Twitter-based Sensing of City-level Air Quality

Polychronis Charitidis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos and Yiannis Kompatsiaris
Information Technologies Institute, Center for Research and Technology - Hellas
Email: {charitidis,espyromi,papadop,ikom}@iti.gr

***Abstract*—This paper investigates the feasibility of estimating current air quality conditions in cities without official air quality monitoring stations based on a statistical analysis of Twitter activity. For this purpose, a framework for collecting and geotagging air quality-related Twitter posts is developed and transfer learning is applied to enable estimations for unmonitored cities using data from monitored nearby cities. Experiments carried out on five cities in the UK and five cities in the US suggest that while Twitter-based estimates exhibit very high accuracy, they are outperformed on average by simple spatial interpolation. However, we find that a meta-model that combines estimates from spatial interpolation with Twitter-based ones increases accuracy in distantly located cities, highlighting the merits of Twitter-based air quality estimation and motivating further work on the topic.**

I. Introduction

Air pollution is a major environmental issue that occurs as a result of human activity (e.g. carbon emissions from cars) or natural processes (e.g. volcanic eruptions) and its implications range from damages to food crops and buildings to diseases in animals and humans. According to the World Health Organization (WHO), air pollution is responsible for an estimated 11.6% of all global deaths in 2012, while 92% of world's population live in places where air pollution exceeds WHO guideline limits. Although the problem is more prominent in countries such as China, a large proportion of European populations and ecosystems are also exposed to air pollution levels that exceed the WHO air quality guidelines according to a 2017 report on air quality in Europe from the European Environmental Agency (EEA) [1].

One of the main areas for action to battle the adverse effects of air pollution is raising people's awareness with respect to air quality and providing them access to real-time air quality information [2], [3]. However, the high costs of installation (€5-30K per monitoring device [4]), maintenance and calibration of reference stations result in sparse monitoring networks that provide measurements only in few locations of large cities. As a result, citizens of smaller urban areas and underdeveloped regions, lack accessibility to air quality information. While low-cost sensors provide a promising alternative means of air quality monitoring in such areas, they are characterized by low robustness and measurement repeatability [4] and, despite low, their cost is not negligible.

In the past, several researchers have tried to address the problems of estimating current air quality in unmonitored locations (*spatial* prediction) and short-term air quality forecasting (*temporal* prediction) using statistical approaches that model the relationships between air pollutants and various explanatory variables such as lagged pollutant observations, wind speed, solar radiation, cloud coverage, air temperature, traffic, etc. (see [5] for a detailed review of such methods).

More recently, the rise of online social networks (OSNs) and the wealth of almost real-time information that they provide about a variety of real-world events and phenomena, has motivated the development of air quality prediction methods that are based on a statistical analysis of the publicly available OSN content. That line of work, is based on the view of OSN users acting as "social sensors" [6] and builds upon previous successes on detecting and tracking real-world events (e.g. flu outbreak detection and tracking [7], earthquake detection [8], wildlife roadkill monitoring [9], etc.) based on a statistical analysis of the content posted on these platforms. Methods of this type have so far been applied for the estimation of city-level air quality in China by analyzing content posted in Sina Weibo (a Chinese microblogging website) with encouraging results [10], [11], [12], [13].

In this paper we present the first, to the best of our knowledge, attempt to perform OSN-based city-level air quality estimations outside China. In particular, our study focuses on cities in the UK and the US that exhibit important differences compared to Chinese cities: a) high air pollution events are less frequent and pronounced, b) their population and, consequently, the volume of OSN content they generate is considerably smaller. In addition, our work is the first to use Twitter as data source for air quality estimation. Although this imposes the development of a Twitter-oriented data collection and mining pipeline, it makes the proposed method applicable worldwide. Besides that, previous works consider only 24-hour temporal bins, while in our work we also consider 6-hour and 12-hour ones. While a 24-hour granularity is useful for a post-hoc analysis, finer-grained estimates provide actionable information and it is therefore important to evaluate the accuracy under this setting. Moreover, with the exception of [10], all previous related works build and evaluate estimation models on the same city, while we adopt a more realistic setup where models are evaluated only on cities that have not been used for training.

Our approach collects air quality-related tweets from the Twitter API using a set of air quality-related keywords and then estimates the location to which they refer using a state-of-the-art location estimation method [14]. In the sequel, all tweets falling in a particular spatiotemporal bin are pooled together to form a single textual document that is represented using a Bag-of-Words (BoW) scheme. This representation forms the basis for the developed air-quality estimation models. Compared to simpler types of features such as the number

978-1-5386-0951-4/18/$31.00

of tweets in each spatiotemporal bin and their polarity (i.e. whether they refer to bad or good air quality) that were used in previous works [11], [12], [13] we found that BoW features lead to better results. Finally, our work is the first that recognizes the multi-task nature of spatial air quality prediction and uses multi-task learning techniques (data pooling, joint feature selection and sample weighting) to build a robust, city-invariant model.

Traditional approaches for spatial air quality prediction include spatial interpolation (e.g. Inverse Distance Weighting (IDW) and variations of Kriging [15]), dispersion models [16] and Land Use Regression (LUR) variants [17]. Among these methods, dispersion and LUR are known to generate robust long-term intra-city predictions (when enough data are available) but spatial interpolation is usually preferred for spatially coarser short-term estimations [5]. Therefore, in our study we compare our city-level Twitter-based estimates with those generated by a spatial interpolation method (IDW). Our experiments show that models based only on Twitter information provide fairly accurate estimates but perform worse than spatial interpolation. However, when spatial interpolation estimates are carefully combined with Twitter-based ones, better accuracy can be obtained.

II. Methodology

Our work aims at producing accurate estimates of current air quality conditions for cities without air quality monitoring infrastructure based on Twitter activity and measurements from nearby cities. To simplify our analysis, we focus on estimating $PM_{2.5}$ but it is straightforward to extend the approach to other pollutants. The developed framework consists of three main components: a) data collection, b) feature extraction, c) multi-task learning. Data collection and feature extraction are described in sections II-B and II-C, respectively, while section II-A gives the problem formulation and presents our multi-task learning approach.

A. Problem formulation and transfer learning approach

Spatial prediction deals with the problem of estimating a quantity of interest in a set of locations on which the quantity is not measured, based on measurements of the quantity in another set of monitored locations. In this context, the quantity of interest is city-wise average $PM_{2.5}$ and C_M/C_U correspond to monitored/unmonitored cities. Due to the high correlation between $PM_{2.5}$ values in nearby locations, the problem can be tackled using simple spatial interpolation techniques such as IDW, which estimates $PM_{2.5}$ in an unmonitored city as a weighted average of the observed $PM_{2.5}$ in nearby cities.

In this paper we follow a model-based approach. For each city $c_j \in C_M$ we construct a set of N training examples $D_{c_j} = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\}$ where $\mathbf{x}^i \in R^d$ is a d dimensional input vector that provides an informative summary of the tweets referring to c_j during the i-th temporal bin (see section II-C) and $y^i \in R$ is the average $PM_{2.5}$ concentration in c_j during the same temporal bin. For cities $c_q \in C_U$, only $\mathbf{x}^i$ are available and our aim is to build a model $h_{c_q} : \mathbf{X} \rightarrow Y$ for each $c_q \in C_U$ in order to estimate the unknown y^i.

The problem at hand can be considered as a special type of transfer learning [18] where there are multiple target learning tasks $h_{c_q}, c_q \in C_U$ (as in multi-task learning [19]) for which labeled data are completely unavailable (i.e. unmonitored cities) while there are plenty of training data for a number of auxiliary tasks (nearby cities with air quality measurements) $h_{c_j}, c_j \in C_M$ that are related to the target tasks.

Transfer learning approach: Assuming that air pollution exhibits a similar statistical dependence with Twitter activity in cities that share common characteristics (i.e. $P(Y^{c_j}|\mathbf{X}^{c_j}) \approx P(Y^{c_i}|\mathbf{X}^{c_i})$ as $sim(c_j, c_i) \approx 1$[1]), we follow a data pooling approach and train a regression model h on $D = \bigcup_{c_j \in C_M} D_{c_j}$ that learns to simultaneously minimize the prediction error on all monitored cities and we therefore expect it to yield accurate predictions for the unmonitored cities as well.

Besides data pooling, we also apply explicit feature selection to ensure that the learned model will be constrained to a subset of the Twitter-based features that are highly correlated with $PM_{2.5}$ in all cities. To this end, we compute the *Pearson correlation coefficient* between each feature X_i and the target Y and keep the k features that exhibit the highest correlation. This lower dimensional feature representation is expected to facilitate learning a more robust, city-invariant model.

Finally, under the assumption that the smaller the distance $d(c_j, c_i)$ between two cities, the higher the similarity of their conditional distributions, we develop a weighted data pooling variant where each training example gets a weight that is inversely proportional to the distance between the city it belongs to and the target city.

B. Data collection

Twitter data: Twitter provides a free API[2] that offers real-time access to a sample of its public data. There are two main methods to retrieve tweets using the API. The "location-based" method, allows retrieval of geotagged tweets around an area of interest while the "keyword-based" method retrieves tweets containing specific keywords regardless of location. Some of the previous works that used Sina Weibo as the source OSN, used the location-based method. In Twitter, however, only a tiny fraction of the posts are geotagged (1.5% according to [20]), which significantly limits the number of tweets about air quality that can be collected[3].

Therefore, we applied a data collection approach that combines keyword-based search with location inference. Concretely, we track a list[4] of 120 English air quality-related keywords that was composed with the help of air quality experts and store all the returned tweets. Since the vast majority of the collected tweets are not geotagged, tweet location should somehow be inferred in order to identify tweets related to a city of interest. To this end, previous works simply use the account's declared location as the post's location, assuming that the two locations will coincide in most cases. Here,

[1]Here we assume that $sim(c_j, c_i) = 1$ if c_j, c_i belong to the same country.

[2]https://dev.twitter.com/streaming

[3]In preliminary experiments we found that, e.g., only about 10 air quality-related tweets per day are retrieved in London.

[4]https://goo.gl/FwB5od

TABLE I
DATA COLLECTION STATISTICS

Cntr.	City	Pop.	#stations	avg. $PM_{2.5}$	#tweets per day
UK	London	8.8M	9	11.8	3972
UK	Liverpool	0.5M	2	7.6	108
UK	Manchester	0.5M	3	9.2	321
UK	Birmingham	1.1M	2	10.2	198
UK	Leeds	0.75M	2	10.1	112
US	New York	8.5M	10	7.9	2564
US	Boston	0.7M	4	8.1	574
US	Philadelphia	1.6M	3	10.0	478
US	Baltimore	0.6M	2	8.5	394
US	Pittsburgh	0.3M	2	10.7	169

we follow a more elaborate approach that employs a recent state-of-the-art location estimation method [14]. This method works by dividing the earth surface into rectangular cells, and then computing the probabilities of each term occurring in each cell, using a very large training corpus of geotagged items. Given an item with unknown location, a probability is computed for each cell based on the item's terms and the center of the most likely cell is used as the item's estimated location. Experiments conducted in [21] (section 2.1.2) show that when the confidence of the method for an item's estimated location is higher than 0.6, this location lies within 10 km of the actual location 94% of the time.

In our task, the location estimation method of [14] is applied as follows. Since we are actually interested in the location that the tweet refers to instead of the upload location, we first check if a location can be estimated with high confidence ($\geq$ 0.8) based on the tweet content and in case it does we use it as the tweet location. Otherwise, similarly to previous works, we use the account's declared location as the tweet location. However, instead of relying on simple text matching (which would preclude location recovery in case of location descriptions referring to, e.g., well-known city districts), we again perform location estimation using the account's location description as input. The number of tweets assigned daily to each city using this approach is shown in Table I.

Air quality data: To collect ground truth $PM_{2.5}$ measurements for the selected cities we use the OpenAQ API[5] and retrieve hourly historical measurements from all stations located within each city's bounding box[6]. The number of stations measuring $PM_{2.5}$ in each city is shown in Table I. To calculate a single hourly $PM_{2.5}$ value for each city, we average the measurements of the respective stations.

C. Feature extraction

To generate a descriptive representation of the tweets assigned to a city c during a temporal bin t (i.e. spatiotemporal bin (c, t)), we use a BoW scheme. First, all tweets are preprocessed by applying tokenization, lowercasing and stopword removal. Then, we create a vocabulary $W = \{w_1, \ldots, w_n\}$ that consists of the n=10,000 most frequently occurring words in a random 1 million sample of the collected tweets. Using this vocabulary, a BoW vector $\mathbf{x} = [x_1, \ldots, x_n]$ is generated to represent all tweets in (c, t), where x_i denotes the number of tweets containing w_i divided by the total number of tweets in (c, t). In addition to this 'current' BoW representation, we also generate *lagged* BoW representations (denoted as BoW^{-j}), where instead of considering only the tweets posted during the current temporal bin t we also consider the tweets of the j previous bins. Lagged BoW representations aim at capturing dependencies between Twitter-posts and air pollution that extend beyond the current temporal bin.

[5]https://docs.openaq.org/

[6]Bounding boxes were obtained from Flickr: https://www.flickr.com/places

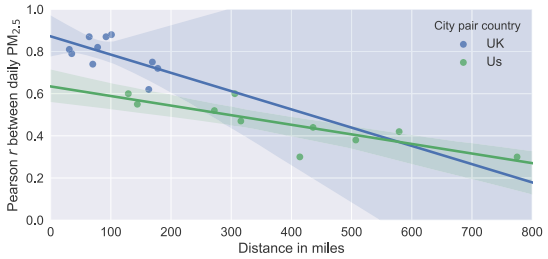

Fig. 1. Scatter plot of distances and Pearson r between average daily $PM_{2.5}$ concentrations for all distinct city pairs in UK and US.

III. EXPERIMENTS

A. Experimental setup

To simulate the spatial air quality prediction task, we collected data for five cities in the UK and five cities in the US (see Table I) for a time period spanning almost a whole year (8/2/2017-18/1/2018). Each city is in turn treated as the test city (hypothetically without air quality measurements) and all the remaining neighboring cities are used for training. For each city, we train and evaluate models able to perform predictions at three different temporal granularities: 6, 12 and 24 hours. This is accomplished by grouping the hourly $PM_{2.5}$ observations into correspondingly sized temporal bins and calculating a single ground truth $PM_{2.5}$ value for each bin as the average of the hourly values. Prediction accuracy for each city and temporal granularity is measured in terms of *Root Mean Squared Error* (RMSE) and macro averaging is applied to calculate country-wise or overall performance (denoted as *aRMSE*). In all our experiments we use *Gradient Tree Boosting* [22] as the regression algorithm, since it is recognized as one of the best off-the-shelf supervised learning algorithms [23] and was found to perform equally good or better compared to other algorithms in a set of preliminary experiments.

B. Baseline performance

Figure 1 shows a scatter plot of the distances and the Pearson r between average daily $PM_{2.5}$ concentrations for all distinct city pairs in UK and US. Clearly, the smaller the distance between two cities, the higher the correlation between their average daily $PM_{2.5}$ concentrations. Given this high spatial dependence, it is not surprising that spatial interpolation methods such as IDW yield highly accurate estimates as shown in Table II. Moreover, we see that a baseline that always predicts the mean $PM_{2.5}$ value per city has a relatively small error which is due to the fact that $PM_{2.5}$ levels in the studied cities are generally low and exhibit small variability.

C. Within-city predictions

Before evaluating spatial $PM_{2.5}$ prediction using our transfer learning approach, we first evaluate the predictability of $PM_{2.5}$

TABLE II
aRMSE OF BASELINE METHODS

	UK			US			Overall		
	6h	12h	24h	6h	12h	24h	6h	12h	24h
IDW	3.79	3.34	3.09	4.12	3.73	3.41	3.96	3.54	3.25
mean	7.00	6.64	6.36	4.60	4.26	4.02	5.80	5.46	5.19

TABLE III
aRMSE OF DIFFERENT TWITTER FEATURES

	#tw	#aqs	#high	all	BoW	BoW^{-1}	BoW^{-2}
6h	5.96	5.93	5.98	5.84	5.15	4.99	**4.97**
12h	6.17	5.98	6.02	5.77	4.96	**4.84**	5.16
24h	5.83	6.11	5.82	5.52	**4.65**	4.96	5.16

in each city using a model trained on Twitter and ground truth data of the city. As already discussed, this represents an unrealistic setup because ground truth data is not available for unmonitored cities. However, it is suitable for assessing the effectiveness of different Twitter features. In this set of experiments, data from each city is split based on time, using odd months for training and even months for testing.

Table III shows the results obtained using models trained on 'current' and lagged BoW features, as well as four simpler Twitter features that we extracted[7]: '#tw' (total number of tweets in each spatiotemporal bin), '#aqs' (number of tweets that provide information on current air quality), '#high' (number of tweets that refer to high air pollution levels) and 'all' (the concatenation of '#tw', '#aqs' and '#high'). We notice that for all temporal granularities, 'all' leads to better accuracy than '#tw', '#aqs' and '#high', suggesting that these features capture complementary information about current air quality. However, we see that the best performance for each temporal granularity is obtained by a BoW variant and, interestingly, we notice that for finer temporal granularities it is beneficial to use lagged BoW features (BoW^{-2} and BoW^{-1} for the 6- and the 12-hour temporal granularity, respectively). Based on these results, subsequent experiments employ the best performing BoW representation for each temporal granularity.

D. Cross-city predictions

We now turn into the main focus of our paper, i.e. spatial $PM_{2.5}$ prediction, and evaluate our transfer learning approach according to the setup described in section III-A. Table IV shows the results obtained when using full-dimensional BoW vectors ('full' column) as well as vectors where only the top-k most correlated features are kept, with (w=1) and without (w=0) sample weighting. First, we observe that the performance of full-dimensional BoW is considerably worse compared to the within-city setup. As expected, the absence of city-specific training data makes the learning task more difficult. With respect to the different transfer learning setups, we see that joint feature selection results in important performance gains in all temporal granularities, with the best results obtained when the top 50 or 100 features are used. Sample weighting, on the other hand, has a less pronounced but consistently positive effect.

Comparing the performance of our Twitter-based estimates with those of IDW, we notice that they do not perform on par. We believe that this result should be largely attributed to the fact that the studied cities exhibited very good air quality conditions for an overwhelming part of the studied period which makes it less likely for people to express their feelings about air quality on Twitter. Our findings match those reported in [10] where IDW was also found more accurate than the proposed approach under good air quality conditions.

TABLE IV
CROSS-CITY aRMSE WITH DIFFERENT TRANSFER LEARNING SETUPS

		full	k=10	k=20	k=50	k=100	k=200	k=500
w=0	6h	5.36	5.48	5.28	**5.21***	5.24	5.29	5.31
	12h	5.21	5.29	5.18	5.12	**5.09**	5.11	5.15
	24h	4.97	4.89	4.78	4.78	**4.75**	4.79	4.86
w=1	6h	5.35	5.47	5.27	**5.21***	5.24	5.29	5.30
	12h	5.21	5.26	5.18	5.11	**5.08***	5.11	5.16
	24h	4.95	4.85	4.77	4.76	**4.73***	4.77	4.84

Despite that, we notice that Twitter-based estimates carry considerable predictive power as they manage to obtain significantly lower error than the mean baseline. Motivated by that, we evaluated a late fusion scheme that combines our Twitter-based estimates with the IDW estimates by learning a meta-model that uses two features: a) IDW estimates for the training cities, b) Twitter-based estimates for the training cities (obtained through inner cross-validation). This model obtains an aRMSE of 4.15, 4.00 and 3.63 for the temporal granularities of 6, 12 and 24 hours respectively. Although its performance is still worse on average compared to IDW, it performs better than IDW in 3 out of 10 cities[8]: Boston, London and Pittsburgh. We notice that these cities are the most distant (on average) to the rest of the studied cities in each country, thus limiting the accuracy of spatial interpolation. This shows that exploiting Twitter information can be beneficial for improving air quality estimates even in cities with good average air quality conditions when they lie far from monitored cities.

IV. CONCLUSION AND FUTURE WORK

We presented a methodology for performing Twitter-based air quality estimations on cities that lack monitoring infrastructure. Our approach was found to provide fairly accurate estimates on a case study involving cities in the UK and the US. Although these estimates are less accurate than estimates obtained through spatial interpolation, we found that by combining the two types of estimates it is possible to improve accuracy in certain cities. In the future, we would like to extend our empirical study to additional air pollutants and to a larger and more diverse (in terms of population, country, air quality levels) set of cities. Moreover, we would like to experiment with more sophisticated textual representations (e.g. [24]) and transfer learning methods. Finally, it would be interesting to study whether better accuracy could be obtained by exploiting the image content of tweets using image-based air quality estimation approaches (e.g. [25]).

ACKNOWLEDGMENT

This work is partially funded by the European Commission under the contract number H2020-688363 hackAIR.

[7] '#aqs' and '#high' were obtained by applying specialized tweet classifiers whose details can be found in [21], section 4.3.

[8] Detailed results per city as well as datasets and code that replicates our experimental results are provided here: https://github.com/MKLab-ITI/twitter-aq

REFERENCES

[1] EEA, "Air quality in europe 2017," https://www.eea.europa.eu/publications/air-quality-in-europe-2017, accessed: 2018-02-06.

[2] Environmental Protection UK, "Healthy air where you live," https://www.healthyair.org.uk/documents/2013/02/healthy-air-community-campaign-pack-2012.pdf, accessed: 2018-02-06.

[3] Unicef, "Understanding and addressing the impact of air pollution on children's health in mongolia," https://www.unicef.org/environment/files/Understanding_and_addressing_the_impact_of_air_pollution.pdf, accessed: 2018-02-06.

[4] N. Castell, F. R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, and A. Bartonova, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?" *Environment International*, vol. 99, pp. 293 – 302, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0160412016309989

[5] H. Taheri Shahraiyni and S. Sodoudi, "Statistical modeling approaches for pm10 prediction in urban areas; a review of 21st-century studies," *Atmosphere*, vol. 7, no. 2, p. 15, 2016.

[6] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, and A. Jaimes, "Sensing trending topics in twitter," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1268–1282, 2013.

[7] V. Lampos, T. De Bie, and N. Cristianini, "Flu detector-tracking epidemics on twitter," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2010, pp. 599–602.

[8] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[9] J.-M. Xu, A. Bhargava, R. Nowak, and X. Zhu, "Socioscope: Spatio-temporal signal recovery from social media," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 644–659.

[10] S. Mei, H. Li, J. Fan, X. Zhu, and C. R. Dyer, "Inferring air pollution by sniffing social media," in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 2014, pp. 534–539.

[11] W. Jiang, Y. Wang, M.-H. Tsou, and X. Fu, "Using social media to detect outdoor air pollution and monitor air quality index (aqi): a geo-targeted spatiotemporal analysis framework with sina weibo (chinese twitter)," *PloS one*, vol. 10, no. 10, p. e0141185, 2015.

[12] S. Wang, M. J. Paul, and M. Dredze, "Social media as a sensor of air quality and public response in china," *Journal of medical Internet research*, vol. 17, no. 3, 2015.

[13] Z. Tao, A. Kokas, R. Zhang, D. S. Cohan, and D. Wallach, "Inferring atmospheric particulate matter concentrations from chinese social media data," *PloS one*, vol. 11, no. 9, p. e0161389, 2016.

[14] G. Kordopatis-Zilos, S. Papadopoulos, and I. Kompatsiaris, "Geotagging text content with language models and feature mining," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1971–1986, Oct 2017.

[15] M. L. Stein, *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

[16] D. G. Fox, "Judging air quality model performance," *Bulletin of the American Meteorological Society*, vol. 62, no. 5, pp. 599–609, 1981.

[17] G. Hoek, R. Beelen, K. De Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs, "A review of land-use regression models to assess spatial variation of outdoor air pollution," *Atmospheric environment*, vol. 42, no. 33, pp. 7561–7578, 2008.

[18] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[19] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[20] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose." in *ICWSM*, 2013.

[21] E. Spyromitros-Xioufis, S. Papadopoulos, A. Moumtzidou, S. Vrochidis, and Y. Kompatsiaris, "hackair deliverable d3.2: 2nd environmental node discovery indexing and data acquisition," Tech. Rep. D3.2, July 2017. [Online]. Available: https://www.researchgate.net/publication/324594192_hackAIR_deliverable_D32_2nd_environmental_node_discovery_indexing_and_data_acquisition

[22] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[23] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[25] E. Spyromitros-Xioufis, A. Moumtzidou, S. Papadopoulos, S. Vrochidis, Y. Kompatsiaris, A. K. Georgoulias, G. Alexandri, and K. Kourtidis, "Towards improved air quality monitoring using publicly available sky images," in *Multimedia Tools and Applications for Environmental and Biodiversity Informatics*, A. Joly, S. Vrochidis, K. Karatzas, A. Karppinen, and P. Bonnet, Eds. Springer, 2018, ch. 5.