Air Trails – Urban Air Quality Campaign Exploration Patterns

Martin Becker
University of Würzburg
Würburg, Germany
becker@informatik.uni-wuerzburg.de

Florian Lautenschlager
University of Würzburg
Würburg, Germany
lautenschlager@informatik.uni-wuerzburg.de

Andreas Hotho
University of Würzburg
Würburg, Germany
hotho@informatik.uni-wuerzburg.de

Abstract

Air pollution in urban areas has become a major issue and has attracted significant public attention. As a consequence, many citizens have started campaigns for measuring the air quality of their personal environment using mobile devices. In this study, we adapt HypTrails — a Bayesian method for comparing hypotheses about human trails — in order to investigate mobility patterns from such campaigns. In particular, we derive an approach to apply HypTrails to continuous, temporally dense navigation paths as is characteristic for GPS tracks. This allows us to directly study the behavioral processes of participants. We showcase our method on the citizen science campaign APIC (the AirProbe International Challenge) yielding promising results: We find differing mobility patterns of users in restricted and unrestricted environments, and extend previous work by showing that roads and road types play an important role explaining the observed paths. This gives first insights into movement patterns of urban air quality exploration. Ultimately, we believe that our approach can help to better interpret data collected in the context of participatory sensing campaigns, and to develop new theories about the motivational processes of volunteers.

1 INTRODUCTION

Air pollution in urban areas has become a major issue and has attracted significant public attention [5, 8]. As a consequence, many citizens have started campaigns for measuring the air quality of their environment [1, 7]. While most such initiatives are based on static measurement stations, some campaigns also use mobile sensorboxes that allow citizens to freely explore urban areas on a large scale. One of these campaigns was the *AirProbe International Challenge (APIC)* which was held as part of the EU project EveryAware[1]. APIC was a participatory sensing campaign mapping air quality in the form of black carbon measurements across four different cities. Due to its mobile nature, it provides the unique opportunity for studying human navigation behavior in the context of exploring air pollution in various urban environments.

In [10] Sirbu et al. have studied mobility patterns in the context of APIC focusing on aspects relevant to participatory sensing campaigns, i.e., activity and coverage (see Figure 1 for an illustration). They found that better spatial and temporal coverage is obtained when volunteers are assigned to specific mapping areas, compared to when no restrictions are imposed. Additionally, when allowed to measure freely, they (i) measure higher pollution levels, and (ii) exhibit differing exploration behavior.

In this study, we also investigate the mobility patterns from the APIC challenge. However, instead of focusing on aggregate statistics like activity and coverage, we directly study the behavioral processes of the participants. Specifically, we propose a novel approach to apply HypTrails [9] — a Bayesian method for comparing hypotheses about human trails — to continuous, temporally dense navigation paths as are characteristic for GPS tracks. This allows us to directly study the behavioral processes of participants in a hypothesis-driven way. We apply this method to the APIC data and confirm the differing mobility patterns of users in restricted and unrestricted campaigns, and show that roads and road types play an important role explaining the observed paths. This illustrates the applicability and versatility of our method while at the same time giving first insights into the exploration patterns of mobile urban air quality sensing.

Overall, we believe that our method paves the way to build more advanced user models, to better interpret data collected in the context of participatory sensing campaigns, and to develop new theories about the motivational processes of volunteers.

[1] http://everyaware.eu, accessed: December 2017

AGILE 2018 – Lund, June 12-15, 2018
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

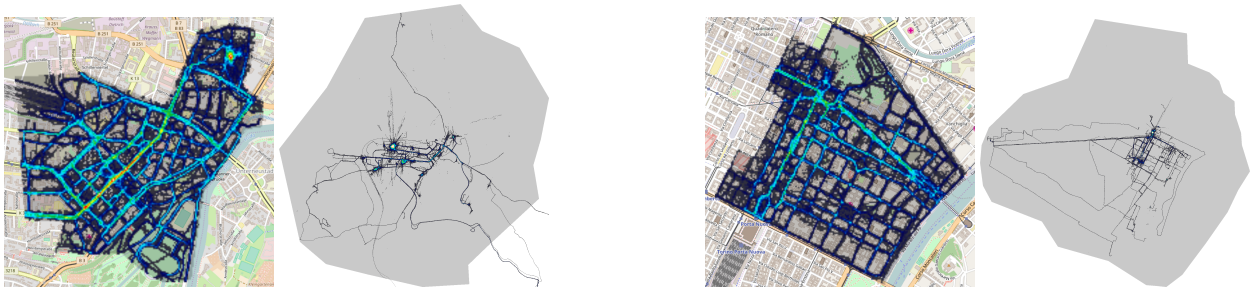
2 BACKGROUND AND DATA

2.1 HypTrails

HypTrails is a framework [9] that allows to formulate and compare hypotheses about human navigation behavior. Such hypotheses usually stem from theory, domain experts, previous experiments, or human intuition and can incorporate a large variety of background information. For example in this work, we investigate the idea that participants of a mobile air quality campaigns follow specific road types depending on the current campaign restrictions. Formally, hypotheses are encoded as transition probability matrices on a discrete state space within a first-order Markov chain framework. They are then compared based on the marginal likelihood $\Pr(D|H,\kappa)$ which represents the probability of the observed data D given a specific hypothesis H and a concentration factor κ. The concentration factor κ represents a measure of how sure the user is that a hypothesis is correct. To establish if one of the hypotheses is better than another we compare the marginal likelihood (the higher, the better) across a range of concentration factors (scaled by the number of states). For more details we refer to [2, 9].

2.2 Navigation data

In the following, we first introduce the APIC challenge from which the navigation data stems. Afterwards we describe how we generate trails from this data which we analyze in this work.



(a) Kassel **(b)** Turin

Figure 1: Urban air quality exploration patterns. These heatmaps represent spatial and temporal coverage during an air quality measurement campaign with mobile sensorboxes for Kassel and Turin (cf. [10]). For each city the left image represents measurements taken while users were restricted to measure in a specific area (phase 2) while the right image shows unrestricted exploration (phase 3). Red dots represent strongly covered areas.

The APIC challenge. The data we use in this work stems from the APIC challenge and is freely available[2]. The APIC challenge was aimed at studying the behavior and perceptions of citizens involved in monitoring air quality, during a large scale international test case. This was organized simultaneously in four cities: Antwerp (Belgium), Kassel (Germany), London (UK) and Turin (Italy). The campaign consisted of three phases, during which volunteer participants were asked to either take part in a web game (phase 1-3) for quantifying their perception on air pollution, or use a sensing device (sensorbox [4]) to measure and explore air pollution (black carbon (BC) concentrations) in their daily life (phase 2-3). In this study, we only focus on the actual measuring activities thus skipping information on the web game (for more information on APIC please see [10]): In phase 2 the measurements started in a predefined area for each of the cities. In phase 3 measurements were continued, however, *without restrictions* on the area to be mapped.

The APIC challenge has successfully involved 39 teams of volunteers across the four cities. Using the EveryAware platform [3], 6,615,409 valid geo-localized data points were gathered during the second and third phase of the challenge (the sensorbox collects one data point per second). Phase 2 was held from the 4th to the 17th of November 2013 and phase 3 took place from the 18th of November to the 1st of December.

Trails. In order to derive "clean" trails from these data points, we apply several pre-processing steps: We only keep measurements with a valid value for GPS accuracy and where the accuracy is better than 10m. Then we group the measurements by device id and sort them by their recording time to attain one trail for each sensorbox. In order to ensure correctly functioning sensorboxes which take one measurement per second, we split these trails whenever the time difference (> 2sec), the distance (> 100m), or the speed (> 50km/h) between two consecutive measurements is greater than a respective threshold.

Then, we generate a discrete state space — which is required to apply the HypTrails approach [9]. In particular, we employ a 200m by 200m grid based on the bounding boxes[3] listed in Table 1. Note that we leave out Antwerp and London because they do not provide enough data to derive decisive results (cf. Section 3.1). We map the points of each trail to the corresponding grid cells and then remove all self-transitions in order to focus on actual exploration rather than static processes. Afterwards we filter all trails which contain only a single entry.

Table 1: Bounding boxes used for discretizing city areas.

	min lon.	min lat.	max lon.	max lat.
Turino	7.6017	45.0080	7.7336	45.1326
Kassel	9.3454	51.2533	9.5650	51.3617

2.3 Road network

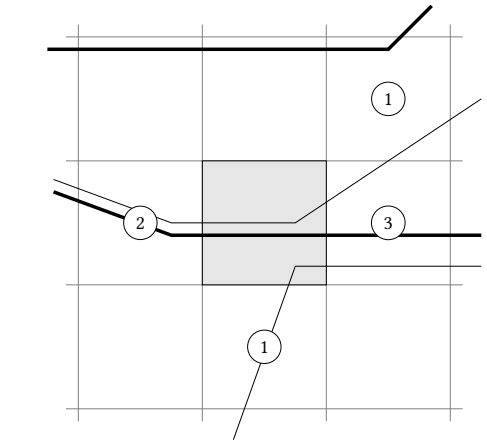
For the hypotheses in Section 3, we use the road network of each city. We extract these networks from OpenStreetMap[4] for each city separately. The corresponding data was downloaded from bbike.org[5]. Using this data we extract roads from the `osm_world_line` table and only retain entries where the field `name`, i.e., the name of the road (required for matching roads across cells as described in Section 3), and the field `highway`, which defines the type of the road, are not `null`.

3 HYPOTHESES

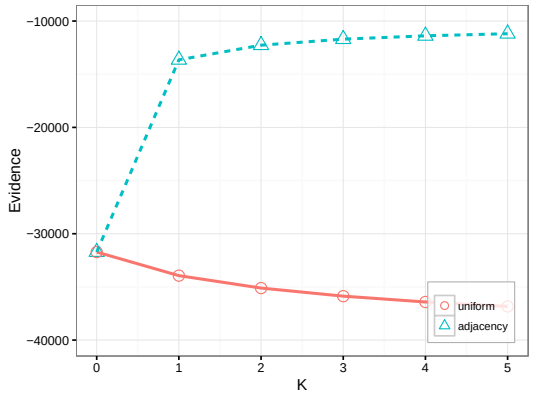
We formulate several hypotheses modeling different aspects of navigation processes. To this end, we define transition functions $\bar{P}$ which are normalized by source state s_i to form transition probability matrices as required by HypTrails.

The uniform hypothesis and adjacency. The uniform hypothesis $\bar{P}_{\text{uniform}}(s_j|s_i) = 1$ provides a "random" baseline every other (informed) hypothesis should be able to outperform. However, since our measurements are continuous (one sample per second), and we are using a state space with 200m by 200m grids, it is highly unlikely that transitions will occur to cells farther away than one cell. Consider for example the gray cell in Figure 2a as the current cell. Only the cells in its immediate vicinity are candidates to navigate to. Thus, we also define the adjacency hypothesis modeling this aspect. In particular we define $adj_i(j)$ to return 1 when the cell

[2]https://www.kde.cs.uni-kassel.de/everyaware/dumps/airprobe, accessed: March 2018
[3]These bounding boxes are based on the corresponding *woeid* ids on the *town* level. For example: https://www.flickr.com/places/info/725003
[4]https://www.openstreetmap.org/
[5]http://download.bbbike.org/osm/bbbike/, file date: 10.08.2017



(a) Grid over street network



(b) Comparison of baseline hypotheses

Figure 2: Hypotheses on AirProbe data. In (a) we illustrate how a road network (e.g. from OpenStreetMap) is used to derive hypotheses which we compare using the HypTrails method. It shows a geo-spatial grid over a road network. The bold lines are residential roads, and the thin lines are footways. The number in each cell represents the count of the roads which also pass the current (grey) cell, i.e., zeros are left out. In this scenario be build hypotheses about behavioral processes (cf. Section 3). For example, assuming a preference to follow roads, the higher the road count, the more likely it is for participants to move from the current cell to the cell with that number. Also, participants may prefer residential roads over footways which we also explore in our hypotheses. In this context, (b) additionally shows the performance of two baseline hypotheses illustrating that limiting transitions to adjacent cells (*adjacency*) is a more plausible baseline than a completely uniform hypothesis (*uniform*) in a scenario with continuous, temporally dense trails.

s_j is one of the adjacent cells of cell s_i (see the eight white cells in Figure 2a), and 0 otherwise. Then the adjacency hypothesis is defined as

$$\bar{P}_{\text{adjacency}}(s_j|s_i) = adj_i(j) \tag{1}$$

Road counts. We further hypothesize that users move according to the road network in a city. That is, given the current cell, we belief that the user will follow some road to one of the adjacent cells. We model this as follows: For each cell we extract the roads present in that cell. Then, given the names of the roads R_i of cell s_i, we count the number of roads $r_{i,j}$ of all the adjacent cells s_j which are also in cell s_i, i.e., $r_{i,j} = \sum_{x \in R_i \cap R_j} 1$. This represents the intuition that the more the roads between the source cell s_i and destination s_j overlap, the more likely a user will move to s_j. For an illustration, see Figure 2a: The cell below the top-right cell contains three roads also present in the grey source cell. Thus, a citizen will more likely go to that cell rather than to the top-right cell containing only one road also present in the (grey) source cell. We formally define the corresponding hypothesis as:

$$\bar{P}_{\text{roads}}(s_j|s_i) = r_{i,j} \cdot adj_i(j) \tag{2}$$

Footway and residential preference. With regard to phase 2 (users are restricted to a limited area) and phase 3 (no spatial restrictions) of APIC, we hypothesize that there is difference in navigation behavior as hinted at in [10]. To address these characteristic properties, we investigate whether the type of the road users prefer to follow changes between the different phases. In this case study, we specifically focus on *residential* roads, as mostly found in cities, and *footways*, which are exclusively reserved for pedestrians and bicycle drivers.[6] Note however, that footways often are found alongside roads, including major roads.

To model a preference for a specific road type, we weigh the different roads individually. Starting with the *residential* category, let $residential(x)$ be 2 if x is a road of the category *residential* and 1 otherwise. Then, we define the weighted sum $w_{i,j}^{\text{residential}} = \sum_{x \in R_i \cap R_j} residential(r)$ to represent the likelihood to move from cell s_i to s_j, where the residential roads are twice as important as all other road types. Consider, for example, Figure 2a where residential roads are bold and footways are thin. Using $w_{i,j}^{\text{residential}}$ instead of $r_{i,j}$, the weight of the cell below the top-right cell would be four instead of three. Formally, we define the hypothesis preferring residential roads as

$$\bar{P}_{\text{residential}}(s_j|s_i) = w_{i,j}^{\text{residential}} \cdot adj_i(j) \tag{3}$$

The hypothesis $\bar{P}_{\text{footway}}(s_j|s_i)$ is defined analogously.

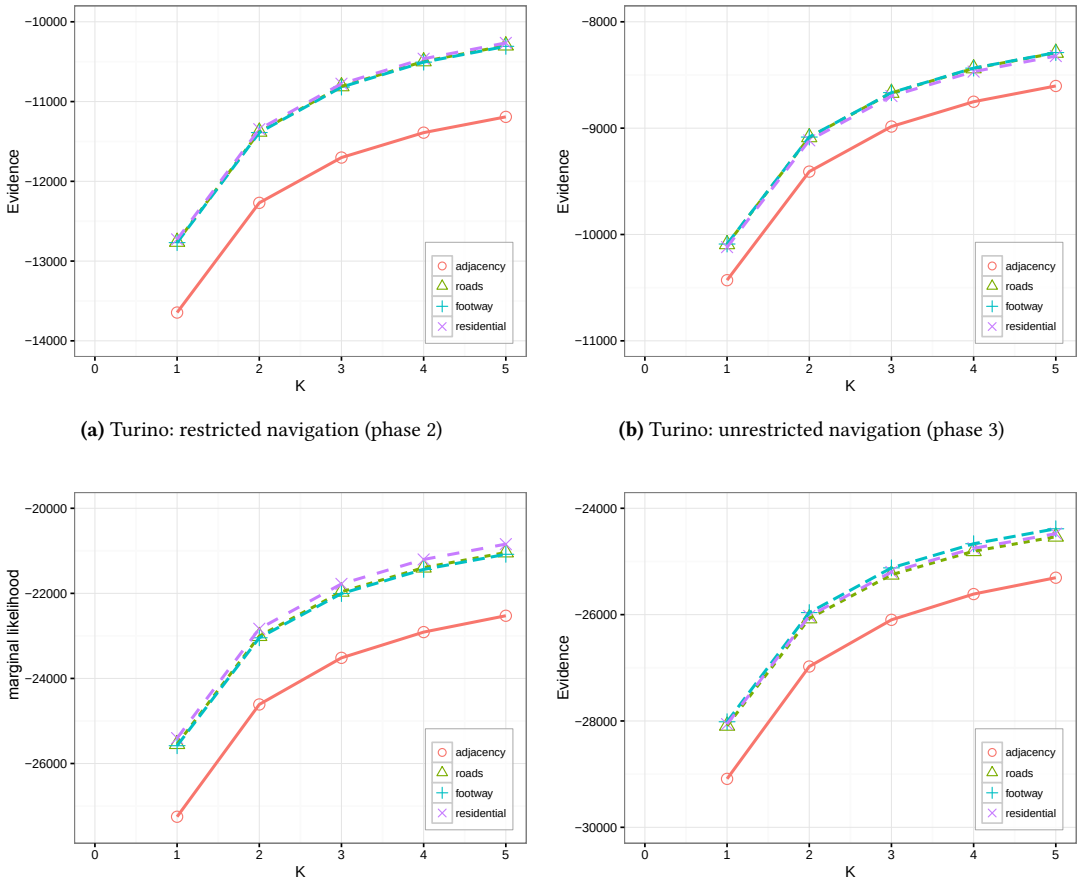
3.1 Results

In the following, we compare the hypotheses introduced in Section 3 on the data from phase 2 and phase 3 of the APIC challenge. Note, that we only report results on Turino and Kassel. For the other cities (Antwerp and London), the general tendencies are the same but due to the smaller amount of data available for these cities, the results are not decisive with regard to the interpretation table of Kass and Raftery [6]. In contrast, the interpretations we report in the following are all backed by decisive differences.

Baselines. We first compare the baselines — defined by $\bar{P}_{\text{uniform}}$ and $\bar{P}_{\text{adjacency}}$ — based on the data from phase 2 in Turin (see Figure 2b). As expected, we observe that — in a continuous setting — it is appropriate to restrict transitions to adjacent cells. That is, $\bar{P}_{\text{adjacency}}$ outperforms $\bar{P}_{\text{uniform}}$ by a large margin.

Roads. Next, we evaluate the performance of weighting the probability of a transitions to an adjacent cells by the number of common

[6] OpenStreetMap defines road categories *residential* and *footway* using the `highway` property. Also see: http://wiki.openstreetmap.org/wiki/Key:highway (accessed: 19.08.2017).



(a) Turino: restricted navigation (phase 2)

(b) Turino: unrestricted navigation (phase 3)

(c) Kassel: restricted navigation (phase 2)

(d) Kassel: unrestricted navigation (phase 3)

Figure 3: Comparison of navigation hypotheses on the APIC data. We compare several hypotheses about human navigation during the two phases of APIC. Generally, we observe that the hypothesis assuming that participants follow *roads* explains the data better than assuming random navigation (*adjacency*). We also find that refined information about street types (*residential* roads or *footways*) can improve on the unweighted roads hypothesis. Furthermore, we observe different preferences for one or the other road type dependent on the phase of the campaign and its objectives. This is in line with our findings in previous work [10] where we have observed different user behavior in the two phases (cf. Section 2.2).

roads with the source cell. The results are shown in Figure 3. On both cities and both phases this hypothesis ($\bar{P}_{\text{roads}}$) shows large improvements on the baseline $\bar{P}_{\text{adjacency}}$. This indicates the general tendency of users to navigate according to the properties of the underlying street network so that more "links" between cells correspond to more people moving to that cell.

Residential roads and footways. Finally, we study the preference for residential roads and footways. We first concentrate on Turino (Figures 3a and 3b). In phase 2, we observe a clear improvement of the hypothesis preferring residential roads compared to weighting all roads equally ($\bar{P}_{\text{roads}}$) or preferring footways ($\bar{P}_{\text{footway}}$). This corresponds to the focused exploration of down-town Turino as visualized in Figure 1. What is hardly visible, is that $\bar{P}_{\text{roads}}$ slightly (but decisively) outperforms the hypothesis preferring footways ($\bar{P}_{\text{footway}}$). In contrast, in phase 3, the preference of residential roads cannot explain the navigation behavior of the users as well as the previously outperformed hypotheses ($\bar{P}_{\text{roads}}, \bar{P}_{\text{footway}}$). Here, however, the footway hypothesis ($\bar{P}_{\text{footway}}$) slightly (but not decisively) outperforms the unweighted roads hypothesis ($\bar{P}_{\text{roads}}$). This shows, that indeed, the navigation behavior between the two phases differs significantly with regard to the preference of residential roads.

Similar results can be observed for Kassel. That is, we observe the same tendencies for phase 2 where the general trend to follow residential roads is stronger than for Turino. In phase 3, things are slightly different. In particular, both, the residential and footway hypotheses outperform the unweighted roads hypothesis. Nevertheless, as for Turino, the footway hypothesis explains the data better than the residential hypothesis.

Comparing the different cities and the different phases two trends are apparent:

(i) A preference for residential roads or footways can improve on the unweighted roads hypothesis.

(ii) Residential roads are preferred in phase 2 while footways explain the navigation behavior better in phase 3.

The former shows that road types generally carry information with regard to navigation preferences, and the latter indicates situational dependencies with regard to the overall goal and strategies of the users. This is in line with the findings in [10] where differing user behavior was observed in phase 2 and 3. One explanation for this may lie in the focus of each phase (cf. Section 2.2): In phase 2 users focused on the city centers trying to cover as much space as possible. Thus, they followed the most common streets in these areas, namely the residential roads. When they were allowed to measure where they wanted to, the focus on the city center decreased, thus reducing the navigation on residential roads. See Figure 1 for a comparison of the respective coverage. The good performance of footways in Kassel (whereas in Turin there where hardly significant differences) may be due to the fact at the users mostly measured air quality while commuting and inherently using large roads which often have an attached footway in Kassel. Further studies will be necessary to clarify the corresponding details. Such work may explore for example the preferences for primary, secondary, and tertiary roads instead of footways in the third phase.

4 CONCLUSION

In this work, we introduced a method to apply HypTrails for studying the underlying processes of exploration patterns in the context of mobile participatory sensing campaigns. In particular, we adapted HypTrails to temporally dense navigation paths in a continuous geo-spatial setting as is characteristic for GPS tracks. To illustrate our approach, we applied this novel method to data from the APIC challenge. The corresponding experiments yield promising results: We show that roads and road types play an important role explaining the observed paths. In addition, the results confirm a difference in navigational characteristics depending on the geo-spatial constraints defined by the APIC challenge as also found in [10].

Overall, this study provides a novel method for understanding behavioral processes in the context of geo-spatial navigation paths. In particular, our approach can be used to build user models, to better interpret data in the context of participatory sensing campaigns, and to develop and compare new theories about the motivational processes of volunteers.

For future work, it may be interesting to extend the transition models as applied in this work, for example, by further investigating the influence of the road network in the context of the data provided by OpenStreetMap, or by formulating heterogeneous hypotheses to explain the overall behavior of users during the APIC campaign [2]. Finally, by employing data from other participatory sensing campaigns as well as subjective information from users may provide the necessary background information to formulate and compare hypotheses that enable further insights into human navigation behavior as well as their incentives and goals in the context of environmental studies.

Acknowledgements This research has been supported by the DFG project “p2map”.

REFERENCES

[1] Martin Adam, Tamara Schikowski, Anne Elie Carsin, Yutong Cai, Benedicte Jacquemin, Margaux Sanchez, Andrea Vierkötter, Alessandro Marcon, Dirk Keidel, Dorothee Sugiri, et al. 2015. Adult lung function and long-term air pollution exposure. ESCAPE: a multicentre cohort study and meta-analysis. *European Respiratory Journal* 45, 1 (2015), 38–50.

[2] Martin Becker, Florian Lemmerich, Philipp Singer, Markus Strohmaier, and Andreas Hotho. 2017. MixedTrails: Bayesian hypothesis comparison on heterogeneous sequential data. *Data Mining and Knowledge Discovery* 31, 5 (Sept. 2017), 1359–1390.

[3] Martin Becker, Juergen Mueller, Andreas Hotho, and Gerd Stumme. 2013. A Generic Platform for Ubiquitous and Subjective Data. In *Conference on Pervasive and Ubiquitous Computing Adjunct Publication (UbiComp '13 Adjunct)*. New York, NY, USA, 1175–1182.

[4] Bart Elen, Jan Theunis, Stefano Ingarra, Andrea Molino, Joris Van, den Bossche, Matteo Reggente, and Vittorio Loreto. 2012. The EveryAware SensorBox: a tool for community-based air quality monitoring. In *Sensing a Changing World Workshop 2*. 1–7.

[5] Lijian Han, Weiqi Zhou, Weifeng Li, and Li Li. 2014. Impact of urbanization level on urban air quality: A case of fine particles (PM2. 5) in Chinese cities. *Environmental Pollution* 194 (2014), 163–170.

[6] Robert E. Kass and Adrian E. Raftery. 1995. Bayes Factors. *J. Amer. Statist. Assoc.* 90, 430 (1995), 773–795. arXiv:http://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1995.10476572

[7] Prashant Kumar, Lidia Morawska, Claudio Martani, George Biskos, Marina Neophytou, Silvana Di Sabatino, Margaret Bell, Leslie Norford, and Rex Britter. 2015. The rise of low-cost sensing for managing air pollution in cities. *Environment international* 75 (2015), 199–205.

[8] David Mage, Guntis Ozolins, Peter Peterson, Anthony Webster, Rudi Orthofer, Veerle Vandeweerd, and Michael Gwynne. 1996. Urban air pollution in megacities of the world. *Atmospheric Environment* 30, 5 (1996), 681–686.

[9] Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. 2015. HypTrails: A bayesian approach for comparing hypotheses about human trails on the web. In *International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1003–1013.

[10] Alina Sirbu, Martin Becker, Saverio Caminiti, Bernard De Baets, Bart Elen, Louise Francis, Pietro Gravino, Andreas Hotho, Stefano Ingarra, Vittorio Loreto, Andrea Molino, Juergen Mueller, Jan Peters, Ferdinando Ricchiuti, Fabio Saracino, Vito D. P. Servedio, Gerd Stumme, Jan Theunis, Francesca Tria, and Joris Van den Bossche. 2015. Participatory patterns in an international air quality monitoring initiative. *PLOS ONE* 10, 8 (08 2015), e0136763.