Project no. 265432

EveryAware

Enhance Environmental Awareness through Social Information Technologies

http://www.everyaware.eu

Seventh Framework Programme (FP7)

Future and Emerging Technologies of the Information Communication Technologies (ICT FET Open)

D4.2: Report on analysis of sensor and subjective data, and comparison of measured vs perceived environment

---

Period covered: from 01/09/2012 to 28/02/2014
Start date of project: March 1st, 2011
Due date of deliverable: Apr 30th, 2014
Distribution: Public

Date of preparation: 28/02/2014
Duration: 36 months
Actual submission date: Apr 30th, 2014
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: Fondazione ISI, Turin, Italy (ISI)
Lead contractor for this deliverable: Vlaamse Instelling voor Technologisch Onderzoek N.V. (VITO)

Executive Summary

Participatory sensing requires the availablity of easy accessible sensor platforms at low-cost. The EveryAware consortium developed sensing platforms for noise and air quality monitoring. Although the sensor platforms are not perfect -see results of D1.2 where significant research went into the testing of the sensors- test cases were developed for citizens to participate in noise monitoring and air quality monitoring. For noise monitoring, a high proportion of the measurements is made in an uncoordinated way, by users of the WideNoise App worldwide. In contrast, the air quality measurements have been made by volunteers participating to highly coordinated monitoring campaigns, basically due to hardware constraints and the preset objective to use participatory air quality data to discover spatial patterns in air pollution which is only feasible in a confined area given the relatively low number of sensor boxes and the high data coverage requirements.

Outline of the document

This document consists of four Chapters. The first Chapter investigates the air quality data that were collected in the Test Cases. The first main conclusion from these tests is that the AirProbe system is a reliable system over the entire monitoring and data communication chain. AirProbe really allows citizens to explore the air quality in their neighbourhoods, to make their own measurements, visualize and interprete them. Combining the sensing activities of the user community allowed to make a spatio-temporal analysis of the air quality in urban areas in Antwerp, London, Kassel and Turin. High resolution maps of BC concentrations were obtained from a two weeks long monitoring campaign and allowed to recognize spatial air quality patterns at street level. Nevertheless, the interpretation of these results prooved to be not straightforward. The data validation, which is a crucial step in any kind of environmental monitoring but particularly challanging in citizen science, is very complex using the current system and the proposed data validation steps resulted in a significant reduction of data volumes.

Chapter 2 is much in line with the results from the noise monitoring presented in D4.1, this time coverring the entire project period. A total of 48406 noise records were made worldwide during this period, which is equivalent to 78.74h of noise monitoring. The importance of coupling subjective information (perception) with the measured data cannot be understated, as it is the former that permits a judgment to be made as to how the measured data is perceived, turning the number from sound to noise, and understanding the perception of noise by communities is in turn vital to policy and decision makers.

Chapter 3 focuses on the subjective data, mainly under the form of **tags**, collected with the applications developed by the Project: WideNoise Plus for noise data and AirProbe for air quality data. The two applications allow users to record data samples and to annotate them with perceptions and tags. The aim here is to analyze the different resources provided by WideNoise Plus and to use the underlying annotated information to build a recommendation framework. For this, we evaluate multiple tag recommendation methods to improve the sensor data collection.

Chapter 4 deals with the analysis of the subjective data obtained by people using our smartphone apps and our web platforms. In particular, the subjective data consist of noise level predictions,

sliders and tags in the case of the Widenoise app, and the strategies adopted by users participating the APIC game/experiment hosted by the XTribe platform. The analysis of the APIC experiment will be covered in the next section due to the intimate relation between users air pollution perception and real measurements. An overestimation of pollutant concentrations in phase one is detectable for all cities. Players located the pollution mainly on main roads and crossroads, while gardens and rivers where perceived as cleaner. In phase three, i.e. as soon as the AirSquare values are made available, they changed opinion substantially. This clearly denotes that they were prone to change their mind.

Dissemination of the results

The results from the noise and air quality monitoring are dissiminated through the project website `http://cs.everyaware.eu/event/overview` where the user community and the general public have free access to the data and summary statistics from the monitoring campaigns.

Contents

List of Figures

Chapter 1

Overview

The analysis of the collected data by WideNoise and AirProbe is reported here. The methodology for data collection is based on the results described in report D4.1 Data Coverage and Interpolation Methods. For WideNoise, a high proportion of the measurements is made in an uncoordinated way, by users of the WideNoise App worldwide. In contrast, the AirProbe measurements have been made by volunteers participating to highly coordinated monitoring campaigns, basically due to hardware constraints and the preset objective to use participatory air quality data to discover spatial patterns in air pollution which is only feasible in a confined area given the relatively low number of sensor boxes and the high data coverage requirements. The distinction between WideNoise and AirProbe measurements has to be taken into account when analysing the data:

	WideNoise	AirProbe
Geographical area	worldwide	confined urban areas
Duration	years	weeks
Coordination level	low	very high
Spatial coverage	low	high within the monitoring area
Temporal coverage	low	high within the monitoring area

This report includes (i) the analysis of sensor data, (ii) the analysis of subjective data and (iii) perceived versus measured environment. The analysis of sensor data includes various information extraction methods for the interpretation of validated sensor data. Because one of the strengths of coordinated mobile monitoring approach as applied in AirProbe is the increased spatial density of measurements, special attention is given to the spatial analysis of the sensor data for discovering spatial patterns (e. g., differences between streets) in air quality. The WideNoise measurements are snap-shots of a highly volatile sound environment at a given location in space and time, and are therefore less suited to perform spatial analysis and comparisons between locations. On the other hand, WideNoise data are more likely made at specific events (i. e., a point measurement at noisy event) and probably more easily tagged or annotated than the AirProbe data. The analysis of subjective data and the comparison of the perceived versus measured environment is therefore based on larger data quantities and probably more representable for the WideNoise data than for the AirProbe data.

Chapter 2

Analysis of air quality sensor data

This chapter provides an overview of the analyzes that have been performed on the sensor data that were collected during the EveryAware project for the AirProbe Intl. Challenge for air quality data. In the Airprobe Intl. Challenge, volunteers from 4 cities, Antwerp, London, Kassel and Turin, conducted air quality measurements with the sensor box for a period of 4 weeks (see D3.2). Two weeks before the start of the monitoring, all the sensor boxes were gathered at specific locations in each city, together withreference devices to perform simultaneous measurements that are used for model calibration (see D1.2). The city-specific models were used to model black carbon concentrations from the sensor measurements. The analyzes of the air quality data is performed for the different cities separately. It is important to stress that the analysis of the air quality data is part of the experimental testing set-up for the use and validation of the integrated EveryAware platform. The air quality maps shown in this report should not be used as validated black carbon maps.

2.1 Data validation

2.1.1 Motivation

Sensor box specific models were used to estimate black carbon concentrations from the sensor measurements (see D1.2). Additionally in Antwerp the volunteers used micro-aethalometers for direct black carbon measurement at a 1-sec resolution. These measurements serve as benchmark measurements for comparison with the estimated black carbon values. The correlation between the measured and modelled black carbon concentration was calculated for the data series of each team seperately. Correlations ranged between -0.17 (team 7) and 0.62 (team 1). Timeseries are plotted in Fig. 2.1. For some of the teams (teams 1, 3, 4, 5, 6, 8 and 10) the correlations are moderate, for others (teams 2, 7 and 9) the modelled black carbon series is completely different from the measured black carbon series.

The difference between modelled and measured BC concentrations could be caused by measurement errors (e. g., sensor failure or inapropriate way of making measurements) or modelling errors. The sensor values of all the three teams where model estimations and BC measurements differed largely show extensive periods with extremely low sensor variability. These periods occur both in indoor as in outdoor environments (Fig. 2.2).

It is clear from the examples above that data validation is a critical step in the interpretation process of the data. Given the high numbers of data and the fact that the data collection itself is largely unsupervized and uncoordinated, it is difficult to develop data validation algorithms that are generally applicable. The validity of each data point that is enterred into the data base is to be checked in a data validation process. Erroneuos measurements that should be identyfied by the validation process are, for example, the measurements that are made during the heating-up of the sensors. Before the start of any data collection, volunteers were asked to switch the sensor

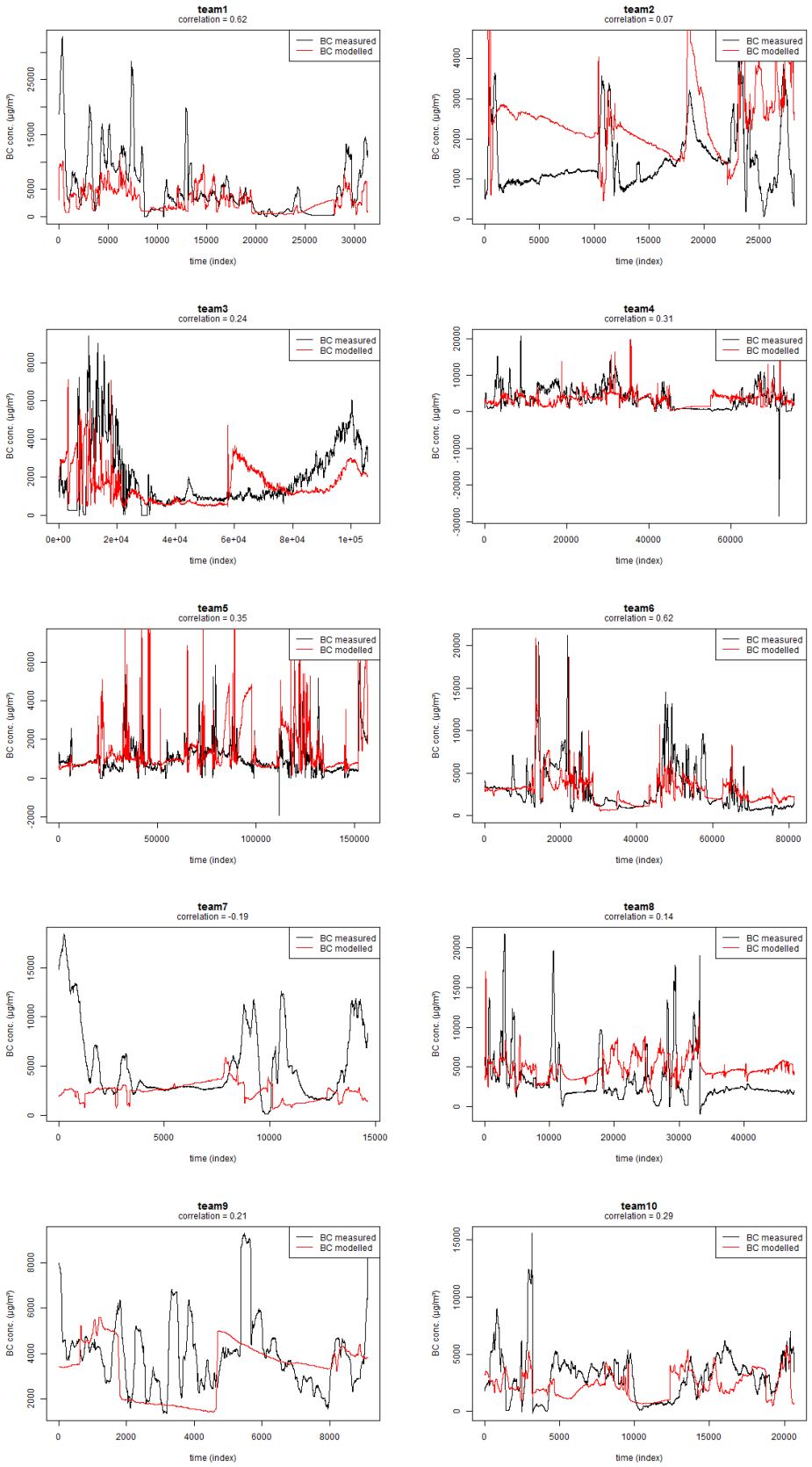


Figure 2.1: Measerud vs. modelled black carbon for the data collected by the teams in Antwerp.

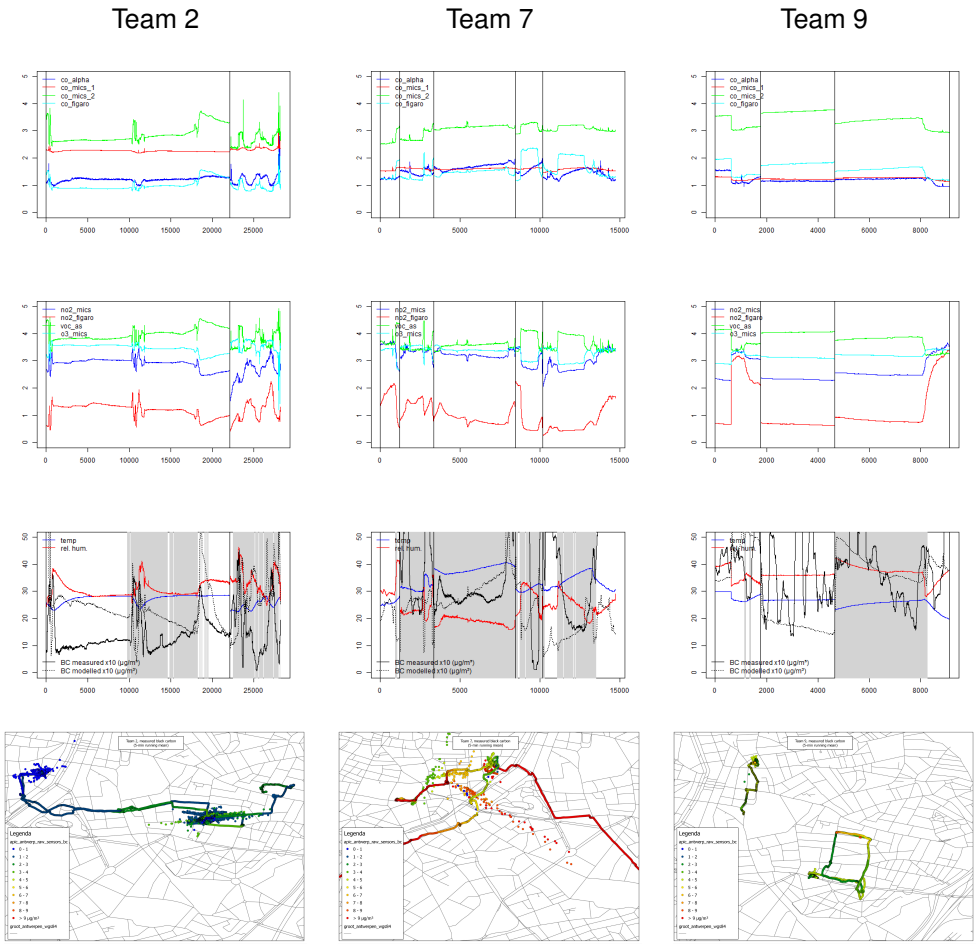


Figure 2.2: Overview of the sensor measurements for teams where measured and modelled BC concentrations differed substantially (teams 2, 7 and 9). The different days of monitoring are indicated by the black vertical lines (note: seperate runs per day are possible), the grey shaded area indicates period with reliable GPS connection.

box on for a period (minimal 30 minutes, preferably longer). The data from this heating-up period have to be discarded form the outdoor air quality analysis. In addition, data that are collected without respecting the heating-up period of sensors typically show the combined effects of fluctuating gas concentrations and a more general increasing or decreasing trend which is due to the sensor heating. These data should be identyfied and treathed differently in the final analysis. Also failing sensors or sensors that go outside their output signal range inevitably lead to a wrong estimation of the black carbon concentration. Failing sensors should be identified and deleted from further analysis. The same holds for failing GPS data. The GPS (like any other measurement device) is not perfect. Occassional fluctuations which are observed in urban environment will lead to a wrong positioning of the air quality measurement, therefore deteriorating the spatial analysis. A validation of the GPS data should be included to decrease these geo-reference errors. The collected data is a mixture of indoor and outdoor measurements (i. e., participants keep on measuring when they enter a shop or tram). These data should not be mixed, i. e., indoor data should be exluded as much as possible from the analysis of outdoor air quality. Finally, the way that the measurements are performed is not controlled. An explaination about a good measurement practice was given to the participants, but it is unknown whether these guidelines were strickly followed all the time. Blockage of the free air entrance by clothes, for example, is a potential source of measurement error.

The data validation process applied in this report consist of the following steps:

1. the identification of errors in the measurement time;
2. the identification of sensor measurements during heating-up;
3. the identification of errors in geo-location GPS values;
4. the identification of failing gas sensors;
5. the identification of errors in the estimated black carbon estimation;
6. the identification of indoor measurements.

2.1.2 Data validation

A data validation process based on the sensor box measurements was carried out. The number of validated measurements is much lower than the original amount of measurements. At the same time, the data quality is increased by the removal of erroneous or uncertain measurements. An overview of the data validation process is given below.

The identification of errors in the measurement time

The identification of incorrect measurement timestamps (i. e., the time when a measurement is made) is done during data processing at the server side. The AirProbe applications sends two timestamps: one from the sensor box (derived from the GPS-signal) and one from the smartphone. The smartphone's timestamp is used as replacement in case the timestamp from the sensor box is missing or invalid (e.g., "1970-01-01") if it is valid by itself. Otherwise, a default timestamp is used to indicate the absence of a useful timestamp.

The identification of sensor measurements during heating-up

After switching on the sensor box, the gas sensors are heated to their working temperature (typically 200 - 250 °C). During this period, which can be as long as 1 to 1.5 hours, the temperature

within the sensor box increases as well (see Fig.). Therefore the sensor box temperature is used as a proxy to screen measurement series for heating-up periods. The variability of the measured temperature within the box is used for the detection of heating-up events. Due to potential high-frequency noise on the temperature signal, a 5 minute average temperature series is calculated ($\mathbf{T_5} = \{T_{5,t1}, \ldots, T_{5,tn}\}$, where $\mathbf{T_5}$ is a temperature data series of 5 minute averaged temperature measurements, and $T_{5,ti}$ the 5 minute average temperature of the i^{th} 5 minute long period of temperature measurements). The difference between the elements of this temperature series is then calculated ($T_{5,t2} - T_{5,t1}$) and positive differences are substituted by "+1", and negative differences by "-1". The running length of this series is determined, and searched for long (longer than 30 minutes, i. e., 6 elements) periods of positive temperature differences at the beginning of the series. These are periods of steady temperature increase such as observed during the heating-up period. The measurements that are made within the heating-up period are flagged.

Potential errors of this identification procedure could occur when the environmental temperature is close to the temperature within the sensor box before and after heating-up of the sensors. In this situation there would not be an increasing temperature in the sensor box, and the heating-up event would not be detected by this methodology. Additionally, some sensors would need less time for heating-up than others. Using the temperature does not allow for identification on the individual gas sensor level. The fact that the black carbon concentration is modelled from the measurements of all the gas sensor (allbeit with different importance) justifies the use of this integrated methodology based on sensor box temperature. The number of measurements that were identified as measurements during sensor box heating-up are given in Table 2.1 and acoount for 19–27% of the total data volume per city.

The identification of out-of-range gas sensors

Sensor values are checked to be within the sensor output range. Records with sensor values lower than 0.05 or higher than 4.95 are flagged as out-of-range records per sensor (maximal read-out range of [0, 5]). The first days of the calibration period prior to the actual test cases were dedicated to check the sensor values for out-of-range events. At that time, the measurement range could be manually adapted to fall within the measurement interval for the individual sensors. Especially the NOx and NO_2 sensors went out-of-range for some sensor boxes, most other sensors did not show this behaviour. The calibration models are constructed on valid sensor box measurements, i. e., all the sensors are within the measurement range, except for e2v MiCS-2710 NO_2 sensors which still frequently went out-of-range. The error on the black carbon estimation by these models is higher when one or several sensors are out-of-range. Of course, sensor importance in the calibration model is an additional important parameter in this sense. Sensors that do not substantially contribute in the model estimation may fall out of range without significant effects. Currently, we did not take sensor importance into account, and used generic threshold values for each sensor of each sensor box. The number of measurements that went out-of-range during the APIC cases are given in Table 2.1. Most of the sensors are out-of-range in at rare occasions (between 0.05 and 6%), but for the e2v MiCS-2710 NO_2 sensor the number of out-of-range events is much larger (35–88%). The drift of the e2v MiCS-2710 NO_2 sensor may explain its low importance in the calibration model (see D1.2).

Identification of failing sensors

Failing sensors often give stable sensor signals when other sensors are fluctuating. The identification of failing sensors is based on the standard deviation of the sensor signals. When the standard deviation of the sensor signals is below a pre-defined threshold, the according records are flagged. The analysis is performed using the standard deviation over a 5 minute data window. The standard deviation threshold is determined as the standard deviation of a data series (300 elements) with

mean 1 and random noise of maximum 0.05 around the mean (=0.003, rounded to threshold value of 0.005). The duration in which the signal has a low variability provides extra information, and flags are coded as: "1" for failure during less than 5 minutes, "2" for periods between 5 and 30 minutes, and "3" for period longer than 30 minutes. A final field is calculated in which the number of failing gas sensors (out of 8) is given per data record.

This methodology of failing sensor detection contains errors. The threshold that is used is set rather subjectively. Increasing or decreasing the threshold could affect the number of records that are identified as from failing sensors. Sensor measurements under stable gas concentrations could be erroneuosly identified as sensor failure. In an urban outdoor environment, the variability is high enough to exceed the threshold, but indoor measurements could possibly lead to a false identification. There is also an overlap with the out-of-range identification. Sensors that go out-of-range will result in a very low standard deviation and will therefore be identified as failing sensors. Numbers of measurements with very low variability are given in Table 2.1.

Identification of errors in geo-location

Several geographical data sources are exploited by the EveryAware platform: location data from the sensor box, location data from the smartphone, location data from WLAN or IP address. In general, the accuracy of the geolocation data is higher when they are taken from sensor box or mobile phone GPS, and lower when the IP address or WLAN data. The difference in accuracy is used in the flagging of the geo location data ("1" for GPS from sensor box or smartphone, "2" for IP or WLAN data, and "3" when location data are lacking).

For data series with location data from WLAN, IP or without location data, an interpolation is performed when these records are preceeded or followed by accurate geo-location data. This could for example occur when the GPS signal is lost for some time. If this period is shorter than 1 minute, a linear interpolation is performed to estimate the geo-location of these records better. The interpolated records are flagged. A similar approach is used for records that make unrealistic jumps. Therefore the data is UTM-projected and the distance between sequential records is calculated. If this distance is unrealisticaly high (we used a threshold of 11 m, which is equivalent to the distance travelled per second at a travelling speed of 40 km/h), the geo-location is estimated by interpolation. This approach is justified by the fact that these jumps only last for a few seconds, the distance over which the interpolation is performed stays limited. The spatio-temporal measurement series over which interpolation of the geographical data is allowed is kept short (max. 1 minute). For longer period, the uncertainty on the estimations becomes too high, especially when linear interpolation of the GPS coordinates is applied whithout taking the street configuration into account. The amount of data with a GPS source other than the sensor box or smatphone GPS ranges between 12% for Antwerp, about 20% for Kassel and Turin, and 40% for Antwerp.

Identification of indoor measurements

Indoor measurements generally show limited variability in temperature and relative humidity signals. The running standard deviations over a window of one minute for temperature and relative humidity were used to identify indoor measurements. If the standard deviation of data subsets of 5 minutes stays below a threshold, these measurements are identyfied as indoor measurements. Additionally, records with a relative humidity below 20%, during heating-up without GPS, with a lot (6 or more) stable sensor signals or with a geo-location data source other than the GPS from sensor box or smartphone are also flagged as potential indoor records. The identification of indoor measurements contains a high degree of uncertainty. The variability threshold for the temperature and relative humidity sensors was set subjectively (as the standard deviation of a 1 minute long data series that was constructed by random sampling of values between 19.995 and 20.05 (steps

of 0.001) with replacement). As indicated in D1.2 teh sensor box could potentially be used to estimate indoor conditions. In this analysis, however, the focus is on the outdoor environment and indoor measurements are eliminated as much as possible.

Table 2.1: Number of measurements affected by the different data validation steps.

	Antwerp	Kassel	Turin	London
Original nr*	283.000	3.200.000	2.018.000	1.314.000
Heating-up	55.000	866.000	383.000	276.000
Out-of-range:				
$\rightarrow$e2v MiCS-2710 NO_2	250.000	1.950.000	1.500.000	465.000
$\rightarrow$ other sensors	19.000	160.000	72.000	16.000
Failing sensors :				
$\rightarrow$CO alpha	20.000	196.000	86.000	46.000
$\rightarrow O_3$	53.000	385.000	116.000	32.000
$\rightarrow$VOC	25.000	429.000	92.000	28.000
Indirect GPS	34.000	614.000	423.000	538.000
Indoor measurements	94.000	1.040.000	333.000	146.000

* Number of measurements (approximation) based on time and location.

Data validation process

The data validation was performed iteratively over the different sessions, where a session is defined as a collection of measurements that are conducted by a unique session of the App (either a new session or a recovered old session). The data within one session are not necessarily sequential in the sense that different data collections from different days could be clustered within one session. The sessions are sensor box specific, so there is no mixing of data from different sensor boxes within one session.

High proportions of data have been identified as heating-up measurements, out-of-range measurements (especially for e2v MiCS-2710 NO_2 sensor) or measurements with unprecise positioning (GPS source other that sensorbox or smartphone GPS). For the analyses presented in the following sections, the data validation was used to reduce the original datasets of each city, applying these rules:

1. measurements are not taken in the heating-up period of the sensors;
2. the sensor values are within the measurement range;
3. the CO alphasense, O_3 and VOC sensor show a variability higher that the sensor noise level;
4. the measurements are made outdoors;
5. the GPS data are from the sensorbox or smartphone GPS.

Finally, data record where the black carbon estimation was outside the reasonable range of 0 to 150 μg/m^3 for an urban environment were excluded from further analyses.

The number of validated outdoor data per city is a large reduction of the original number of measurements. The validated data sets contain 31%, 38%, 53% and 45% of the original data for Antwerp, Kassel, Turin and London, respectively. The histograms of the estimated black carbon concentrations for the original datasets versus the validated datasets are given in Fig. 2.3. The main difference is a shift toward higher black carbon concentrations for the validated data. The bimodal pattern that is observed in Antwerp, is not seen at other cities. For Kassel, the distribution

of black carbon concentrations is quite uniform compared to other cities, whereas for Turin and London oposing patterns are observed. In Turin most of the measurements are below 10 μg/m^3, in London most values are greater than 15 μg/m^3. Averaged black carbon concentrations are 10.8, 12.7, 8.8 and 16.1 μg/m^3 for Anwterp, Kassel, Turin and London, respectively. There is a significant difference in the measured black carbon concentration at the four cities (Wilcoxon test, $P < 0.01$).

2.2 Data interpretation and visualization

2.2.1 Temporal analysis of the measurements

The temporal analysis of the measurements is conducted similarly for the four cities. Following analyses are performed: (i) a day-by-day analysis of the black carbon concentration, and (ii) an assessment of the black carbon concentration at different hours of the day. Given the nature of the data –which are spatially and temporally explicit– it is difficult to neglect the spatial component and focus on the temporal component exclusively. A (substantial) part of the variability in the temporal analysis is caused by spatial effects (e. g., measurements at busy and low traffic streets). In the temporal analysis we assume that citizens made measurements at a comparable mixture of streets and other locations, so that the variability in the measurements caused by spatial effects is similar from day to day and for the different hours of the day.

A large day-to-day variation in black carbon concentration is observed at all four cities (Fig. 2.4). For London, black carbon concentrations are generally high, and here the day-to-day variation is smaller, but still significant between certain days. The range of the boxes (interquartile ranges) also differ substantially between days. For some days, the interquertile range is only 1–2 μg/m^3, for other days this range can be 10 μg/m^3 or more.

For the assessment of the variability in black carbon concentration at different hours of the day, the estimated black carbon concentrations were first normalized for the variability between days by dividing black carbon values by the daily averaged values. The typical bimodal pattern with elevated concentrations during morning and evening rush hours is slightly visible in Kassel and Turin (Fig 2.5). For Antwerp the black carbon estimations between 8 and 9 am are significantly lower than for later hours. The highest values are observed at 13 and 18-19 pm. In London, the lowest values are observed between 8-9 am, i. e., during the timeslot when the highest values are expected based on traffic intensity.

2.2.2 Spatial analysis of the measurements

The spatial analysis of the sensor box measurements form the Airprobe International Challenge is performed at different levels of detail. First, an analysis is made at the street level, where the estimated black carbon levels are compared between different streets in the study area at the four cities. The calculation steps involved in the attribution of a measurement point to a street are: (i) a projection of the measurements to the closest street in an area of approximately 100 m around the measurement point. The street data layer is an open street map layer from the four cities, and (ii) the street names are added as an extra column to the validated data sets. Some data points could not be attributed to a street because they were to far from the closest street. These records were withheld from the spatial analysis. Consequently datasets were further reduced in size, the datasets contained approximately 50.000, 400.000, 305.000 and 320.000 elements for Antwerp, Kassel, Turin and London, respectively. The streetwise analysis includes the 25 streets or squares where the number of measurements were the highest.

Differences in the average black carbon concentration between streets were observed at the four cities. In Anwterp, the highest concentrations were observed in a residential area with moderate

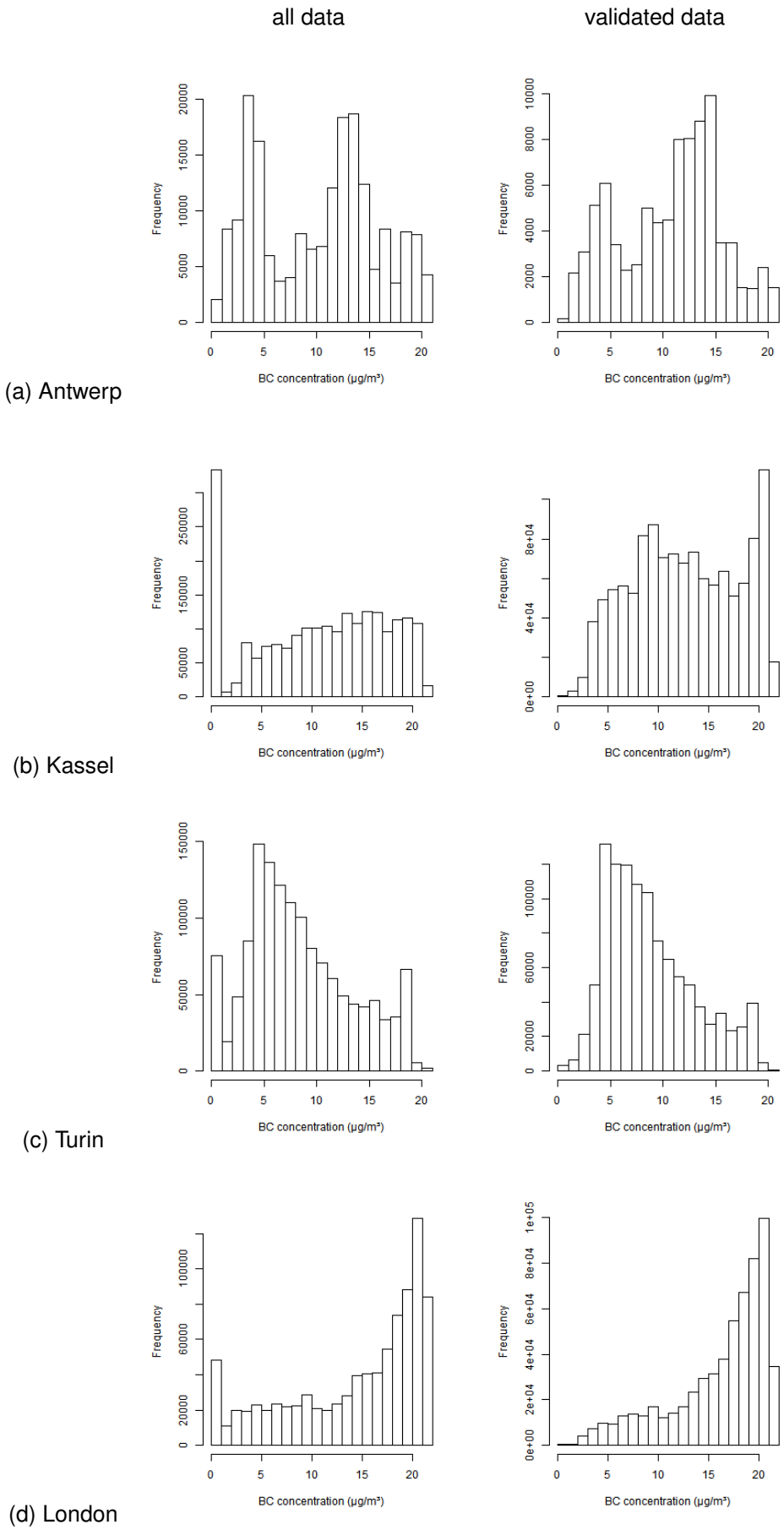


Figure 2.3: Histograms of estimated black carbon concentrations, based on all the measurement and the validated measurements per city.

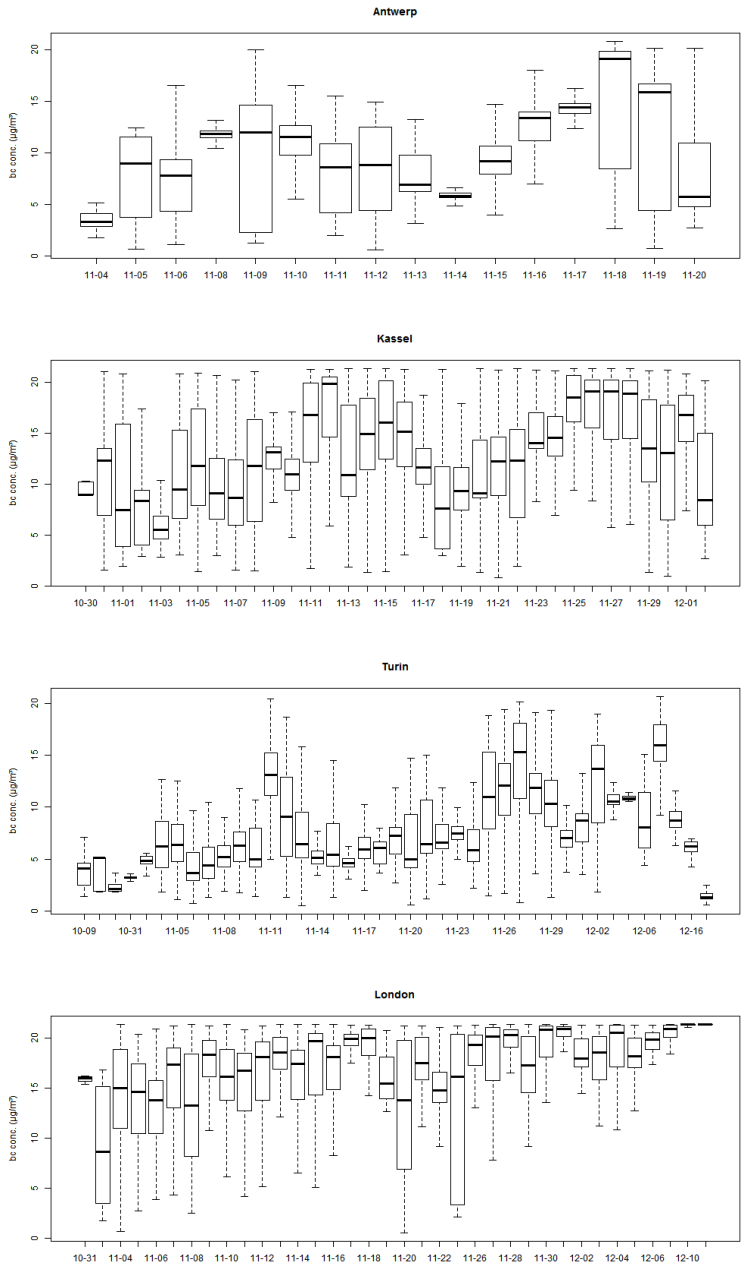


Figure 2.4: Day-to-day boxplots of the black carbon concentration at the four different cities during the Airprobe International Challenge.

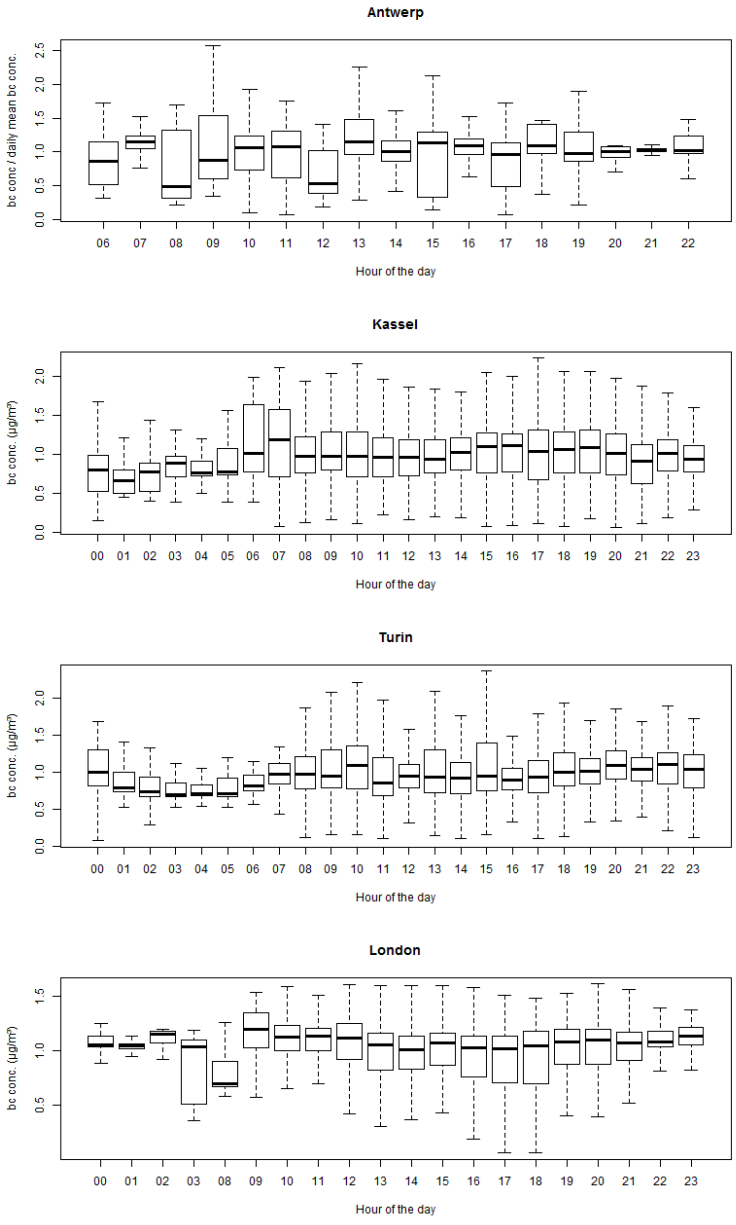


Figure 2.5: Black carbon concentration in function of the hour of the day at the four different cities during the Airprobe International Challenge.

traffic (Ramstraat, Grote Beerstraat, etc.). The estimated black carbon concentration in this area is considerably higher than at the main roads crossing the area (Plantin en Moretuslei and Turnhoutsebaan) (Fig. 2.6). In contrast, the number of extreme peak concentrations is highest at roads with the highest traffic intensity (Plantin en Moretuslei and Turnhoutsebaan) in the area. The highest peak events are observed in Dolfijnstraat and Tweelingenstraat, both are residential streets with moderate traffic.

In Kassel the range of street averaged black carbon concentration is lower than in Antwerp. The highest concentrations were observed in the “Akazienweg” and the “Kurt-Schumacher-Straße”. The lowest concentration is found at “An der Ahna”. Exposure to peak events of >20μg/m^3 is most frequently observed at “Georg-Forster-Straße”, “Wilhelmshöher Allee”, and “Kurt-Schumacher-Straße”. There was also construction work during the challenge in the “Fünffensterstraße” that caused high concentration values.

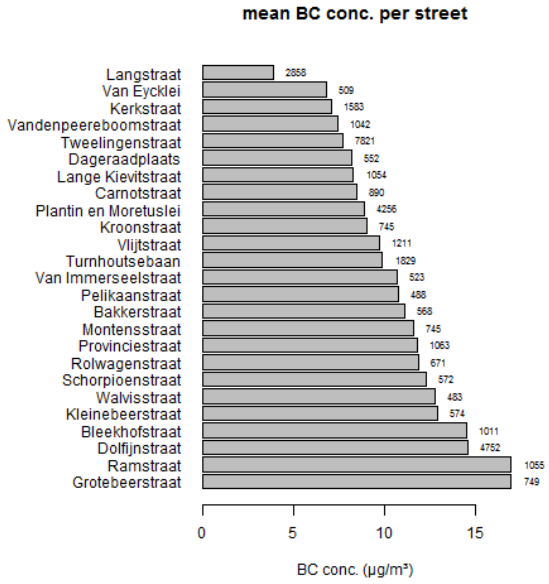
In Turin, the average black carbon concentration was the highest in Viale Primo Maggio, where the number of peak concentrations was also very high. This may be due its position at the intersection of two main arteries of the city, corso San Maurizio and Regina Margherita, which were, however, outside the mapping area. Via Accademia Albertina, a street reserved for bus transit, showed highest number of peaks. Piazza Carlo Felice, on the other hand, is a large square with a green area and no traffic, showing lowest number of peaks and a low average BC concentration.

In London, the black carbon estimates were generally low. Indeed, the variation in black carbon concentration of 25 streets with the highest measurement density was very limited. Most peak events are encountered at the Barbican Highwalks, which are walkways at levels relatively high above the road. Looking at the average map, however, in the London case, there was only one area of high BC value detected (around 8.5-10 on the scale) - close to a small section of Farrindgon Road, which is a main road passing through the edge of the area. GPS errors in the data mean that the high value cannot definitively be attributed to this road. Two additional areas of medium value (between 2 and 5.5 on the scale can be seen around Commercial Street and Whitechapel Road, again major roads passing through the area. Interestingly, average readings of about 4 can be found in the Thomas Moore Residential Gardens, which is perhaps an area where lower readings would be expected.

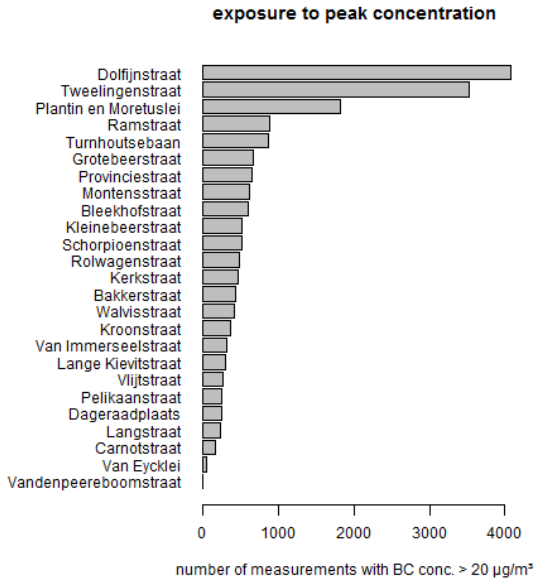
A data aggregation to fixed point within streets (at distances of 20 m) was performed using a Gaussian smoothing function. All the measurements within 30 m from the predefined fixed points were weighted and averaged to a single value that is attributed to the fixed point. Locations with less than 5 measurements were withheld from this analysis. The maps with estimated black carbon concentrations for each of the four cities is given in Fig. 2.10. In Antwerp, the highest concentrations were observed in a residential neighbourhood (zone A in Fig. 2.10) in between the two main entrance roads to the area (B in Fig. 2.10). Given the higher traffic intensity at the main entrance roads this result was unexpected. At a low traffic area with a traffic free square and calm surrounding streets (C in Fig. 2.10) at the South side of the main entrance road Plantin en Moretuslei (lower B), estimated black carbon concentrations are low. Some busy crossroads and street canyons (D in Fig. 2.10) had about the lowest black carbon concentration in the area.

For Turin, the limited traffic zone is clearly visible on the map in Figure 2.11 (zone A), where lower BC values are obtained. Two pedestrian streets, however, show high pollution levels (marked B in the figure). This could be due to frequent car queues at the intersections of these streets with those perpendicular, where pedestrian traffic significantly slows down the cars. The area around the train station, usually busy with busses and other traffic also display high BC values (area C). Standing out are also Corso Regina Margherita (D) and corso Re Umberto (E) with high pollution levels and the streets in the vicinity of parks (area F) for low pollution levels.

For Kassel, the pedestrian streets around “Obere Königsstraße” with lower BC values are clearly visible in Figure 2.10 (b) (zone A). The street “Auedamm” around the city park “Karlsaue” has also quite good pollution values (zone B). Areas with high traffic volume [“Wilhelmshöher Allee” (C),

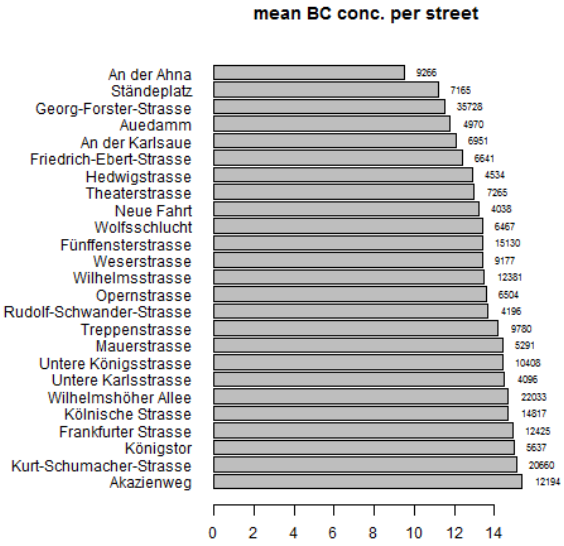


(a)

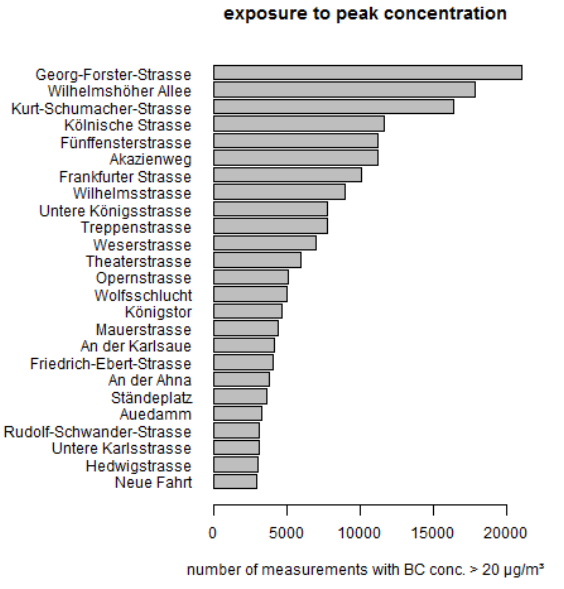


(b)

Figure 2.6: Average black carbon concentration per street in Antwerp (a) and the number of peak events (black carbon concentration >20μg/m^3) at these streets (b).

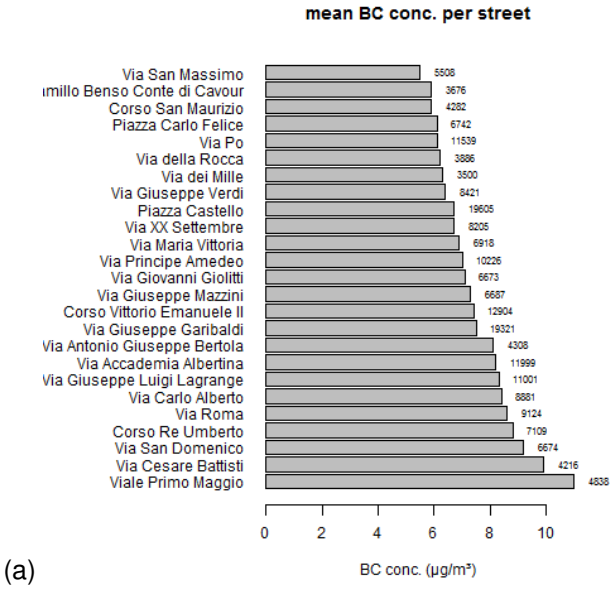


(a)

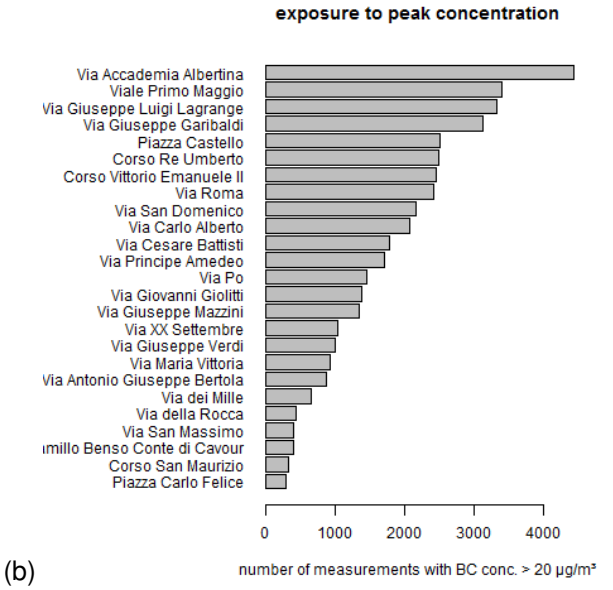


(b)

Figure 2.7: Average black carbon concentration per street in Kassel (a) and the number of peak events (black carbon concentration >20μg/m^3) at these streets (b).

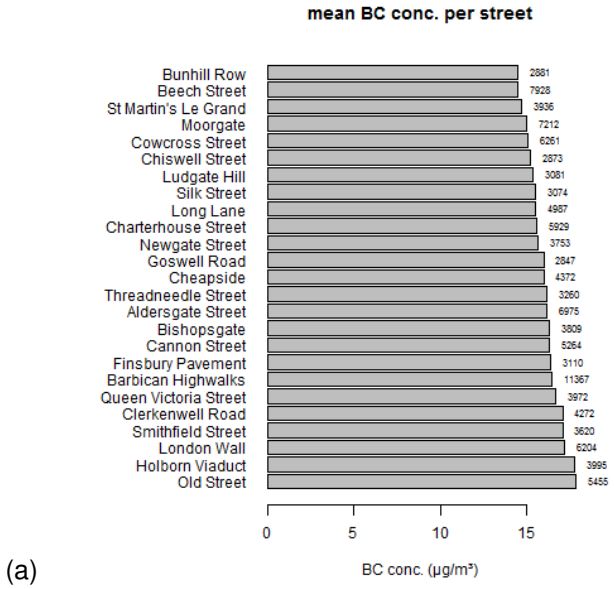


(a)

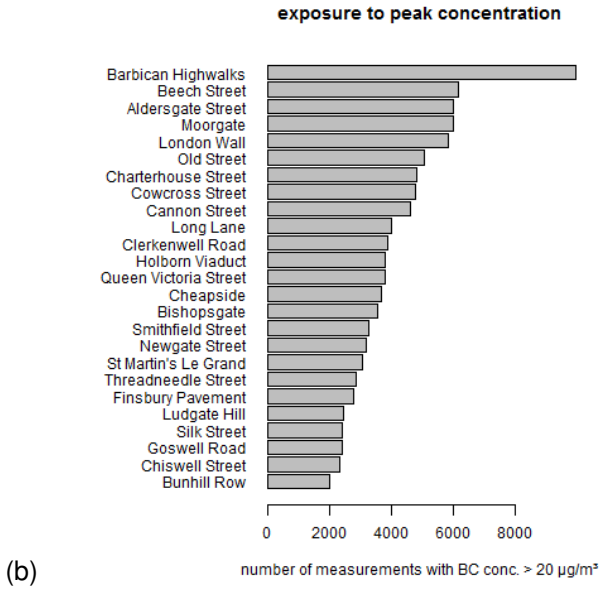


(b)

Figure 2.8: Average black carbon concentration per street in Turin (a) and the number of peak events (black carbon concentration >20μg/m^3) at these streets (b).



(a)



(b)

Figure 2.9: Average black carbon concentration per street in London (a) and the number of peak events (black carbon concentration >20μg/m^3) at these streets (b).

"Frankfurter Straße" (D), "Weserstraße" (E), "Schützenstraße" (F)] are also well covered.

For London, examining the lower end of the scale, again unexpected values are found - in particular, average readings of between 0 and 2.5 along a major route through the area (Aldersgate Street, which is the A1 road going north from London, towards the North East of England). Similar low readings can be found along two other major routes - the A1211 and the A501 (Moorgate) (in the UK context, an A-road is a major route having speed limits of between 100 and 112 kmph). (See Section 4.6, Deliverable D6.3 for the participants' views on the readings obtained).

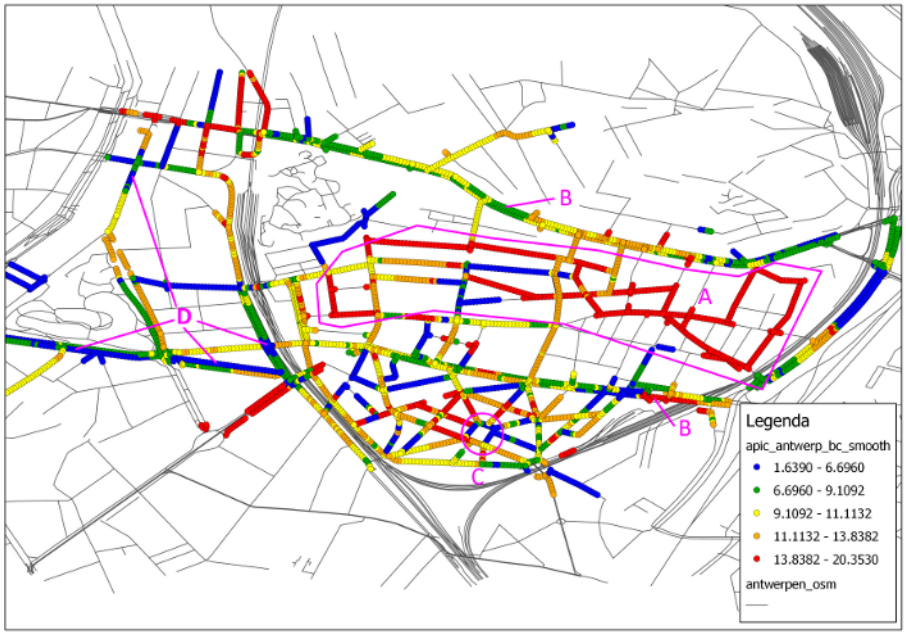
2.2.3 Visualization tools from the web-platform

Air quality data are freely accessible from the website `http://airprobe.eu/`. Different web pages (Explore, Understand, Collect and Compare) are designed to assist and guide citizens from the data collection to the data interpretation (see D2.2). Different visualization tools have been implemented for optimal use. Citizens can track their own measurments and measurment activities, analyse personal exposure data from time series or spatial representations of measurements from a Google Earth plugin. Furthermore, collective data, i. e., all the data that are recorded by the EveryAware Air Quality sensing platform are vizualised on a open street world map (OSM). The air paraameter shown on this map is black carbon, and a link to additional information on black carbon is provided. From the air quality data, a black carbon head map is constructed which is used as additional map layer. Finaly, a point layer with statistics about number of measurements (counts), range of dates when measurements where made (from, until), mean estimated black carbon concentration and the number of estimated black carbon values is given. Point statistics differ at different zooming levels by changes in the aggregation of measurements. The colours of the points is according to the mean estimated black carbon concentration. A screenshot of the map is given in Fig. 2.12. Air quality data are freely accessible from the website `http://airprobe.eu/`. Different web pages (Explore, Understand, Collect and Compare) are designed to assist and guide citizens from the data collection to the data interpretation (see D2.2). Different visualization tools have been implemented for optimal use. Citizens can track their own measurments and measurment activities, analyse personal exposure data from time series or spatial representations of measurements from a Google Earth plugin. Furthermore, collective data, i. e., all the data that are recorded by the EveryAware Air Quality sensing platform are vizualised on a open street world map (OSM). The air parameter shown on this map is black carbon, and a link to additional information on black carbon is provided. From the air quality data, a black carbon head map is constructed which is used as additional map layer. Finaly, a point layer with statistics about number of measurements (counts), range of dates when measurements where made (from, until), mean estimated black carbon concentration and the number of estimated black carbon values is given. Point statistics differ at different zooming levels by changes in the aggregation of measurements. The colours of the points is according to the mean estimated black carbon concentration. A screenshot of the map is given in Fig. 2.12. Further technical details about the visualization platform can be found in D2.2.

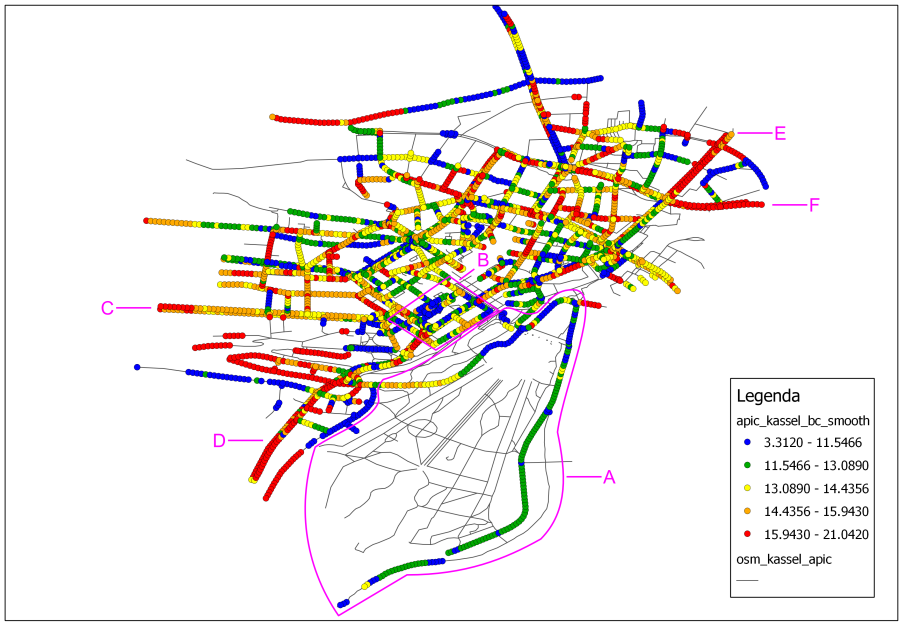
2.3 Conclusions from the air quality case studies

Case studies were deployed in Antwerp, Kassel, Turin and London. Participants were equipped with sensor boxes to collect air quality data in a confined area of about 2-4 km^2 within these cities. From our experience with mobile air quality monitoring including experiments and analyses performed under Task 4.1 in the First Reporting Period of the project we learned that repeated measurements at the same location are needed to increase the representativity of the air quality mapping. Therefore we defined study areas of a feasible size to monitor in a two weeks period by 10-12 sensor boxes.

Air quality data were cleaned and validated by implementing data validation procedures. Data val-

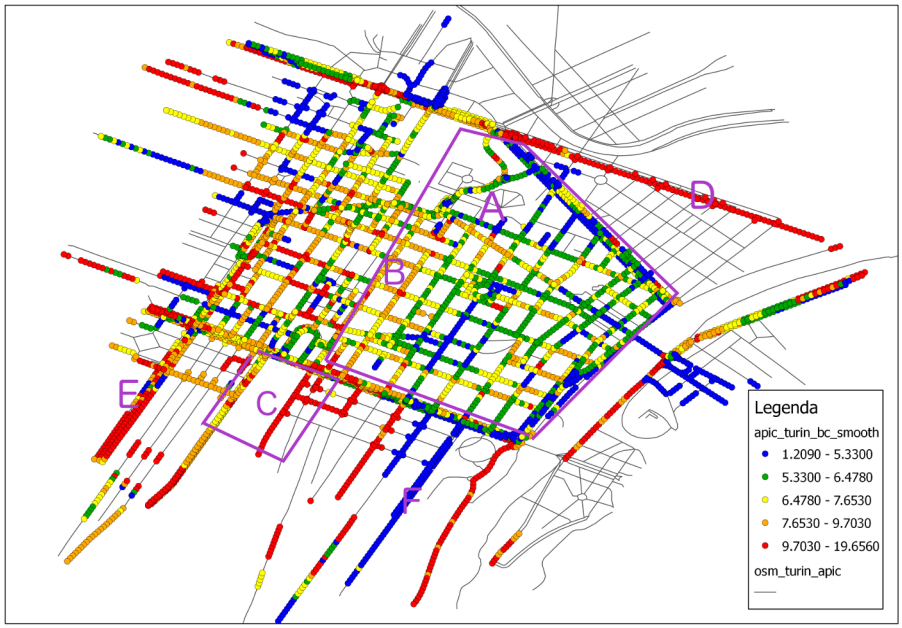


(a)

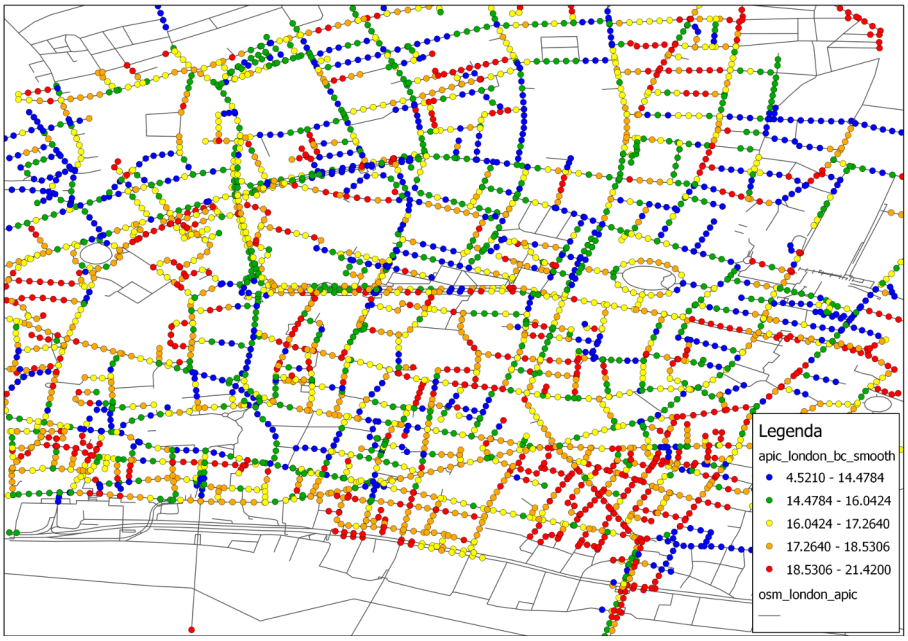


(b)

Figure 2.10: Maps of the smoothed black carbon concentration in Antwerp (a) and Kassel (b).



(c)



(d)

Figure 2.11: Maps of the smoothed black carbon concentration in Turin (c) and London (d).

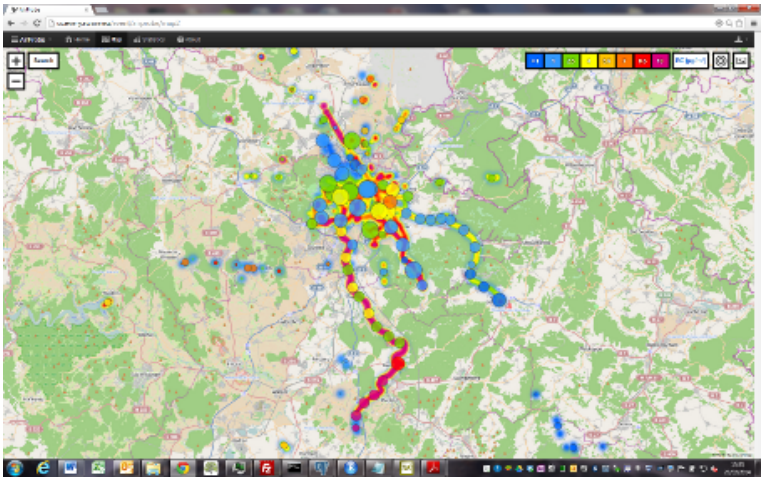


Figure 2.12: Visualization of air quality measurements on the project webpage (zoom of Kassel, Germany, and surroundings).

idation was needed because of sensor box hardware issues (sensor heating, GPS failure, sensor failing) and issues related to mobile monitoring (e. g., mixing of indoor and outdoor measurements, inappropriate handling of the sensor box). The data validation resulted in a large reduction of measurement data in all the four cities during the test cases. However, the validation procedures could not be properly evaluated and it was decided not to implement them in the EveryAware system. Further developments should direct towards methods to decrease hardware issues and toward an automated data validation protocol.

The spatio-temporal analysis of sensor data showed that the EveryAware platform is suitable to collect, transfer, store and visualize air quality measurements. High resolution maps of BC concentrations were obtained from a two weeks long monitoring campaign and allowed to recognize spatial air quality patterns at street level. Nevertheless, the map with modelled black carbon shows much less variability than the map with observed black carbon concentrations, probably due to slower sensor response times and also absolute BC concentrations differed substantially compared to reference maps.

Chapter 3

Analysis of noise sensor data

Although the main focus of the second phase of the EveryAware project has been on Air Quality, the work commenced in Phase 1 on general noise capture via the WideNoise App has continued throughout Phase 2, with a particular focus on the continuing Large Scale Case Study at Heathrow Airport. While Deliverable D6.3 [EveryAware, 2014c] provides detail on the noise measurements with respect to the continuation of the Large Scale Case Study around London's Heathrow airport (comparing results obtained to the overall dataset), this section summarizes the noise data captured through the 3-year EveryAware project as a whole, whether through a specific project or otherwise for further analysis, the reader is referred to [Becker et al., 2013] where an in-depth interim analysis of the noise data is provided).

3.1 Quantitative Noise Results

Figure 3.1 shows the map of the distribution of the decibel measurements taken by WideNoise-enabled phones throughout the project, as clustered points. Figure 3.2 shows this data as a gridded structure. A total of 48406 points have been added to the database overall, with 21520 of these since the interim report.

Table 3.1 provides more detail about the dataset:

Figure 3.3 shows the number of measurements per device over the entire period of the EveryAware project. As can be seen, a significant proportion of the total devices used (14310 of 15293) took fewer than 5 measurements, confirming the trend observed in D4.1 [EveryAware, 2014b]

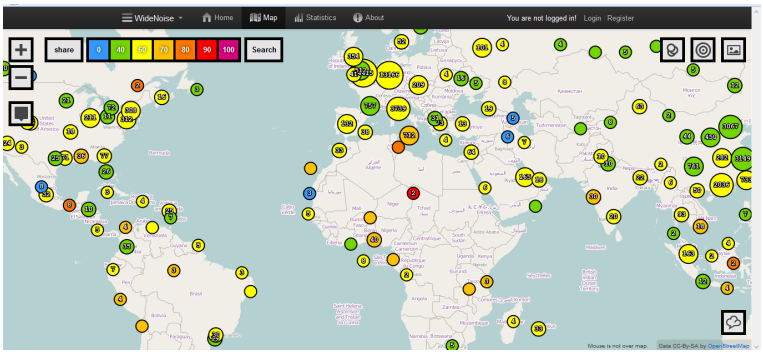


Figure 3.1: WideNoise Data Captured - World Overview (clustered points).

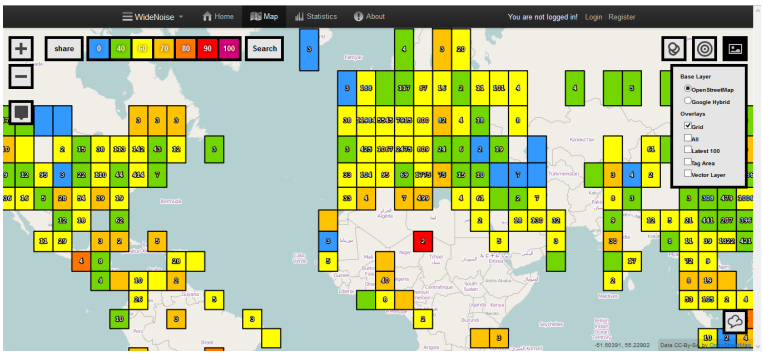


Figure 3.2: WideNoise Data Captured - World Overview (grids).

Data	
Location	worldwide
Measurements	
Number of measurements	48406 (24886)
Number of measurements with geo-coordinates	38228 (17011)
Number of Measurements with geo-coordinates from ip	10178 (7627)
Number of measurements with perceptions	16104 (8015)
Coverage	
Overall duration of measurements	78.74h(41.5h)
Decibel Statistics	
Average	63.93(63.94)
Standard deviation	19.28(19.27)
Minimum	0 (0)
Maximum	119.89 (119.89)

Table 3.1: Worldwide WideNoise Summary (figures in brackets refer to the values at Month 18).

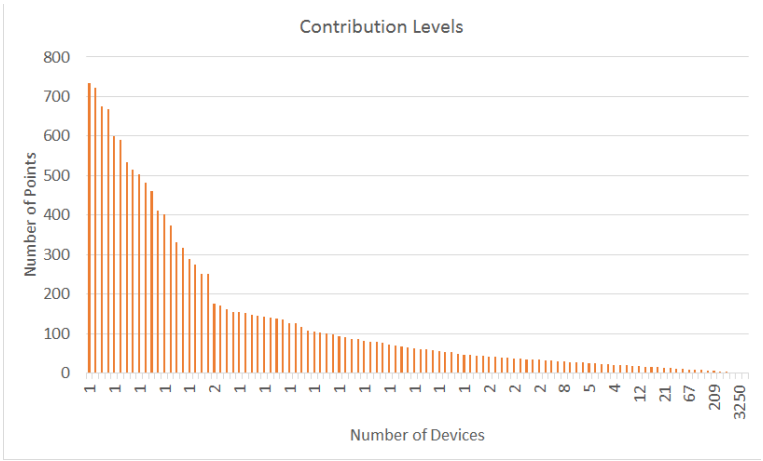


Figure 3.3: Number of Devices Versus Number of Points

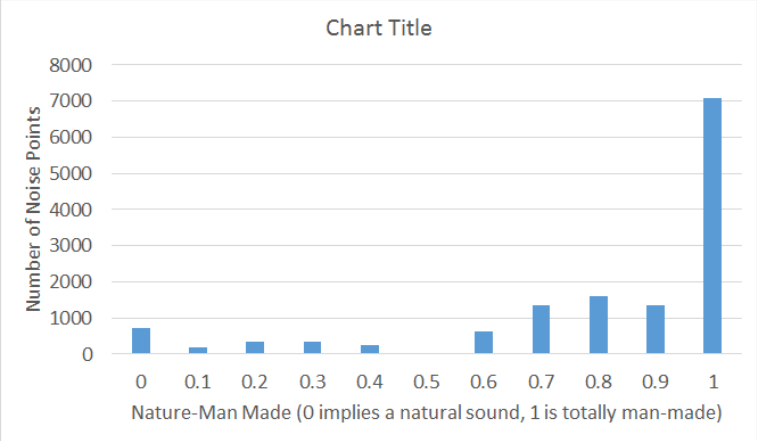


Figure 3.4: User Perception - Natural versus Manmade Sounds.

3.2 Relating Qualitative and Quantitative Data

As noted in the interim report [EveryAware, 2014b], unlike with Air Quality data it is not possible to directly compare the noise data measured using the WideNoise tool with existing noise models, as the latter are not based on direct measurement but rather are created through the consideration of measured traffic flows and other contributing environmental factors. There are thus no specific interpolation methods appropriate for noise data measured as points in space and time. However, the WideNoise App does permit the comparison of subjective and objective data, and facilitates a greater understanding of how perceived noise differs or is similar to measured noise.

Figures 3.4, 3.7, 3.5 and 3.6 show the choices made by users for their perceptions relating to whether they love or hate the sound, were alone or in a social situation when the reading was taken, whether the environment was calm or hectic and whether it was natural or man made. As noted in Deliverable D3.1 [EveryAware, 2014a] the sliders used to capture these values permit users to move from left to right on the screen, making extreme values on the scale (0 or 1) perhaps more easy to select. The information presented here should therefore be reviewed in this light.

A total of 13870 people submitted a value for the Nature/Man-Made option (note that this figure exclues those who left the slider at its default value as it is not possible to distinguish between those users who didn't submit a value and those who deliberately opted to set the value as half way between the two). As can be expected, in Figure 3.4 the majority of users selected 'man made' when evaluating the sound. This is most likely due to the fact that many of the noise measurements (and in particular the targeted Campaigns) took place in urban areas.

A total of 11512 people submitted a value for the Calm/Hectic option (as above this figure excludes those who left the slider at its default value). Although the extreme range is less dominant, in Figure 3.5 the majority of users selected 'hectic' when evaluating the sound. As suggested in Deliverable D3.1 [EveryAware, 2014a] this may be due to the fact that users make measurements when noise has become annoying - e.g. when a plane is overhead or perhaps when they are in a hectic environment.

A total of 11370 people submitted a value for the Alone/Social option (as above this figure excludes those who left the slider at its default value). In this case, however, there is a far less significant different between the extreme ends of the scale.

A total of 11500 people submitted a value for the Love/Hate option (as above this figure excludes those who left the slider at its default value). Although the extreme range is again less dominant, in Figure 3.7 the majority of users selected 'hate' when evaluating the sound. As suggested in Deliverable D3.1 [EveryAware, 2014a] this may be due to the fact that users make measurements

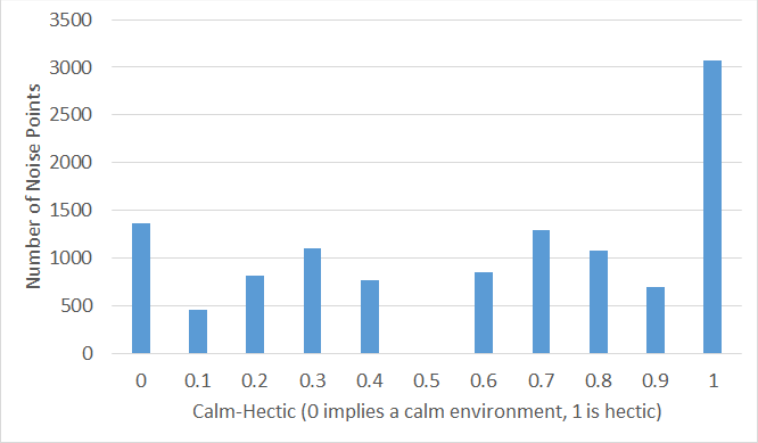


Figure 3.5: User Perception - Calm versus Hectic Environment.

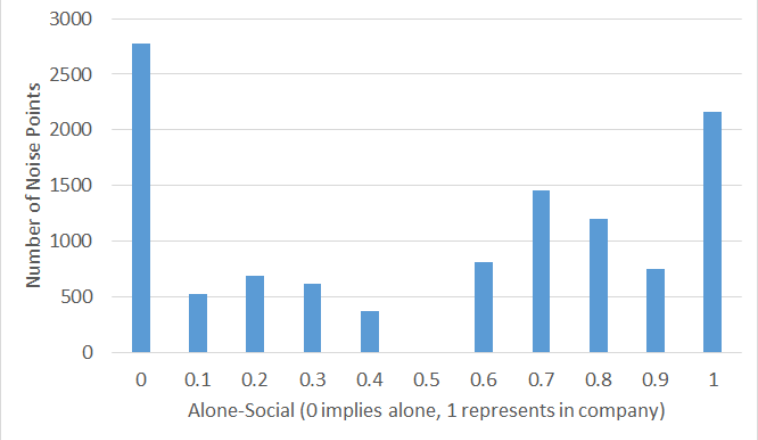


Figure 3.6: User Perception - Are You Alone or in a Group.

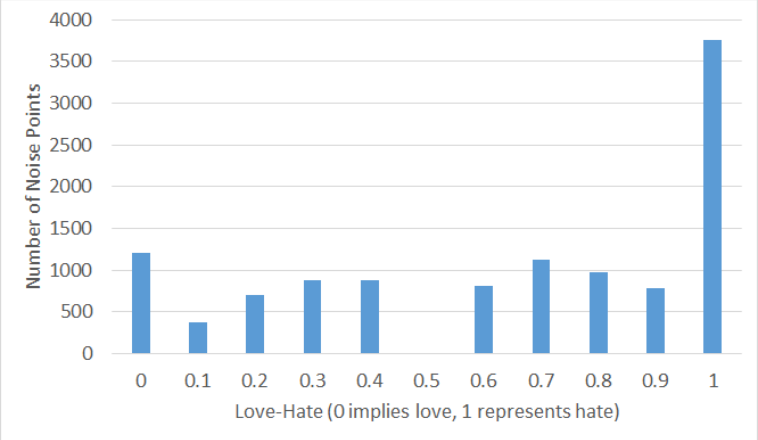


Figure 3.7: User Perception - Love versus Hate the noise.

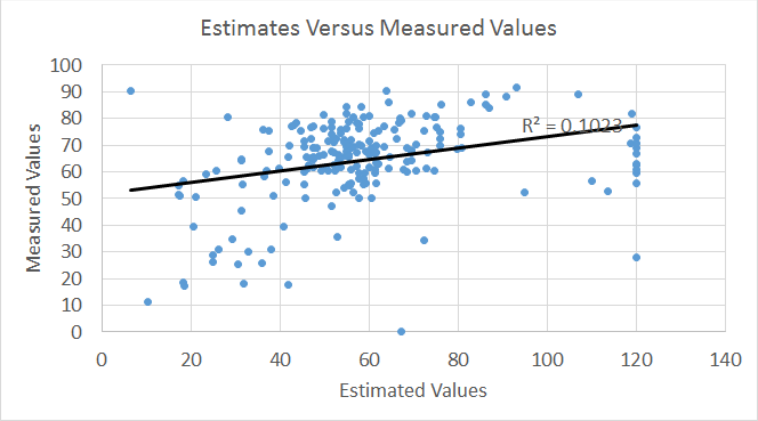


Figure 3.8: Estimated versus Measured Noise.

when noise has become annoying - e.g. when a plane is overhead or perhaps when they are in a hectic environment.

Figures 3.8 examines the difference between the perceived (as guessed by the user) and measured noise values (with measurements where the user did not submit an estimate or submitted an estimate of 0 eliminated). Although an approximately linear trend can be observed, R^2 is 0.1023, indicating a relatively poor match.

3.3 From Measurement to Policy

The results obtained through the use of the WideNoise App in the EveryAware project highlight once again the relevance of a coordinated campaign of noise measurement to ensure at least minimum coverage across a specific location. Given the varying accuracy issues relating to the data, the importance of coupling subjective information (perception) with the measured data cannot be understated, as it is the former that permits a judgment to be made as to how the measured data is perceived, turning the number from sound to noise, and understanding the perception of noise by communities is in turn vital to policy and decision makers. A trade-off is therefore required going forward between automated data capture and continual measurement of noise (such as that carried out by the AirProbe App) and offering the opportunity to participants to capture noise that is relevant to them and tag this as such. While automation may be convenient, it is the manual effort made to capture and tag noise that highlights its importance.

It should also be noted that even if more accurate devices were made available, and higher levels of space/time coverage could be obtained, the question remains open as to the required number of points, their distribution across space and time and subsequent interpretation and/or interpolation methods to make the result comparable to the official maps, given their radically different sources. This is a key open challenge for any Citizen Science noise monitoring project.

Chapter 4

Analysis of Subjective Data

This chapter provides an overview of the subjective data analysis that have been performed on the data that were collected with WideNoise Plus for noise data. **What about the few tags collected in connection with AirProbe? Please LUH comment on this.** The application WideNoise Plus allows users to record sound samples and to annotate them with perceptions and tags. The app documents and maps the soundscape all over the world. The procedure of recording includs the assignment of tags. We analyze the difference resources provided by WideNoise Plus and use the underlying opinionated information to build a recommendation framework. Therefore, we evaluate multiple tag recommendation methods to improve the sensor data collection.

4.1 Constraints

There are some special conditions for mobile sensing that must be addressed when choosing recommender algorithms. WideNoise Plus is most often used outdoors without regard to Internet connectivity. Thus, the application must be able to produce recommendations only from data that has been stored on the device and the elements of the current record (the measured noise, the location, and the user's perceptions). Furthermore, producing recommendations should only consume as little power and runtime as possible. Otherwise, the increased battery drain and long waiting time would discourage users from taking further measurement.

We compare several approaches against each other in our experiments. We describe them in the remainder of this section and discuss their advantages and drawbacks regarding resource consumption as well as their suitability for the mobile environments.

4.2 Methods

Most Popular Tags (MPT)

A very simple recommendation method is to always suggest those tags that have been assigned the most often so far. This yields a non-personalized recommender that will serve as a lower baseline in our comparison of algorithms.

The only input data that would have to be provided for the app are just those top most popular tags. Since also nothing has to be computed, the application would require only very little storage and almost no processing time at all. Therefore, this method would be the best-case in terms of resource requirements.

However, it is expected to be rather bad with regard to the quality of the recommendations, since it is just a static list of the same tags for each record. Table 4.1 shows the list of the current most popular tags. While there are some country specific tags like the Italian word "esterno" (outdoor

scene), there are some international ones like “garden” or “car” that are likely to occur all over the world. Therefore, this recommendation strategy is considered an adequate baseline for our evaluations.

Table 4.1: The 10 most popular tags in the dataset

Amount	Tag
573	garden
557	esterno
549	heathrow
525	aeroplane noise
271	voci
187	car
181	antwerpen
157	plane
151	street
133	arriva

Most Popular Tags by User (MPTU)

Another very simple recommendation method is to suggest those tags that have been used by the given user the most often so far. This yields a personalized recommender that recommends tags that are known to the user and in a language they understand.

It is also very suitable for the mobile devices, as only the user profile and no other training data has to be stored. Using the pre-ordered list of the user’s tags, the algorithm is similarly fast as the global most popular tag recommender. However, this algorithm has a severe cold start problem as it cannot produce tags for new users.

Proximity-Based Approach (Prox)

An approach that uses the location information provided by the location sensor is to recommend tags that have been used so far at the given location or nearby. Prox is thus a context-aware recommender that will recommend tags that likely describe the location like for example “airoplane noise”, which has been used near airports. Therefore, a proximity-based prediction is likely to have good performance.

The algorithm has stronger requirements than the previous ones. Either the whole dataset (all recordings in any location) must be stored on the device or an Internet connection is required beforehand in order to query for records that have been taken roughly near the user’s current location.

In our experiments we will use the k-Nearest-Neighbors algorithm [Ricci et al., 2011, page 129–131] to find the nearby tags. This ensures that this approach always recommends tags even if they are taken from faraway places. For our experiments we manually choose a value of 42 for k, since this showed good results in a subset in the training data.

The distance between two locations can be calculated with a number of methods like the Manhattan, the Euclidean, or the great circle distance. The Manhattan distance is rather inaccurate although very easy to compute. The Euclidean distance is much better in terms of accuracy, but with the price of a higher computational effort. However, compared to the actual air-line distance, the accuracy is getting worse for locations further away from the equator. Finally, the great circle distance is very precise, but is the most expensive with regard to computations. We use the Euclidean distance (Prox-ED) and the great circle distance (Prox-GCD) due to its higher accuracy.

Perception-Based Approach (Perc)

An approach that uses WideNoise Plus's perception values is able to recommend tags that are associated with the same mood (e.g., "love"). This yields a context-aware recommender that will recommend tags that describe the user's perception of the noise, location, etc. (e.g., "noisy plane spoiling peace"). There are four scales with a range from -5 to +5 each with steps of size 1 to express the corresponding perception:

- Feeling: Ranges from "hate" to "love" and expresses whether the user enjoys the recorded noise or whether it was unpleasant.
- Disturbance: Ranges from "hectic" to "calm" and expresses how disturbing the recorded noise was perceived by the user.
- Isolation: Ranges from "alone" to "social" and expresses how much company the user had.
- Artificiality: Ranges from "man-made" to "nature" and expresses whether the recorded noise was caused by humans, machines, or nature.

The method is suitable for mobile devices, as only an aggregated list of tags for each possible perception combination has to be stored. Using the pre-ordered lists of the perception's tags, the algorithm has to combine those lists that are the most similar to the given perception setting. A perception vector p' is considered similar to the current perception vector p if no perceptions differ more than a given threshold d, i.e., if $||p - p'||_\infty \leq d$.

In our experiments we will set the (initial) threshold to $d = 1$ and increase it by one in cases where no such measurement p' exists and thus nothing could be recommended.

Clustering-Based Most Popular (Clus)

This approach, presented by [Abbasi et al., 2009] uses the location information of the location sensor to cluster the records and assign the most frequent tags ordered by decreasing user frequency of a cluster's records to that cluster during a preprocessing step. Recommended are those tags that have been used in the cluster of a given location so far. This yields a context-aware recommender that will likely recommend tags that describe the location. This algorithm is similar to Prox, but, since the records are clustered, the computational effort and the amount of input data is lower.

It is thus suitable for mobile devices, as only the precomputed ranked list of tags of each cluster have to be stored. For each new record, the distance to all clusters has to be computed to select the ranked tag list of the cluster closest to the user.

In an offline preprocessing, the resources are clustered using k-Means and the most frequent tags for each cluster are determined. k-Means requires the number of cluster k as an input parameter as well as a distance computation function. For k we use the rule of thumb proposed by [Mardia et al., 1979, page 365]:

$$k \approx \left(\frac{n}{2}\right)^{\frac{1}{2}}$$

Hereby, n refers to the number of resources to be clustered and the Euclidean distance is used as distance function. Clusters are represented by their centroids and in the recommendation phase, we use the Euclidean distance for distance calculation.

During our experiments we discovered that, in our scenario, it is better to choose the absolute tag frequency during clustering phase rather than the user frequency. We will present the result for user frequency (i.e., Clus-UF) and absolute tag frequency (i.e., Clus-AF) separately during our evaluation.

Hybridization

To improve performance, multiple recommenders can be combined in hybrid recommenders. Such a combination can improve the results by combining several aspects, e.g., to yield a location-based approach that also is influenced by the given perceptions.

The suitability for our scenario depends on the algorithms that are combined. In this paper we will analyze 3 combinations between most popular tag by user on the one hand and either the perception (Perc-MPTU), proximity (Prox-ED-MPTU), or clustering (Clus-ED-MPTU) approach on the other hand. We use most popular tag by user as it produces personalized recommendations with only little computational effort. In order to keep the computational effort small we chose the Euclidean distance-based versions of Perc and Prox.

All involved algorithms compute their individual rankings. For a tag we compute a score as an unweighted linear combination [Burke, 2002] of the inverse ranks according to the following equation:

$$score(t) = \left(\frac{1}{\text{rank}_1(t)} + \frac{1}{\text{rank}_2(t)} \right)^{-1}$$

Hereby, $\text{rank}_1(t)$ and $\text{rank}_2(t)$ are the positions of the tag t in the rankings of the two combined algorithms.

4.3 Dataset and Experiments

In this section we introduce the dataset of our analysis and how it was assembled as well as the metrics we use for the evaluation in Section 4.4.

4.3.1 Dataset

The basis for our experiments is the full set of WideNoise Plus records with at least one tag, collected between December 14, 2011 and June 12, 2013. After the removal of records that had been submitted for testing by the developers, the collection consists of 5,434 reports collected by 546 users that contain 1,151 distinct tags and 9,255 tags in total. The following further preprocessing steps were applied to the tags: All tags have been lower-cased and some encoding issues have been resolved manually (e.g., we replaced “wrzburg” with “würzburg”).

Before we describe the experiments on tag recommendations, we observe a few statistical properties of the datasets. Figure 4.1 shows the distribution of the tag frequency. The distribution tends to be fat tailed.

Figure 4.2 shows the distribution of the number of tags assigned to one record. The maximum number of assigned tags is 8 and we therefore pick it as the maximum number of recommended tags in our experiments. On average, one WideNoise Plus record has 2.45 tags assigned to it.

Figure 4.3 shows the distribution of the number of tag assignments per user. The most active user assigned 2,461 tags and the average number of tag assignments per user is 33.92. However, we have a fat tail of users that made just one tagged records and then stopped using this feature.

4.3.2 Evaluation

We evaluate the different recommendation algorithms in an offline experiment. We split the full dataset into training and test data using a time split after 70 % of the records leaving 3,805 records for the train phase and 1,629 records for the evaluation phase. In that way, we stay close to the actual scenario: The WideNoise Plus app runs on a mobile device and must produce recommendations from the data on the device. While it is not possible to send training data record by record

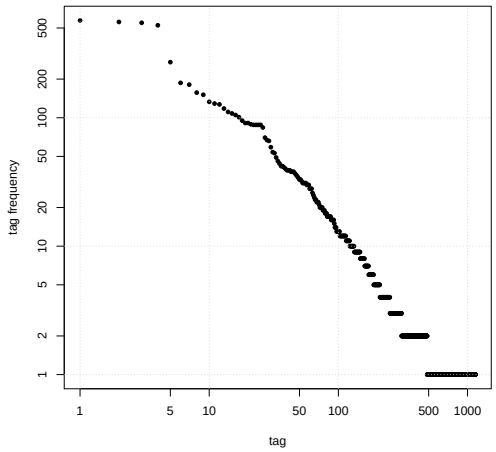


Figure 4.1: Distribution of the tag frequency on a log-log scale. The elements on the x-axis are the 1,151 unique tags, ordered by decreasing frequency.

to an application, it is very well conceivable, to update the app with training data in larger regular intervals. A consequence of this procedure is that the test data set contains users and tags that do not occur in the training data. Again, this is close to the real scenario, where often users take measurements over only a short time span and thus do not have large user profiles to be used for training. This closeness to the real-world scenario was the decisive element for a time split and against other methods like cross validation procedures, where random samples of the full data set are selected as test data.

The algorithms are trained and then used to produce a ranked list of recommendations for each record in the test dataset comprising the user, the sensor measurements (longitude, latitude and noise level) and the four perceptions. To evaluate the performance we measure the predictive power of recommendations, i.e., for every record of the test data, precision, recall, and F_1 measure are computed. For these three metrics, the number of recommended tags has to be set to some fix number k. To pay tribute to the size of mobile devices and following the findings above on the maximum number of assigned tags, we let k run from 1 through 8 and compute the score at each level. Thus, if A is the set of tags that were actually assigned to the record and P is the set of the top k recommended tags, then precision and recall are defined as follows [Ricci et al., 2011, page 109]:

$$\mathrm{Precision}(A, P) = \frac{|P \cap A|}{|P|}$$
$$\mathrm{Recall}(A, P) = \frac{|P \cap A|}{|A|}$$

The F_1 measure is the harmonic mean of precision and recall:

$$\mathrm{F}_1(A, P) = 2 \cdot \frac{\mathrm{Precision}(A, P) \cdot \mathrm{Recall}(A, P)}{\mathrm{Precision}(A, P) + \mathrm{Recall}(A, P)}$$

Theoretical upper bound In the experiments, we will compare not only different algorithms against each other, but also to a theoretical “perfect recommender”. This upper bound demonstrates, how much room for improvements is left for further, possibly more advances methods in future work. The bound is constructed by recommending those tags for a record that have actually been used for it as long as these tags occur in the training data. It is clear that no real algorithm

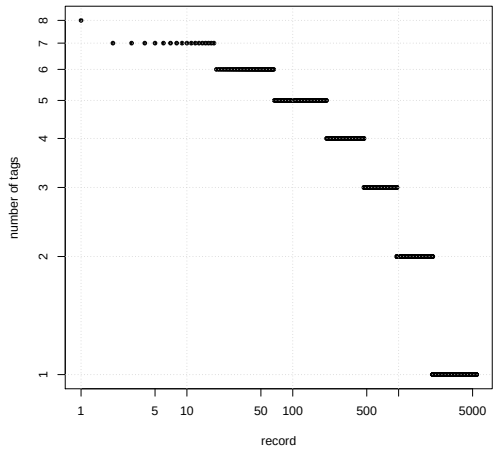


Figure 4.2: Distribution of tags per record on a log-log scale. The x-axis represents the dedicated records and the y-axis represents the number of tags assigned to such a record.

Table 4.2: F_1 measure for WideNoise Plus

number of tags	MPT	MPTU	Perc	Prox-ED	Prox-GCD	Clus-UF	Clus-AF	Perc-MPTU	Prox-ED-MPTU	Clus-AF-MPTU	upper bound
1	0.002	0.180	0.084	0.213	0.204	0.049	0.165	0.165	0.234	0.186	0.433
2	0.028	0.175	0.113	0.203	0.204	0.054	0.192	0.204	0.227	0.229	0.450
3	0.043	0.162	0.100	0.181	0.182	0.047	0.178	0.193	0.206	0.204	0.384
4	0.036	0.153	0.098	0.170	0.169	0.064	0.164	0.165	0.190	0.188	0.327
5	0.032	0.149	0.091	0.153	0.162	0.057	0.155	0.146	0.168	0.176	0.284
6	0.030	0.135	0.084	0.148	0.146	0.051	0.138	0.139	0.163	0.161	0.251
7	0.027	0.124	0.082	0.136	0.134	0.046	0.128	0.132	0.149	0.149	0.225
8	0.026	0.116	0.076	0.126	0.125	0.042	0.120	0.125	0.139	0.139	0.204

– which of cause has no knowledge about the user's actually chosen tags – can beat that upper bound (unless it can produce tags that have never been used before).

4.4 Results

In the evaluation we use the global most popular recommender as baseline. Every more sophisticated recommender should achieve better results than that. Additionally, we include the values for a perfect pseudo-recommender that predicts just those tags that are actually used and present in the train dataset.

As described in Section 4.2, there are language specific tags in the list of the global popular tags (see Table 4.1). Further, besides the larger geographic areas, there are also tags that will occur only in certain particular areas like the tag "heathrow", which would be relevant only around London. It is therefore to be expected, that the most popular baseline will achieve rather low results.

Figure 4.4(a) shows the results for precision and recall. The F_1 measure results are shown in more

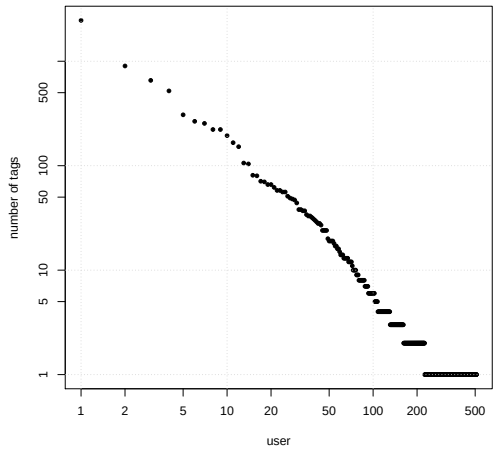


Figure 4.3: Distribution of the number of tag assignments per user. The x-axis represents the users and the y-axis represents the number of tag assignments of these users.

detail in Figure 4.4(b) and Table 4.2. In the discussion, we focus on the scores that are obtained for the recommendation of two and three tags respectively, since the average amount of assigned tags in the dataset is 2.45. Compared to the baselines, we observe, that all algorithms successfully outperform the most popular tags recommender, but also – comparing to the theoretical upper bound – that there is plenty of room for improvements.

An interesting results is that the personalized MPTU approach yields a very good score. It is already better than the computationally intensive Perc approache, but slightly worse than Clus and Prox.

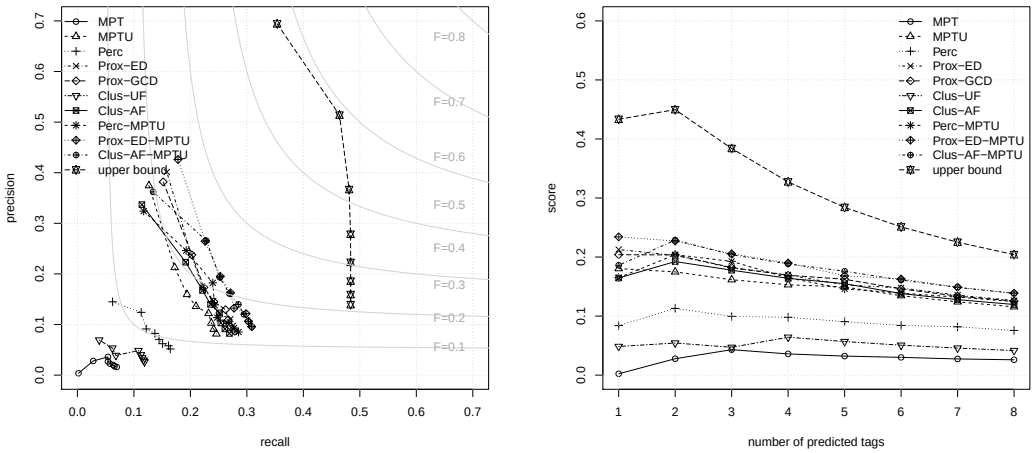
It is very interesting that in comparison to the use of the Euclidean distance, the great circle yields almost the same results. For our scenario, this is good news, as similar recommendations are produced with less computational effort. The use of clustered locations (i.e., Clus) yields similar results as Prox-ED and Prox-GCD, but is computationally less expansive.

Looking at the hybridization results, we see that all algorithms profit from the merge with MPTU. Prox-ED benefits far more from MPTU than Clus-AF and achieves the best results among all investigated algorithms – approximately already half of the maximal possible score.

To evaluate the suitability for mobile devices we measured the runtime it took each recommender to predict the tags for the whole evaluation dataset. Figure 4.5 depicts the computation time for every algorithms[1].

The computational effort of the great circle distance is not acceptable considering the almost same performance. While Prox-ED-MPTU achieved the best recommendation quality, it requires a lot of computation time. Still one has to consider that the analysis was conducted on a relatively powerful computer and that the times will increase with a growing dataset. The runtimes can therefore only be used as indicators, since smartphones have much less computation power and would therefore take much longer.

[1]The evaluation was conducted on a Lenovo ThinkPad X220 with an Intel Core i7-2640M (2.80 GHz), 8 GB RAM, Windows 8 Professional 64-bit, and Gnu R 2.15.3 64-bit.



(a) Recall vs. precision. The grey lines are isoquants for several F_1 scores.

(b) F_1 measure.

Figure 4.4: Evaluation results for WideNoise Plus.

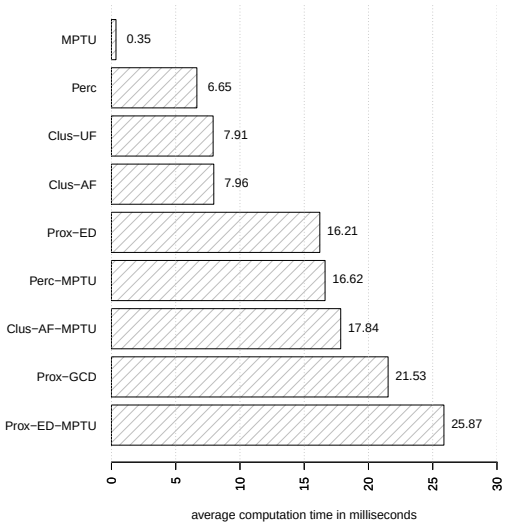


Figure 4.5: Average recommender runtime.

Chapter 5

Perceived versus measured environment

In this chapter we present and discuss the first results of our method of subgroup discovery aimed at obtaining interesting descriptive patterns in ubiquitous data. In particular, by applying our method to the data collected by the users of our Widenoise application, i.e. average sound level measurements, subjective perceptions and tags, we carry out a factor analysis to detect which environmental parameters have the strongest impact on user's environmental perception. In doing that, we provide a novel graph-based analysis approach for assessing the relations between the obtained subgroup set, and for comparing subgroups according to their relations to other subgroups.

In the last section we present the general statistics of the AirProbe web game, where players had to guess the air pollution level of their city by relying on their perception of the environment. Moreover, we perform a detailed analysis of player behavior and their game strategy change in response to the actual air pollution (black carbon levels) measurements taken with our sensor box.

5.1 Overview

For subgroup analytics, we first obtain a set of the top-k subgroups for a specific target variable. Typically, an efficient subgroup discovery algorithm needs to be applied. In our experiments, we apply the SD-Map* [Atzmueller and Lemmerich, 2009] algorithm for efficient subgroup discovery, which is suitable for sparse tagging data [Atzmueller and Lemmerich, 2013]. After that, the set of subgroups needs to be assessed and put into relation to each other.

The proposed approach especially focuses on this specific step: It considers a relation between subgroups such that their "connections" according to this relation can be modeled as a graph. More formally, given a certain criterion implemented by a relation function $rel : I \times I \rightarrow \mathbb{R}$ we obtain a value estimating the relationship between pairs of subgroups, identified by their respective subgroup descriptions. Possible relations include, for example, geographic distance, or semantic criteria. In our application setting, we focus on the latter, since we will use the given perceptions for noise measurements as semantic proxies for subgroup relatedness.

For assessing our result set of subgroups R, we obtain the rel-value for each pair of subgroups (u, v). After that, we construct a *subgroup assessment graph* G_R for R: The nodes of G_R are given by the subgroups contained in R. The edges between node pairs (u, v) are constructed according to the respective $rel(u, v)$ value: If the respective value between the subgroup pair is zero, then the edge is dropped; otherwise, an edge weighted by $rel(u, v)$ is added to the graph.

It is easy to see that – depending on the applied relationship function rel – this construction process can result in a fully connected graph which is hard to interpret. Therefore, a refinement of this process utilizes a certain threshold τ_{rel} which is used for pruning edges in the graph. If the relation

"strength" $rel(u, v)$ between a subgroup pair (u, v) is below the threshold, i. e., $rel(u, v) < \tau_{rel}$ then we do not consider the edge between u and v, such that the edge is dropped. By carefully selecting a suitable threshold τ_{rel} the resulting subgroup network can then be easily inspected and assessed.

Typically, the situation becomes interesting when the graph is split into different components corresponding to certain clusters of subgroups. We will discuss examples of constructed networks below. For selecting a suitable threshold, a *threshold-component* visualization can be applied, see Figure 5.5 for an example. This visualization plots the number of connected components of the graph depending on the applied threshold. Then, the "steps" within the plot can indicate interesting thresholds that can be interactively inspected. A related visualization plots the used threshold against the graph density for obtaining a first impression of the ranges of suitable threshold selections.

5.2 Applied Dataset

We utilize data from WideNoise Plus application between December 14, 2011 and June 12, 2013. WideNoise Plus allows the storage of noise measurements using ubiquitous mobile devices, and includes sensor data from the microphone given as noise level in dB and data from the location sensors (i.e., GPS-sensor, GSM- and WLAN-locating) represented as latitude and longitude coordinate as well as a timestamp. Furthermore, WideNoise Plus captures the user's perceptions about the recordings, expressed using the four slider feeling (love to hate), disturbance (calm to hectic), isolation (alone to social), and artificiality (nature to man-made). In addition, tags can be assigned to the recording. In our analysis, we utilize the following objective and subjective information for each measurement:

- Objective: Level of noise (dB).
- Subjective perceptions about the environment:
  - "Feeling" (hate/love) encoded in the interval $[-5; 5]$, where -5 is most extreme for "hate" and 5 is most extreme for "love".
  - "Disturbance" (hectic/calm), encoded in the interval $[-5; 5]$, where -5 is most extreme for "hectic" and 5 is most extreme for "calm".
  - "Isolation" (alone/social), encoded in the interval $[-5; 5]$, where -5 is most extreme for "alone" and 5 is most extreme for "social".
  - "Artificiality" (man-made/nature), encoded in the interval $[-5; 5]$, where -5 is most extreme for "man-made" and 5 is most extreme for "nature".
- Tags, e. g., "noisy", "indoor", or "calm", providing the semantic context of the specific measurement.

The applied dataset contains 5,237 data records and 1,056 distinct tags: the available tagging information was cleaned such that only tags with a length of at least three characters were considered. Only data records with valid tag assignments were included. Furthermore, we applied stemming and split multi-word tags into distinct single word tags.

Figures 5.1-5.4 provide basic statistics about the tag count and measured noise distributions, as well as the value distributions of the perceptions and the number of tags assigned to a measurement. As can be observed in Figure 5.1 and Figure 5.4, the tag assignment data is rather sparse, especially concerning larger sets of assigned tags. However, it already allows to draw some conclusions on the tagging semantics and perceptions. In this context, the relation between (subjective)

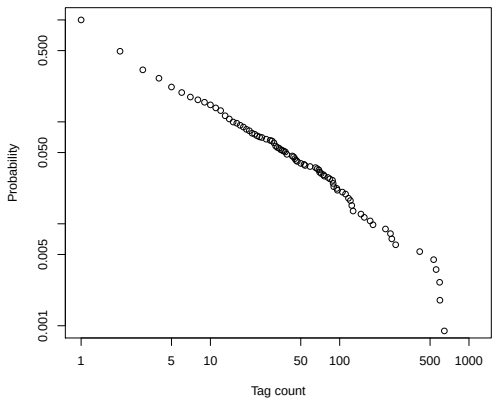


Figure 5.1: Cumulated tag count distribution in the dataset. The y-axis provides the probability of observing a tag count larger than a certain threshold on the x-axis.

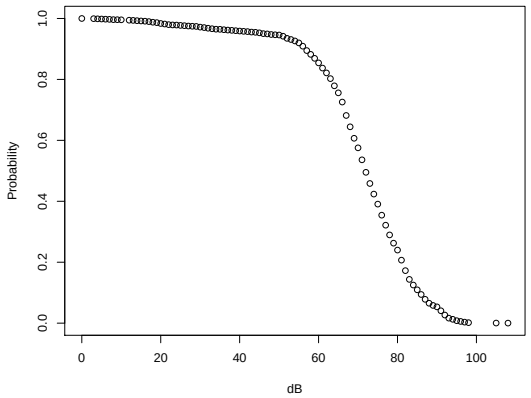


Figure 5.2: Cumulated distribution of noise measurement (dB). The y-axis provides the probability for observing a measurement with a dB value larger than a certain threshold on the x-axis.

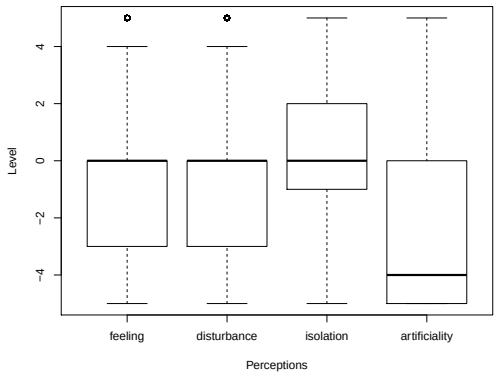


Figure 5.3: Overview on the value distribution of the different perceptions.

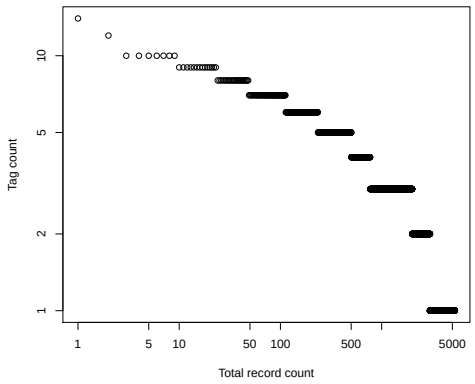


Figure 5.4: Distribution of assigned tags per resource/data record.

perceptions and (objective) noise measurements is of high interest. Therefore, we present first analysis results of interesting patterns in the case study described below. We focus on the relation between semantics and perceptions as indicated by the different subjective perception values.

5.3 Case Study: First Results and Discussion

In the following, we present first analysis results in the context of the WideNoise Plus data. According to the proposed approach, we applied subgroup discovery for the target variable *noise (dB)* focusing on subgroups with a large deviation comparing the mean of the target in the subgroup and the target in the whole database. We applied the simple binominal quality function. Table 5.1 shows the resulting 20 patterns combining the two top-10 result sets.

In the table, we can identify several distinctive tags for noisy environments, for example, *craft*, *aircraft*, *plane*, *heathrow AND plane* which relate to Heathrow noise monitoring [Atzmueller et al., 2012] for more details. These results confirm the basic analysis in [Atzmueller et al., 2012]. For more quiet environments, we can also observe typical patterns, e. g., focusing on the tags *indoor*,

Table 5.1: Patterns: 1-10 - target: large mean noise (dB); 11-20 - target: small mean noise (dB); Overall mean (population): 70.12 dB. The last two columns include the node degree in the subgroup assessment graph, for $\tau_{rel} = 0.90$ and $\tau_{rel} = 0.95$.

id	description	size	mean dB	feeling	disturbance	isolation	artificiality	deg (t=0.9)	deg (t=0.95)
1	craft	67	92.10	-3.06	-3.21	3.21	-4.61	4	1
2	air	72	89.72	-3.07	-3.10	2.97	-4.57	4	1
3	arriva	252	78.64	-0.02	-0.01	0.01	0.00	9	8
4	plane	415	76.26	-3.47	-2.61	-0.59	-3.75	5	0
5	heathrow AND plane	31	87.81	-4.61	-4.48	-0.32	-4.65	3	2
6	runway	107	79.62	-3.78	-3.45	-1.45	-3.94	3	2
7	runway AND plane	92	79.92	-3.75	-3.67	-1.38	-3.78	3	2
8	aeroporto	13	94.08	-5.00	0.00	0.00	0.00	6	1
9	ciampino	16	91.13	-4.06	0.00	0.00	0.00	10	2
10	departure	14	92.50	-0.71	0.57	-0.29	-1.36	11	8
11	home	124	45.58	1.10	1.31	-0.96	-0.99	9	7
12	bosco	17	35.35	3.29	3.53	-1.65	1.88	0	0
13	indoor	111	56.69	0.81	0.71	-0.17	-1.29	9	8
14	office	172	59.78	0.10	0.68	-0.35	-1.68	11	9
15	borgo	12	31.33	3.00	3.25	-1.00	1.67	0	0
16	background	35	48.06	0.40	2.11	-2.46	-0.97	10	9
17	work	74	55.76	-0.49	0.19	-0.35	-1.86	11	5
18	indoor AND background	22	44.32	0.55	1.91	-2.14	-0.73	10	8
19	kassel	96	58.67	-0.17	0.64	0.17	-1.41	10	9
20	work AND background	23	47.43	0.61	1.74	-2.00	-0.74	10	8

background and *work*, and combinations. Some further interesting subgroups are described by the tags *bosco* (forest) and *borgo* (village). These also show a quit distinct perception profile, shown in the respective columns of Table 5.1. This can also be observed in the last two columns of the table indicating the degree in the subgroup assessment graph (see below): The subgroups described by *borgo* and *bosco* are quite isolated.

In order to analyze subgroup relations with respect to the perceptions, we apply the Manhattan similarity as our assessment relation rel. We measure the similarity using the averaged perception vectors of the respective subgroup patterns, with normalized values in the interval $[0; 1]$. Using the Manhattan distance, we consider the overall "closeness" of the vectors; alternatively, the cosine similarity would focus on similar perception "profiles", i. e., uniformly expressed perceptions.

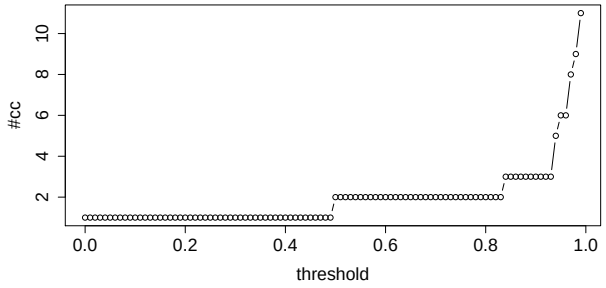


Figure 5.5: Thresholded connected component plot based on a minimal rel value.

For determining appropriate thresholds τ_{rel}, Figure 5.5 shows a threshold vs. connected component plot using the Manhattan similarity defined above. Then, appropriate thresholds can be selected by the analyst. As can be observed in Figures 5.6-5.7 the respective networks for thresholds 0.90 and 0.95 show a distinct structure. Starting with the lowest threshold $\tau_{rel} = 0.90$ we can already observe the special structure of patterns 12 and 15. At this level, the remaining graph stays connected. With threshold $\tau_{rel} = 0.95$, several clusters emerge – the "Heathrow cluster" $(5, 6, 7)$, as well as the large cluster covering most of the *lower noise* patterns. However, this cluster also contains some patterns from the *higher noise* patterns (3, 8, 9, 10), which are rather unexpected and therefore quite interesting for subsequent analysis. The connecting subgroup patterns can then be simply extracted by tracing the connections in the graph.

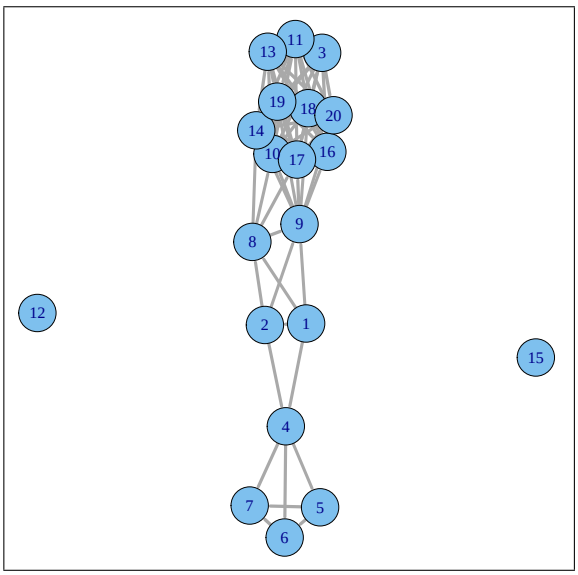


Figure 5.6: Assessment graph: $\tau_{rel} = 0.90$.

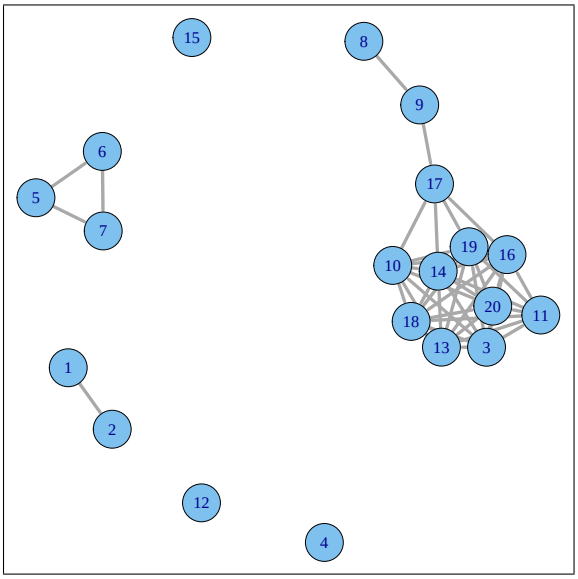


Figure 5.7: Assessment graph: $\tau_{rel} = 0.95$.

5.4 The AirProbe web-game

In this and in the following sections we report general statistics for data gathered through the AirProbe web game. This web application, implemented on the XTribe platform, allowed to gather geolocalized subjective opinions from our volunteers during the AirProbe International Challenge. The dataset we analyze here contains data from October 21st, 2013 to December 23rd, 2013. We report in Table 5.2 and Table 5.3 some general statistics on user participation.

With more then $80,000$ annotations (i.e., both new and edited airpins) and more than 300 participants in the four cities, the case study provides sufficient material to successfully analyze and study the underlying opinion dynamics. The only exception is the city of Antwerp where a modest dataset has been gathered, so that in some of the following analyses those data will be put aside. In the next sections we report a set of basic analyses, focused on the main entities of the web game.

5.4.1 Players and session

Players overall activity

We report in the left part of Fig. 5.8 the number of daily users during the period covered by our dataset. The experiment lasted 6 weeks, 2 weeks for each phase, with additional few more days of data collected after the end of the case study. We can see that the participation has been practically constant (o slightly decreasing) for the whole duration of the experiment. In order to understand whether users were faithful or continuously replaced by new users we investigated their playing

Table 5.2: General stats. (DoP means day of play, i.e. the number of all the day of play of each players.)

Phase	Last Day	Users	Sessions	DoP	Tiles	AirPins Add. (Mod., Del.)	AirSquares
1	11-03	320	2648	1318	1503	35387 (1755, 147)	0
2	11-17	132	1605	967	801	27614 (4545, 532)	0
3	12-01	105	1206	810	245	9998 (6552, 794)	2939
Extra	12-23	64	332	220	42	667 (158, 41)	175
Tot	12-23	341	5780	3315	2591	73666 (13010, 1514)	3126

Table 5.3: Cities stats. (DoP means day of play, i.e. the number of all the day of play of each players. AP means AirPins, AS AirSquares.)

Phase	City	Last Day	Teams	Users	Sessions	DoP	Tiles	AP Add.	Mod.	Del.	AS
1	Antwerp	11-03	2	13	55	47	36	193	17	1	0
2	Antwerp	11-17	2	7	51	39	23	499	24	14	0
3	Antwerp	12-01	2	3	28	27	13	473	14	300	34
Extra	Antwerp	12-23	2	5	33	29	4	44	3	19	12
1	Kassel	11-03	6	55	572	296	351	9325	427	35	0
2	Kassel	11-17	6	37	724	334	371	16326	2532	215	0
3	Kassel	12-01	5	31	529	289	143	5371	2274	169	1623
Extra	Kassel	12-21	5	21	121	72	25	423	20	1	56
1	London	11-03	10	137	1004	513	586	13757	760	56	0
2	London	11-17	10	55	550	419	266	6825	1050	272	0
3	London	12-01	10	42	396	329	47	2851	2483	297	670
Extra	London	12-19	8	28	112	68	11	183	98	19	47
1	Turin	11-03	11	65	781	301	425	11333	517	44	0
2	Turin	11-17	9	33	280	175	141	3964	939	31	0
3	Turin	12-01	8	29	253	165	42	1303	1781	28	612
Extra	Turin	12-19	3	10	67	51	2	17	37	2	60

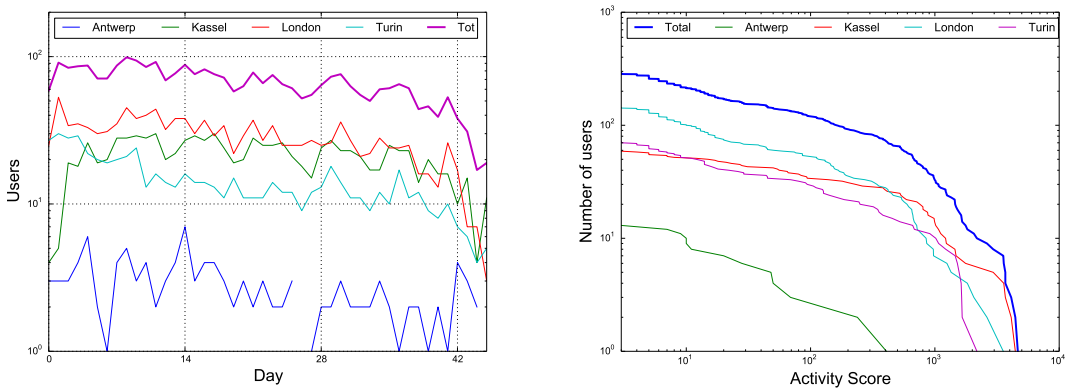


Figure 5.8: On the left, the number of active users for each day of the experiment starting from 09:00 of 2013-10-21. On the right, the Activity Score cumulative graph. The Activity Score is defined for each user as the number of actions performed in the game (counted actions are: game start; revenue, bonus and achievements claim; AirProbe purchase, edit or delete; Tile or AirSquare purchase). For each value of the Activity Score, the graph shows the number of users with a greater score.

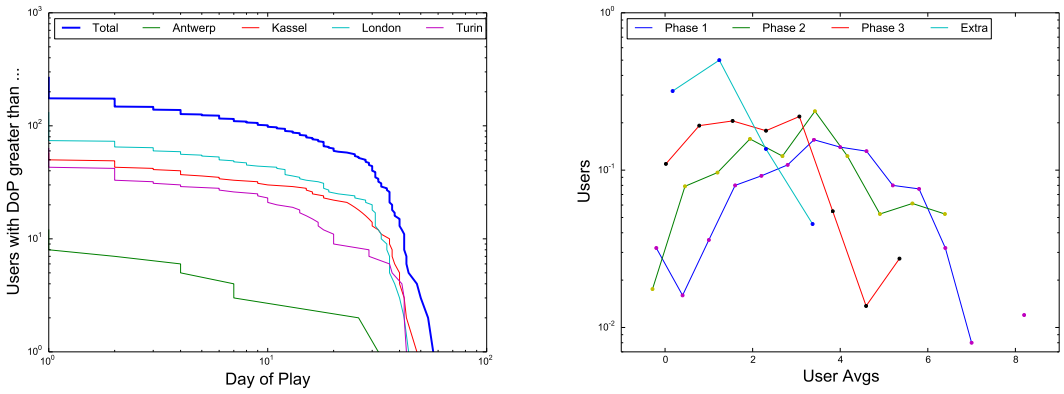


Figure 5.9: On the left, the Day of Play (DoP) cumulative graph. For each value of the DoP, the graph shows the number of users which played at least that number of day. On the right, the distribution of users averages (how many user had a certain average) for each phase.

habits. The right part of Fig. 5.8 shows user activity. We defined an Activity Score for each user which consists of the number of actions performed by a user during her game history. Actions accounted for are: game start; revenue, bonus and achievements claim; AirProbe purchase, edit or delete; Tile or AirSquare purchase. The mentioned graph reports on the x axis the Activity Score and in the y axis the number of users with an Activity Score greater than the relative x axis value. It is fairly visible in all curves (except Antwerp) a change in the slope, approximately at the score value 10^3. This threshold helps us to understand the composition of our ensemble of players. At a global level, ~ 100 of players were very active, having performed thousands of actions, while the rest of the users were more 'occasional players'. The existence of a significative ratio of motivated users is exactly what we needed in order to monitor the evolution of the opinion during the experiment. In order to confirm the existence of a core of users that actually played for almost all the case study, we elaborated a cumulative Day of Play graph, reported in the left part of Fig. 5.9. We calculated for each user the number of days in which she played at least once (DoP). Then, similarly to what we did in the right part of Fig. 5.8, we reported on the x axis the DoP and on the y axis the number of users with a DoP equal or greater. Again, it is quite visible a sort of threshold more or less near 30 DoP (the whole case study lasted 42 days), with a corresponding value of ~ 100 users. We can thus affirm that, in the ensemble of our users, we managed to gather an important core of players that actively contributed to the case study with constance and motivation.

Basic statistical analysis

In this preliminary analysis, we simply determined the average value of the AirPins for each day, neglecting the spatial positions, and then we measured the corresponding distribution in each phase. In the right part of Fig. 5.9 we report the distributions. We observe a noticeable shift in the distributions, pointing out that the opinion is changing coherently. We shall analyze in the following sections the causes, the dynamics and the consequences of this shift.

Once that the nature of our community of volunteers is clear, we can analyse their in-game habits. We define a session of play as the consecutive time the user spent with a browser open on the game interface page. The session time is measured as the difference between the timestamp of the login and the timestamp of the last action. This loose definition does not distinguish between the active playing time and the time the game window has been left open in the background, so it is not a precise measure of the attention dedicated by our users, so that we will consider the session

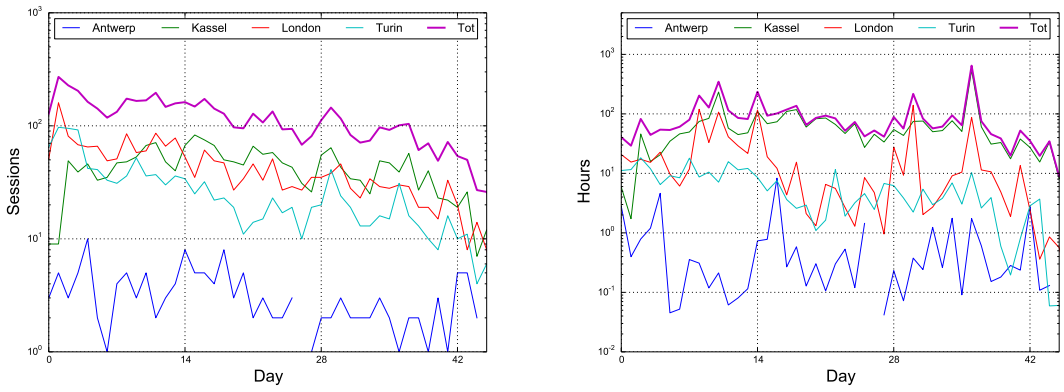


Figure 5.10: On the left, the daily number of sessions. On the right, the daily number of hours of play.

time a sort of proxy for it.

In the left part of Fig. 5.10, we report the number of sessions registered during each day of play. This number is decreasing during the experiment, probably because of the loss of users that can be observed in the left part of Fig. 5.8. This loss of players is compensated by the large motivation of the the remaining users. This is quite evident in the right part of Fig. 5.10, where we consider the total time spent on the game each day, calculated as the sum of all the duration of the sessions for that day. This number stays substantially constant for the whole case study, witnessing an essentially uniform effort for all the challenge duration.

5.4.2 Game dynamics

The main interaction of players with the game is through the AirPin (AP), which is the expression of players opinions about air pollution, annotated on a map. Players, after buying tiles of the map with virtual credits, had to place their estimations of air pollution in terms of μg/m^3 of black carbon. This annotation (AirPin) was performed by clicking on a point of the map sufficiently far from previous annotations, and by specifying the desired value with a slider. The daily number of APs added, modified or removed, is reported in Fig. 5.11, respectively in the top, in the bottom left and in the bottom right part. The magnitude of the daily AP input flow is around 10^3 for the whole challenge fluctuating a lot. The daily quantity of AP modifications or deletions gives us hints about the annotation behavior of players. Both pictures, in fact, show peaks after day 14 and day 28, when the phase and the rules of the game were modified. In phase two the revenues were harder to get, while in phase three AirSquares were introduced. The two peaks show that players acknowledged that something changed in the game and adapted their annotations to the new game conditions. The value assigned to APs could be chosen between 0.00 and 10.00 μg/m^3 and could be selected continuously with a slider. The general usage of these values in the three phases is reported in the top left part of Fig. 5.12, while the other graphs report the situation for Kassel, Turin and London (clockwise). In the first phase, we observe an important peak located in the middle of the scale. In each city we observe how players estimate center around 5 μg/m^3 (only in Kassel we detect an additional peak near 0, probably because of the presence of wide gardens included in the game area). This trend can be explained by considering that the proposed measurement unit of AirPin values was unfamiliar for almost all players. Therefore, when guessing values on an unknown scale, it seems reasonable to be attracted toward the center of it. This choice is not correct in this case, since black carbon concentrations in urban areas are in average around 2 μg/m^3. In fact, in each graph of Fig. 5.12 we can observe how, in the following phases,

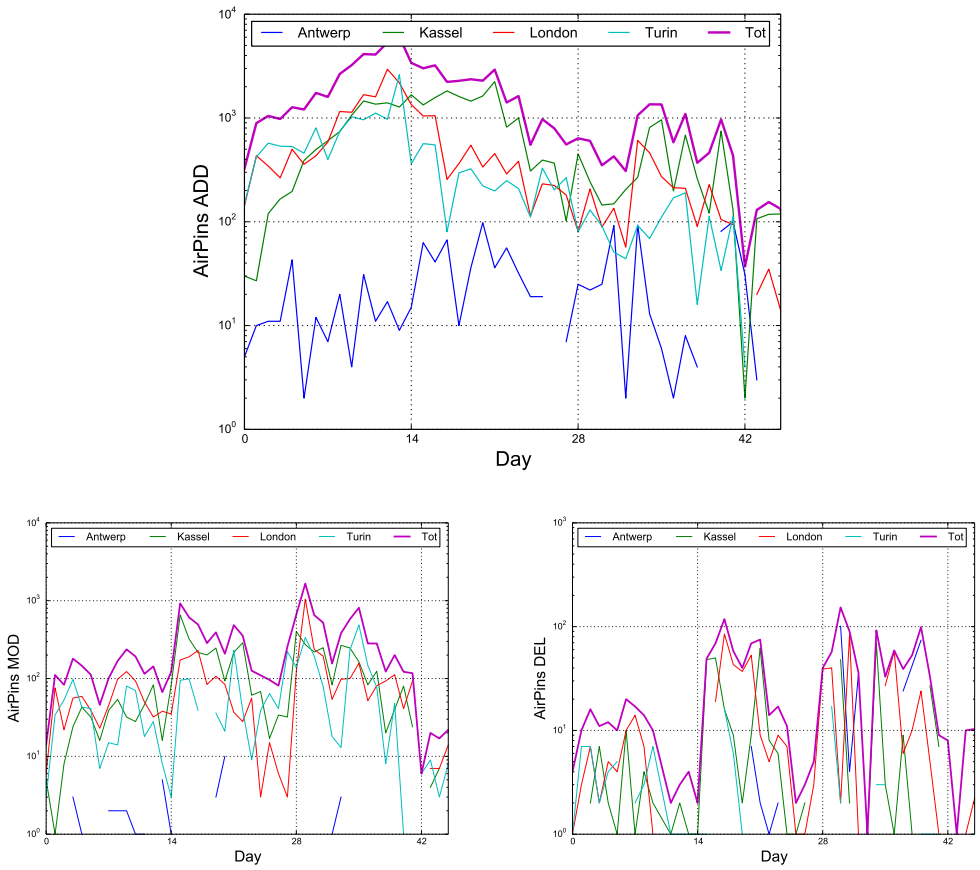


Figure 5.11: On the top part, the daily number of AP added. On the bottom left and right, respectively, the daily number of AP modified or deleted.

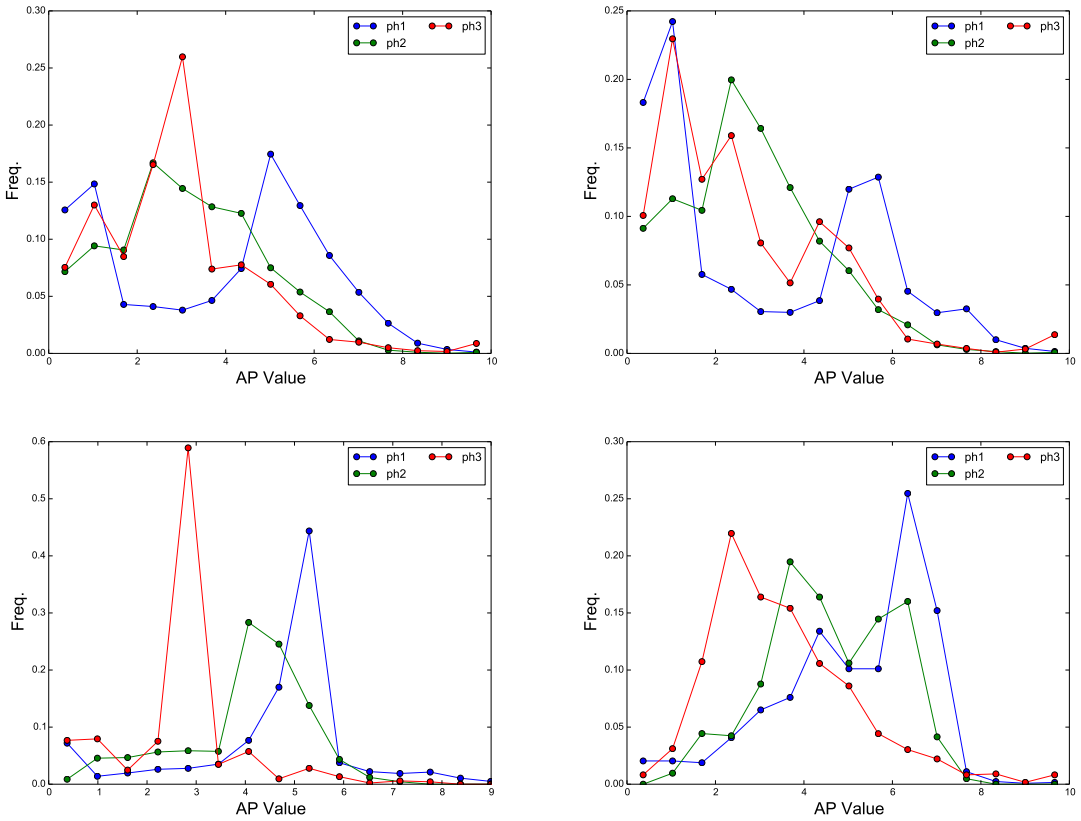


Figure 5.12: Clockwise, from the top left: the usage of the scale in the overall, for Kassel, for Turin and for London in each phase of the challenge.

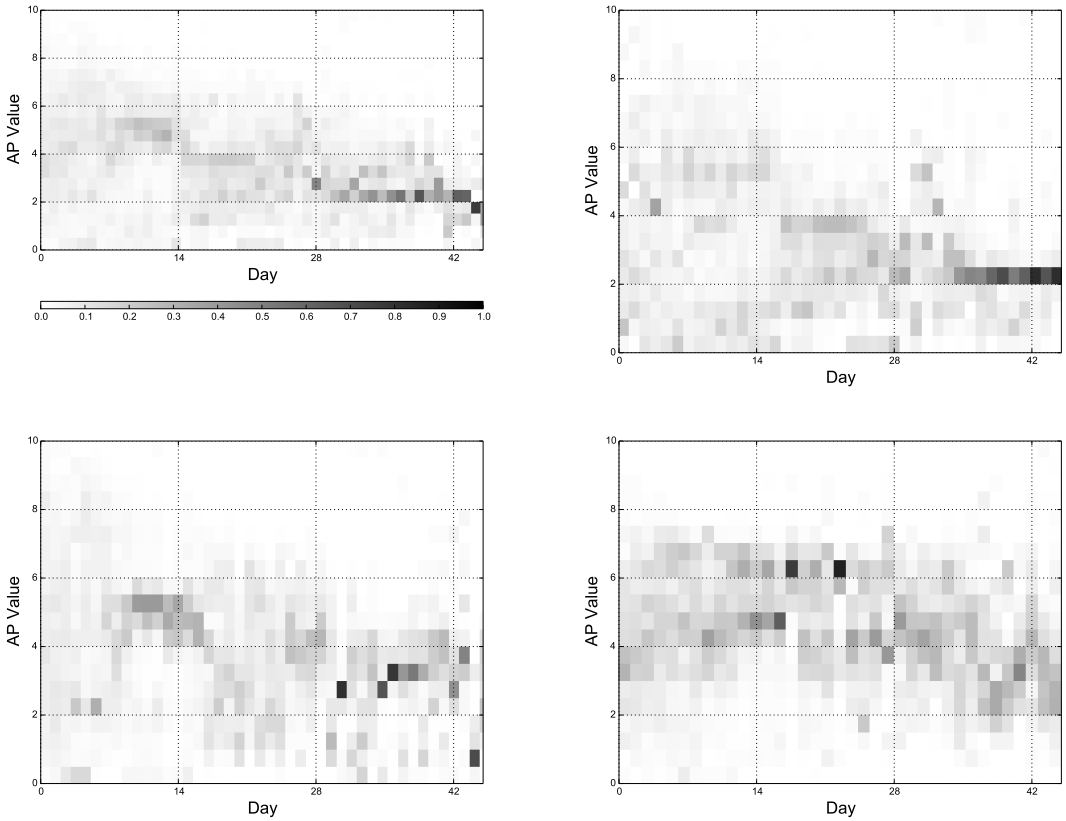


Figure 5.13: Clockwise, from the top left: the daily density graph in the overall, for Kassel, for Turin and for London. In these graphs, each column represents the usage density histogram of the scale for a given day. The color corresponds to the ratio of opinions in the corresponding bin (0.0 is white, 1.0 black). Bins size is $0.5\ \mu g/m^3$.

where the actual measurements collected on the field started to seep out, the distributions shift towards lower values. By comparing the figures in Fig. 5.12 it is also interesting to note that in phase one the city of Turin is perceived by its citizens as the most polluted, followed by London. This was an expected result, since Turin is known to be one of the most polluted European city, due to industrial production, urban traffic and a particular conformation of mountains and hills around. The evolution of the distribution of the air quality opinion is elaborated in Fig. 5.13 as density graphs. These graphs show how the daily opinion in the challenge for Kassel, Turin, London and in total is distributed. We observe that the average of the distribution shifts toward lower AP values at the beginning of each phase. Also, a sort of collapse of the distribution is visible for the last phase, where citizens possibly learned the black carbon concentration values of their environment. We would like to point out that Fig. 5.13 is portraying the evolution of the perception of the air quality, but we are mainly interested in the relation between the perception and the *real* values. More precisely, we are interested in the difference between estimated black carbon concentrations and sensor box measurements, to which players were exposed by buying AirSquares. Moreover, users have the opportunity to be informed on values probably perceived as the true values, even though in aggregate form (AirSquares show the average between the measures of the sensor boxes in a certain area). Are we sure that they are actually learning something instead of just copying? To understand this we need to monitor the evolution of the difference between values of AirPin added (or modified) and the value of the AirSquare containing it. This analysis will be thoroughly described in Deliverable 5.2. Here, we report the heat map of the AirPin values in the

three phases for Kassel, London and Turin in Fig. 5.14, Fig. 5.15 and in Fig. 5.16 respectively. As already pointed out, the overestimation of pollutant concentrations in phase one is detectable for all cities. Players located the pollution mainly on main roads and crossroads, while gardens and rivers where perceived as cleaner. In phase three, i.e. as soon as the AirSquare values are made available, they changed opinion substantially. This clearly denotes that they were prone to change their mind. In deliverable 5.2 we will try to understand whether this opinion shift simply follows the AS values or is a consequence of a sort of *virtual awareness*.

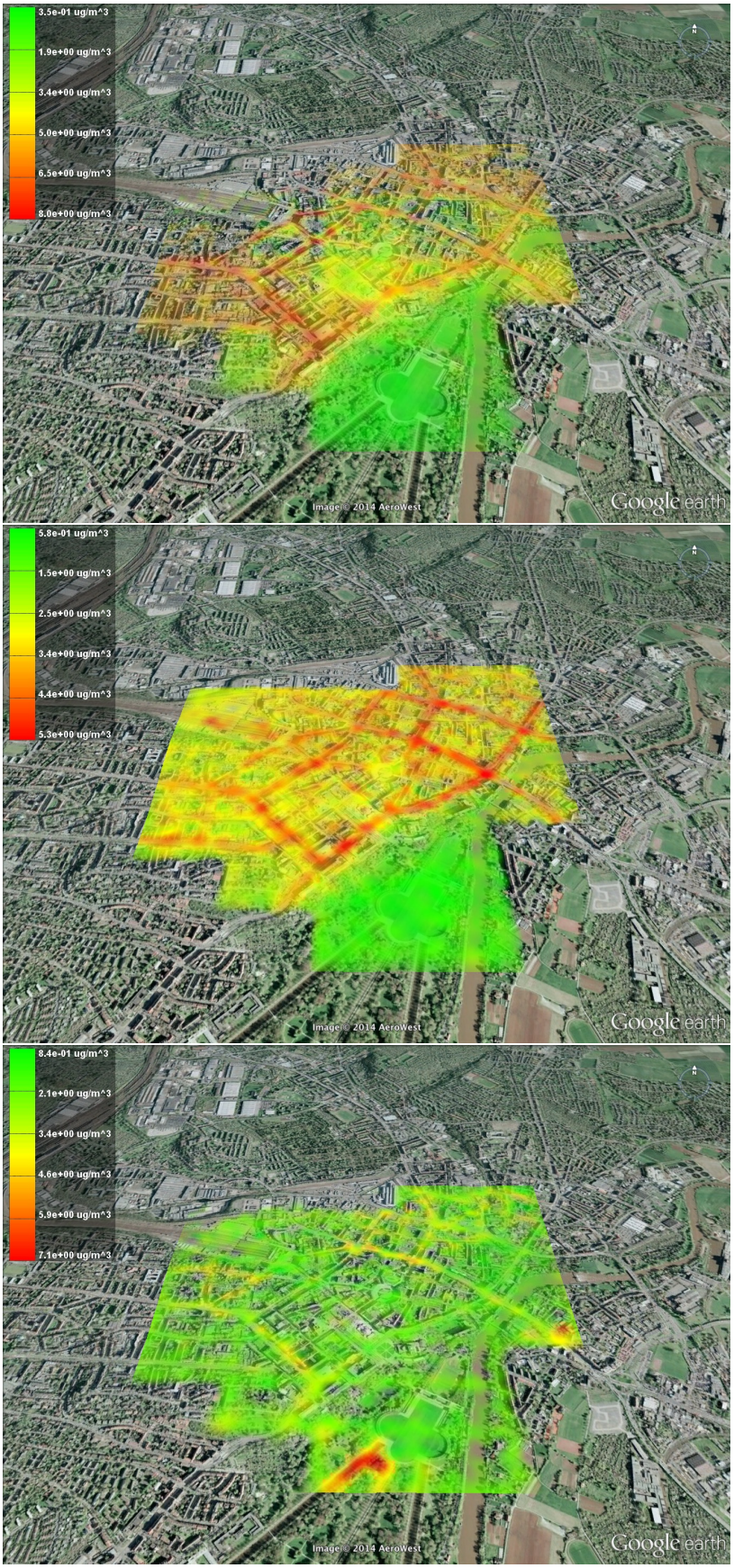


Figure 5.14: From the top, for phase 1, 2 and 3: the heat map of AP values for Kassel. Values in the key are, as usual, $\mu g/m^3$ of Black Carbon. The opacity is an related to the number of AirPins in that point.

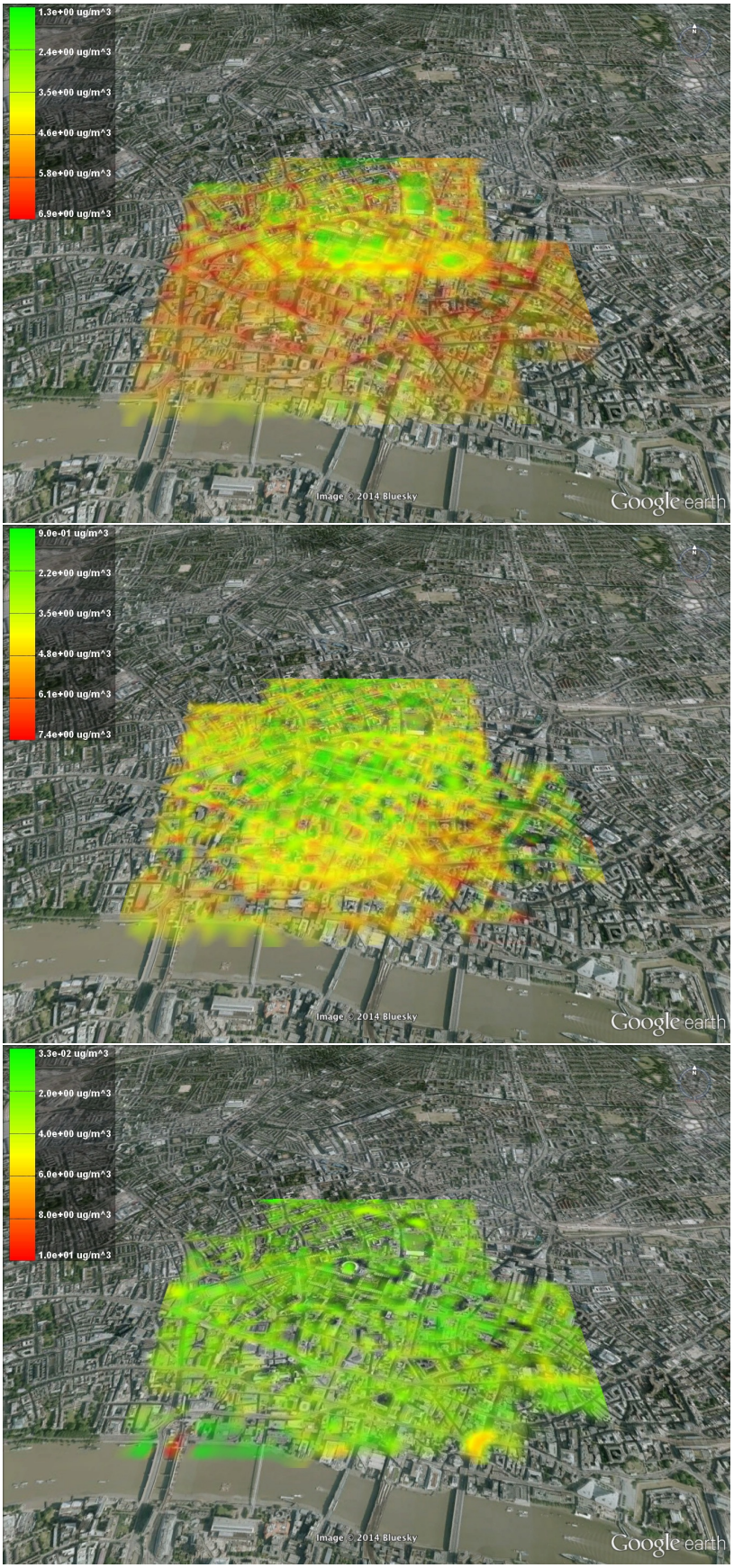


Figure 5.15: From the top, for phase 1, 2 and 3: the heat map of AP values for London. Values in the key are, as usual, $\mu g/m^3$ of Black Carbon. The opacity is an related to the number of AirPins in that point.

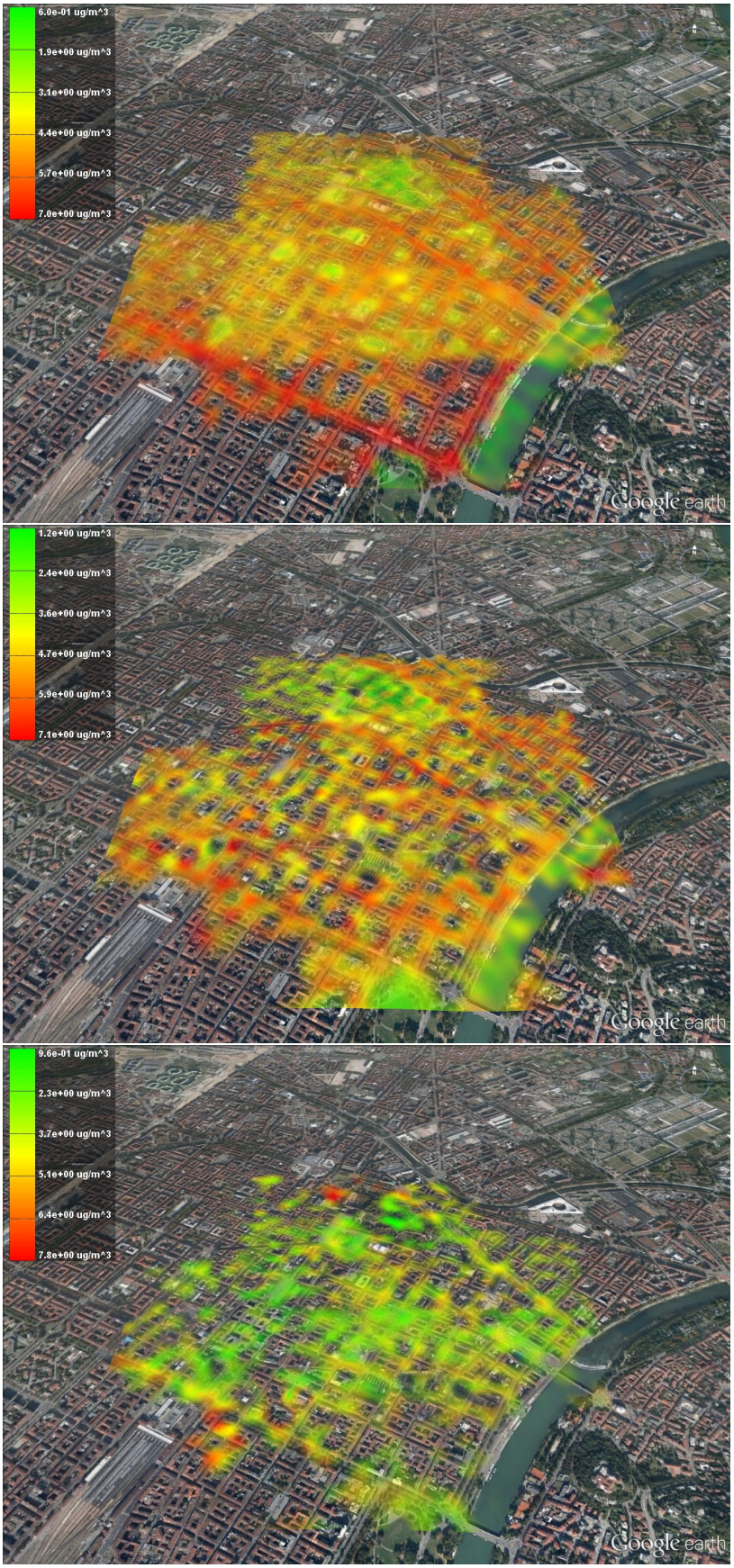


Figure 5.16: From the top, for phase 1, 2 and 3: the heat map of AP values for Turin. Values in the key are, as usual, $\mu g/m^3$ of Black Carbon. The opacity is an related to the number of AirPins in that point.

Bibliography

Rabeeh Abbasi, Marcin Grzegorzek, and Steffen Staab. Large scale tag recommendation using different image representations. In *Semantic Multimedia: 4th International Conference on Semantic and Digital Media Technologies, SAMT 2009*, volume 5887 of *Lecture Notes in Computer Science*, pages 65–76. Springer, 2009. ISBN 978-3-642-10542-5. doi: 10.1007/978-3-642-10543-2_8.

Martin Atzmueller and Florian Lemmerich. Fast subgroup discovery for continuous target concepts. In *Proc. ISMIS 2009*, LNCS, 2009.

Martin Atzmueller and Florian Lemmerich. Exploratory pattern mining on social media using geo-references and social tagging information. *IJWS*, 2(1/2), 2013.

Martin Atzmueller, Martin Becker, Stephan Doerfel, Mark Kibanov, Andreas Hotho, Björn-Elmar Macek, Folke Mitzlaff, Juergen Mueller, Christoph Scholz, and Gerd Stumme. Ubicon: Observing social and physical activities. In *Proc. IEEE CPSCom*, 2012.

Martin Becker, Saverio Caminiti, Donato Fiorella, Louise Francis, Pietro Gravino, Mordechai Haklay, Andreas Hotho, Vittorio Loreto, Juergen Mueller, Ferdinando Ricchiuti, Vito D. P. Servedio, Alina SÃőrbu, and Francesca Tria. Awareness and learning in participatory noise sensing. *PLOS ONE*, 2013. DOI: 10.1371/journal.pone.0081638.

Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November 2002. ISSN 0924-1868. doi: 10.1023/A:1021240730564.

EveryAware. Report on the everyaware platform performance in the pilot studies, deliverable d3.1, 2014a.

EveryAware. Report on data coverage and interpolation methods, deliverable d4.1, 2014b.

EveryAware. Final report on participation fostering activities, deliverable d6.3, 2014c.

Kanti V. Mardia, John T. Kent, and John M. Bibby. *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press, 1. edition, 1979. ISBN 978-0-12-471252-2.

Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors. *Recommender Systems Handbook*. Springer, 1. edition, 2011. ISBN 978-0-387-85819-7. doi: 10.1007/978-0-387-85820-3.