is available from GSV images for predicting these measures.

Third, the way in which one integrates CNN based approaches with GP also made a difference. CNN-I-GPR and CNN-I-GPPK yielded higher performance increase compared to CNN-I-GPC. Both of these models aim to learn the residual error structure with different means. Their performance were similar across the different measures.

4 Conclusions

In experiments with multiple SES indicators from London, i.e. income, crowding, and education, we find that GP-based methods for explicitly incorporating spatial information with GSV imagery can substantially improve prediction and measurement performance. The focus here was on making predictions for the city where we also do the training. Transferability of learned features to different target cities using CNN-I only was investigated in previous work. Investigation of transferability of learned spatial correlation structures to different target cities will be very valuable especially in places where measurement data on SES indicators is not available or outdated - we leave this to future work. Additionally, GPs conveniently give Bayesian posteriors that are useful for quantifying uncertainty. In future work, we also plan to investigate predictions from GPs not just for RMSE but also in terms of how well-calibrated the uncertainty intervals are.

References

Greater London Authority GLA. Better health for all londoners: Consultation on the london health inequalities strategy, 2017.

Greater London Authority GLA. Inclusive london: The mayor's equality, diversity, and inclusion strategy, 2018.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Jessica E Steele, Pål Roe Sundsøy, Carla Pezzulo, Victor A Alegana, Tomas J Bird, Joshua Blumenstock, Johannes Bjelland, Kenth Engø-Monsen, Yves-Alexandre de Montjoye, Asif M Iqbal, et al. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14 (127):20160690, 2017.

Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Transfer learning from deep features for remote sensing and poverty mapping. *arXiv preprint arXiv:1510.00098*, 2015.

David B Lobell. The use of satellite data for crop yield gap analysis. *Field Crops Research*, 143: 56–64, 2013.

Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *AAAI*, pages 4559–4566, 2017.

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, page 201700035, 2017.

Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 779–785, 2014.

Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576, 2017.

Ian Seiferling, Nikhil Naik, Carlo Ratti, and Raphäel Proulx. Green streets- quantifying and mapping urban trees with street-level imagery and computer vision. *Landscape and Urban Planning*, 165: 93–101, 2017.

Daniel R Richards and Peter J Edwards. Quantifying street tree regulating ecosystem services using google street view. *Ecological indicators*, 77:31–40, 2017.

Neal Jean, Sang Michael Xie, and Stefano Ermon. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. *arXiv preprint arXiv:1805.10407*, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

GPy. GPy: A gaussian process framework in python. `http://github.com/SheffieldML/GPy`, 2012.