Pervasive and Mobile Computing 9 (2013) 681–694

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Participatory noise mapping works! An evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring

Ellie D'Hondt [a,*], Matthias Stevens [b], An Jacobs [c]

[a] BrusSense, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[b] ExCiteS group, University College London, Gower Street, London WC1E 6BT, UK
[c] IBBT-SMIT, Vrije Universiteit Brussel, Pleinlaan 9, 1050 Brussels, Belgium

ARTICLE INFO

Article history:
Available online 4 October 2012

MSC:
68U99
68N01
62P12

Keywords:
Environmental policy
Noise mapping
Participatory sensing
Citizen science
Mobile phones

ABSTRACT

Participatory sensing enables a person-centric collection of environmental measurement data with, in principle, high granularity in space and time. In this paper we provide concrete proof that participatory techniques, when implemented properly, can achieve the same accuracy as standard noise mapping techniques. We do this through a citizen science experiment for noise mapping a 1 km^2 area in the city of Antwerp using NoiseTube, a participatory sensing framework for monitoring ambient noise. At the technical side, we set up measuring equipment in accordance with official norms insofar as they apply, also carrying out extensive calibration experiments. At the citizen side, we collaborated with up to 13 volunteers from a citizen-led Antwerp-based action group. From the data gathered we construct purely measurement-based noise maps of the target area with error margins comparable to those of official simulation-based noise maps. We also report on a survey evaluating NoiseTube, as a system for participative grassroots noise mapping campaigns, from the user perspective.

1. Introduction

Pervasive computing naturally lends itself to many of environmental sustainability's core challenges. For one thing, it provides numerous possibilities for better monitoring the physical world [1], by way of miniaturised sensors which are becoming ever more pervasive, performant and cheap. At the same time pervasive technologies are inherently connected to their users, which means that they can play an important role in awareness building. The realisation that pervasive technology can provide a direct link between data and people has crystallised into the idea of *participatory sensing*, which appropriates everyday mobile devices such as cellular phones to form interactive sensor networks that enable public and professional users to gather, analyse and share local knowledge [2,3]. Its potential for data collection at a high granularity in space and time, very difficult for scientists to achieve otherwise, together with its potential for citizen involvement at all layers of society, make this an extremely promising tool for monitoring and managing the environment [4,5]. One particularly interesting application of these ideas is to construct participatory, measurement-based environmental maps. Indeed, while there are already important efforts going on to monitor the effects of pollution, these are mostly in terms of computer models and urban statistics, while very little actual measuring is involved. The resulting maps give an average but not at all a complete view on the situation, entirely missing local variations. Crowdsourcing through participatory sensing

* Corresponding author.
E-mail addresses: Ellie.DHondt@vub.ac.be (E. D'Hondt), Matthias.Stevens@vub.ac.be (M. Stevens), An.Jacobs@vub.ac.be (A. Jacobs).

1574-1192/$ – see front matter
doi:10.1016/j.pmcj.2012.09.002

techniques alleviates these issues and allows for a person-centric approach to gathering and visualising actual data on a much larger scale.

Any successful participatory sensing application lies at the nexus of three essential ingredients: *technology*, *data* and *people*. For environmental monitoring the data aspect is of particular importance. Indeed, already today there is much demand from citizens, activist groups and even city administrations for using participatory technologies to tackle local pollution issues.[1] Of course this has only been possible because the underlying technologies have been much publicized, in academic literature [6,7] but also in the media. However, it is an open question whether the data gathered with these technologies is at all accurate. This is crucial to the cause of anybody wishing to set up environmental monitoring actions, as inaccurate data would not only affect the action involved but also the success and credibility of any future actions. The issue is especially sensitive because authorities often assume, out of convenience or out of conviction, that data collected by citizens by means of participatory sensing techniques are not nearly as accurate as data gathered by officials using conventional assessment methods. This article focuses on determining what exactly the data quality is that one can achieve, given a concrete case (noise), an operational participatory sensing system (NoiseTube), and an actual measurement campaign designed with these issues in mind. We pay special attention to standards, norms and conventional noise assessment methods, with the aim of giving a first qualitative comparison between the existing and proposed approaches. We stress that our current focus is not on the technology, nor on large-scale deployment, which do not make sense before the issue of data quality is settled.

We choose noise as a concrete case for several reasons. First, noise pollution is representative both in that it is tightly correlated to other types of pollution and in terms of the structure of the regulatory framework behind it. Second, it is a very actual topic of concern for citizens and authorities, and indeed there are currently important efforts going on to monitor its extent and assess its effects [8,9]. Third, there has been considerable progress in technologies for participatory noise monitoring, with applications such as EarPhone [10], NoiseSPY [11], WideNoise and our in-house NoiseTube application [7,12] taking the lead. Our choice for working with NoiseTube is motivated not only by pragmatic reasons but also because as the only open-source, publicly available project, targeting several platforms and involving not only a mobile application but also a web-based community platform, we feel it is the most complete solution to date [12]. Tried and tested on an individual-user level, it allows us to focus on the data quality aspect rather than on the technology.

The structure of this paper is as follows. We first give an overview of what official noise pollution norms and assessment techniques encompass in Section 2. In Section 3 we set up our measuring equipment according to these norms, and report on extensive calibration experiments and subsequent testing of our equipment, in and outside the lab. This section already provides half of the work towards profiling participatory sensing data quality, as it thoroughly improves the accuracy of mobile phones as sound level[2] meters (by correcting for systematic errors), while also assessing their precision as stand-alone devices (by evaluating the spread of random errors). Next we report on the various implementation steps for a 2-phase realistic citizen science experiment for noise mapping a 1 km^2 area, and analyse the resulting maps. To do this we collaborated with 13 volunteers from a citizen-led activism group, equipping each of them with a mobile phone for measuring and allowing extensive feedback sessions on results, experiences and usability. This coordinated mapping campaign allows us to answer the other half of the question on data quality, namely whether large datasets can make up for the inferior accuracy of mobile phones with respect to professionally used equipment. The outcome of our experiment is a set of purely measurement-based, statistically analysed noise maps of the target area, comparable to official noise maps in terms of data quality. We also include a qualitative comparison with the latter, highlighting similar patterns as well as differences and arguing how these underline the credibility of our approach. Although participatory noise maps have been reported earlier in literature [10,11], they are either much more limited in scale or they do not take the data quality aspect nearly as serious as we do in the work presented here. Finally, because adoption by the wider public is so important for participatory techniques to take off, we organised a user survey of which results are presented in Section 5. While the group is too small to draw any kind of general conclusion, the feedback gathered is still extremely useful to fine-tune future experiments. To conclude, we discuss the validity and value of participatory noise maps in Section 6, where we also provide general guidelines on noise mapping protocols and report on our experience with extending the participatory approach to the level of large-scale adoption.

2. Noise mapping today

The European Union is at the forefront when it comes to noise assessment [13]. Indeed, there has been a tradition in noise mapping for many years, at least at the level of individual member states. On the basis of this tradition European-wide regulations were defined in the form of European Union Directive 2002-49-EC [8]. This directive has in turn incited further developments in noise assessment methods, and has essentially led to an inventory of noise maps for large cities throughout Europe. The United States, though its Noise Control Act dates back to 1972, does not mandate noise mapping at a federal or

[1] This is obvious from our own experience, as we are regularly contacted by these actors for setting up actions, and from the success of organisations such as Mapping for Change and the Public Laboratory for Open Technology and Science (PLOTS).

[2] Sound level is the physical parameter measured, while noise, i.e. unwanted sound, is the term more commonly used in environmental assessment contexts.

state level. As a result it has been slower to catch on to regional noise mapping, and community-wide noise maps are rare. Nevertheless, the same methods that are applied in Europe can be implemented in the context of the United States, and this is becoming a reality at the time of writing [13]. The situation for other developed countries, such as Canada or Japan, are similar to that of the United States, while developing countries typically lack national legislation for noise assessment and even noise control. For these reasons, and because our research is based in Europe, we take the assessment methods which lie at the basis of the European Noise Directive (END) to be the norm. Below we describe in short what this directive entails, so as to provide policy context to our research, and, more importantly, place our approach within that context.

The END dictates that as of 2007, cities with a population of over 250,000 have to estimate the number of citizens exposed to average yearly sound levels of 55–75 decibel (dB), in bands of 5 dB, and over 75 dB, and this at 4 m above the ground on the most exposed facade of their home. Separate numbers are required for road, rail and air traffic and for industrial sources, where only major roads (>6 million vehicle passages per year), railroads (>60,000 train passages per year) and airports (>50,000 movements per year) should be considered. Sound levels obtained are to be presented graphically on strategic noise maps. Exposure numbers should be updated every 5 years and moreover, as of 2012, smaller cities (> 100,000 inhabitants) are also to be taken into account. The most visible part of the implementation of this directive has been the creation of strategic noise maps. These show the aforementioned sound levels as a colour-coded overlay on city maps. More concretely, the main noise indicators for noise mapping are L_{day}, L_{evening}, L_{night} and L_{den} (day–evening–night), obtained by averaging sound levels for corresponding time intervals over the period of one year. To be more precise, the sound level required is the so-called *A-weighted equivalent continuous sound pressure level* $L_{\text{A,eq}}$, the quantity put forward as the international standard for environmental noise assessment [14]. $L_{\text{A,eq}}$ is itself an average, this time of instantaneous air pressure differences (i.e. sound) over arbitrary periods of time:

$$L_{\text{A,eq}} = 10 \cdot \log_{10} \frac{1}{T} \int \frac{p^2}{p_0^2} dt \quad [\text{dB(A)}], \tag{1}$$

where p_0 is the reference sound pressure, taken as 20 μPa. The subscript A denotes that the equivalent sound pressure level is to be weighted with the *A-weighting curve*, which approximates the sensitivity of the human ear to different frequencies (low and high frequencies are attenuated). By choosing time intervals appropriately equivalent sound pressure levels can be computed for periods of interest, such as peak hours or during the night time. In this way, L_{day}, L_{evening} and L_{night} are found by taking the following time intervals in Eq. (1): a 12 h day period, a 4 h evening period, and an 8 h night period respectively.[3] The composite day–evening–night sound pressure level L_{den} is a logarithmic,[4] weighted average over these three periods, as follows,

$$L_{\text{den}} = 10 \cdot \log_{10} \left(\frac{1}{24} \left[12 \cdot 10^{\frac{L_{\text{day}}}{10}} + 4 \cdot 10^{\frac{L_{\text{evening}}+5}{10}} + 8 \cdot 10^{\frac{L_{\text{night}}+10}{10}} \right] \right) \quad [\text{dB(A)}]. \tag{2}$$

By adding 5, respectively 10, to the evening and night values one accounts for the increased annoyance caused during such periods. In summary, the maps that are produced as a result of the END present L_{den} and L_{night} values, for train-, traffic-, airport- and industry-related sources of noise, in colour-coded bands of 5 dB(A). Though not mandatory, sometimes a composite map for all sources considered is provided. Note that the END states specifically that "the values of L_{den} and L_{night} can be determined either by computation or by measurement" [15, Annex II.2]. In practice, however, sound levels are always computed because of scalability issues—indeed, it has hitherto been unfeasible to consider measurement devices at all places and times. Instead, sound maps are simulated on the basis of statistics on sources present, source-specific propagation rules and information on the urban layout. This approach is in fact embodied in the established norms which per construction are phrased in terms of four types of sources of sound, and then only those sources larger than a given size. Measurements are only used as a means to initialise the model, for which data is gathered from professional-grade, highly accurate equipment typically mounted on rooftops or next to specific sources of noise in representative areas of the city. As an example, the Brussels Region, which is over 160 km^2 large and counts over one million inhabitants, relies on 19 measuring stations (one of which is mobile).

The simulation approach has been extremely useful as it has allowed assessment of background noise in the absence of physical data of adequate granularity. However, by focussing on 4 sources of sound only, one ignores all incidental sounds such as those originating in construction works, traffic jams, passage of emergency services, manifestations, or neighbourhood noise, all of which are known to be highly annoying to city dwellers. Simulated maps are not only incomplete in acoustic terms but also geographically (large roads only) and temporally (average day and night in the year). They also lack contextual factors, to which individual sound experience is highly sensitive, and typically date back a number of years. Simulated maps are useful for authorities to obtain an idea of global trends in the urban soundscape, thus providing an indication of actual citizen exposure. They do not, however, capture person-centric exposure levels or model local variations very well. For this reason they have remained of little interest to citizens, who either perceive the obvious (i.e. that his street is loud), or, worse, cannot link the mapped area to their own experience (street not considered, local variations ignored).

---

3 Typically a 24 h period is split up as follows: 7 am–7 pm (day), 7–11 pm (evening), 11 pm–7 am (night), though member states may define these periods independently to account for local and cultural differences.

4 Logarithmic averages are used when computing the average sound level over a period in terms of sound levels of consecutive time intervals of that period.

3. Mobile phones as sound level meters

If we are to take participatory sensing seriously as a method to produce noise maps, the first thing to do is to evaluate the equipment used. Moreover, if we want to compare results to official, simulation-based noise assessment methods, it is essential to play by the same rules insofar as possible. As mentioned above, the END allows the use of measurement equipment to establish sound levels directly [15]. The norms for how these measurements should proceed are again very much written with simulation models in mind, providing separate methods for each source considered, and dictating the measurement of quantities that are directly compatible with propagation formulae. For example to assess traffic noise, measurement equipment should involve: (1) A-weighting, (2) direct read-out of sound levels in dB(A), (3) computation of $L_{A,eq}$ over arbitrary time intervals, (4) calibration, (5) spherical sensitivity and wind protection, (6) wind speed, (7) wind direction, (8) speed of passing vehicles and (9) measurement at 4 m above street level. Requirements (1)–(5) correspond to the international standard for sound level meters (SLMs) [16]. Requirements (6)–(9) are tailored to the simulation model, allowing to compute meteorological corrections and propagation parameters related to vehicle speed, as well as ensuring compatibility with computed sound levels. In a participatory setting facilitating easy adoption is crucial and the focus is on personal exposure levels during daily life. For this reason requirements (5)–(9) do not make sense: we want all users to be able to carry out measurements with the equipment they already own (ruling out wind and speed measurements), and we want to determine what users are exposed to as they traverse their daily lives (i.e. wherever they may find themselves). Requirements (1)–(2) are satisfied through the NoiseTube software, which shows A-weighted equivalent sound levels per second; requirement (3) is achieved by post-processing of these individual values. This leaves us with requirement (4), to which we pay special attention here.

Calibration is a comparative procedure by which one device's measurement readings are compared with those of a trusted, reference device, which is considered to produce correct readings. This gives insight in systematic errors which can then be corrected for. This can substantially increase a device's *accuracy*, the degree of closeness of device measurements of a quantity to that quantity's true value. While calibration is in itself a well-known procedure, it is not typically carried out for mobile phones. For this reason we have put much effort in calibration experiments, determining frequency characteristics as well as white noise characteristics for 11 mobile phones. Although others have proposed systems which, like NoiseTube, allow mobile phones to be used as SLMs, to our knowledge this is the first time calibration of such a system has been carried out so extensively. Earlier work typically targets only one phone with one type of sound, be it a test signal [10] or white noise [11]. An improvement is the work in [17], yet still no systematic overview of the frequency-sound level domain is given. None of the above motivate their choice of calibration method. Our experiments serve as a way to improve the accuracy of mobile phone microphones. This is important in the context of this article because it is an important factor in improving the accuracy of participatory noise maps. But these experiments also serve as a way to obtain insight into what kind of characteristics one can expect and give precise arguments in favour of particular calibration techniques. They are also important in the larger context of participatory environmental sensing in order to counter the oft-heard critique that mobile phones simply do not stand up to the standards of more professional equipment.

Calibration in the lab. For the experiments conducted in this article we used a set of 10 Nokia 5230 mobile phones (referred to as Nokia 5230 #i or just phone #i for the *i*-th phone henceforward), a reasonably cheap model which nevertheless satisfies all requirements for running NoiseTube. Calibrating these phones as SLMs requires three ingredients: a controlled environment to carry out sound level measurements, a sound generator, and a professional SLM to compare our mobile phones against. Experiments were carried out in an anechoic chamber, the preferred environment for calibration. Sound was generated at particular frequencies using an HP Agilent 33 120A waveform generator, and as white noise (an equal composition of all frequencies) using a Brüel and Kjær Type 1405 noise generator, each of which was connected to an amplifier (Brüel and Kjær Type 2706) and a speaker to produce the actual sound in the anechoic chamber. Finally, the reference ("correct") SLM set-up consisted of a MicroTech K250 high-end condenser microphone, an LMS Scadas III data acquisition station, and a Windows PC running the LMS Test.Lab software. Sound generators, LMS station and PC were positioned outside the anechoic chamber, while amplifier, speaker, microphone and mobile phones were positioned inside it. NoiseTube ran continuously on the mobile phones that were being tested, with sound recorded at the highest sampling rate possible, in this case 48 kHz. The general procedure is to produce sound of a particular type for a time interval of one minute, readout "true" sound levels directly from the PC, while those of our mobile phones were obtained by taking the NoiseTube measurements log file, identifying the interval of interest, and averaging over the obtained measurements.

In our experiments we have studied three types of variability: phones, frequencies, and sound levels. While targeting one particular model, we have not acquired our phones in one batch but rather partly new, partly second hand, which constitutes a realistic set with different usage histories and dates of manufacture. Before deciding whether a frequency-independent method (such as white noise calibration) is sufficient, we subjected two of our phones to tests spanning the full frequency range of the human ear. Concretely, we produced pure tones at every third octave band from 50 Hz to 20 kHz, for a total of 27 tones. This sequence was traversed a total of 7 times, taking sound pressure levels at 1 kHz to be fixed at 60–90 dB with steps of 5 dB. When jointly plotting these 7 traversals (as sound levels w.r.t. frequencies) for phone #0, phone #1 and the reference, we noticed the following. First, there is no significant difference between the two phones tested, which led us to focus on the characteristics of one phone only. Fig. 1 compares the results for phone #0 with the reference microphone. Next, there is a clear dependence on sound levels, witnessed by a varying spread between different traversals which is not present in the reference graph. On the other hand, the frequency characteristics of the phone (i.e. the shape of the traversal)

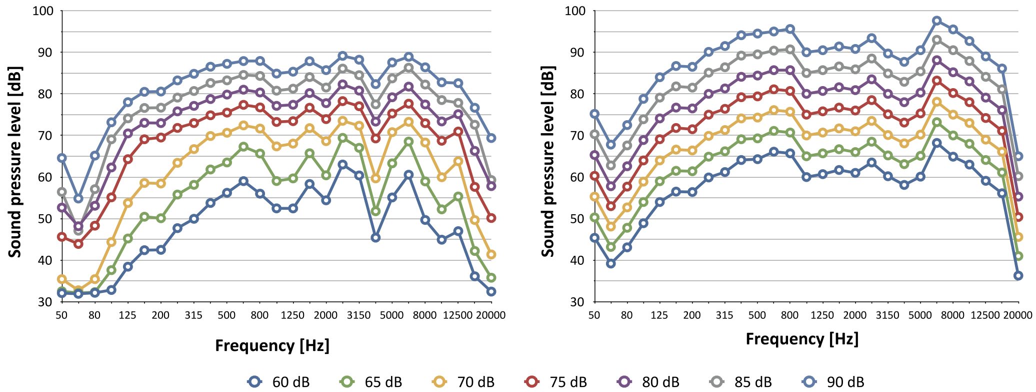

Fig. 1. Outcomes of frequency-dependent calibration experiments, giving dB values in terms of frequency. Nokia 5230 #0 phone on the left, reference microphone on the right.

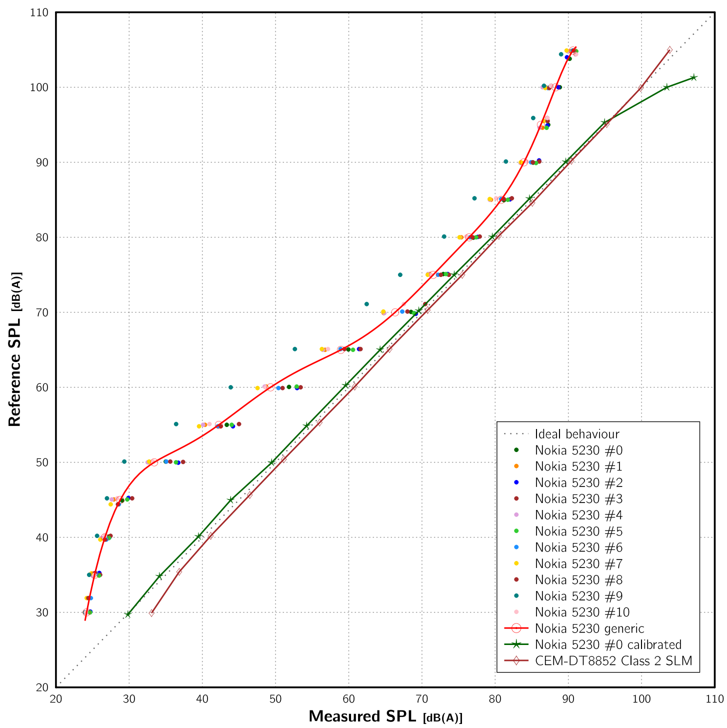

Fig. 2. Calibration points for equipment used. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are quite similar to that of the reference microphone, at least in the domains that are most important in urban areas (i.e. excluding low sound levels and frequencies). Calibrating phones in terms of frequency entails developing digital filters, which is a difficult procedure. This holds all the more in a participatory setting, where one has to consider the fact that few users will be able to calibrate their equipment themselves. But our data shows that we are entitled to take a pragmatic stance, as sound levels have a more significant impact than frequencies do. This justifies switching to a calibration technique which is independent of frequencies, in the sense that one calibrates against a test sound that is a fixed mix of all frequencies in the range considered. Again there are different options, the most common being white noise or pink noise, which has a decreasing frequency spectrum. Pink noise is the spectrum of choice when there is a focus on low frequency sounds, which is not particularly the case in an urban setting. Therefore, we decided to focus on white noise as it is the standard followed by many institutions.

The procedure for white noise calibration is similar to the procedure above, where instead white noise is generated at sound levels of 30–105 dB in 5 dB-steps. Fig. 2 shows sound levels as measured by the 11 Nokia 5230 handsets used in our mapping experiments (coloured dots). These points are obtained as averages over 1 minute intervals. The standard deviation over these intervals – which gives an indication of random errors, and thus precision – is low: an average of 0.15 dB(A) over the full range of sound levels, with maximal values (of up to 0.37 dB(A)) at low reference values. This low variability indicates

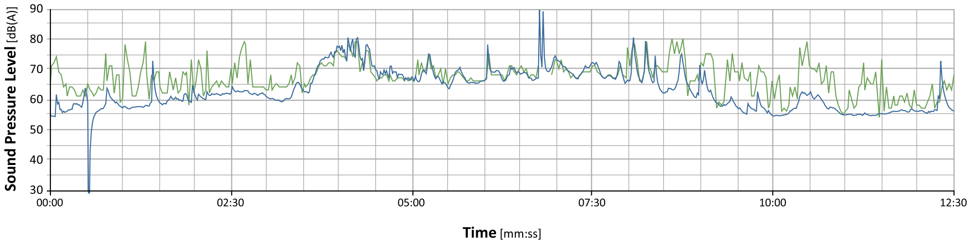

Fig. 3. Excerpt of comparison graph for Nokia 5230 #0 (green) vs. DT8852 (blue) sound level measurements during an outdoor walk. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

high precision of handsets when used as SLMs, at least in this controlled environment. Fig. 2 also shows a polynomial regression fit of the calibration points (bright red line), which represents the average behaviour of the 11 phones tested. Again we see that phones behave quite similarly—with the exception of phone #9 which is about 5 dB off for typical urban sound levels with respect to the other phones. Note that this is not an issue since we calibrate each phone independently. Concretely, we compensate for the average measurement offsets of each individual phone through a correction term which is found by linear interpolation between calibration points for that phone. This process occurs at the level of the NoiseTube software, which keeps track of calibration data and which ensures that the right correction term is added to measured values before storage and on-screen display. After thus calibrating all phones we checked the accuracy of phone #0 by exposing it to the same levels of white noise, the results of which can be seen as the green line in Fig. 2. This experiment shows that decent calibration methods can bring back systematic errors to 1.07 dB(A), decreasing to as little as 0.56 dB(A) for sound levels under 100 dB, at least in a controlled environment.[5] Frequency runs of the calibrated phone also showed vast improvements with respect to the situation before calibration.

Validation in the field. In a real-world measuring campaign phones do not find themselves in an anechoic chamber with pure white noise coming in from one particular direction. Rather, they are exposed to the elements and surrounded by sounds of arbitrary complexity. As a result one should expect a decrease in accuracy in such a setting. To get an idea of real-world performance of calibrated phones, we compared the sound levels as measured by phone #0 with those measured by a CEM-DT8852 Class 2 SLM [18] during an outdoor walk in the target area. This SLM advertises an accuracy of ± 1.4 dB, but we found that – at least for white noise – it performs even better, with an average absolute error of just 0.77 dB (see the dark red line in Fig. 2). Fig. 3 shows an excerpt of the sound levels measured during this walk. It is immediately apparent that accuracy of the phone (w.r.t. the SLM) varies substantially, wind being an important factor the effects of which can clearly be observed in the first few minutes. The effect of wind is not constant because, as the building density in the area and our walking direction changes along our itinerary, so does our exposure to the wind. On top of this, unlike the DT8852 the phone does not have a windscreen, so that fairly moderate wind speeds, such as the 15 km/h observed at the nearby Antwerp airport on that day, do have an effect. We also notice that the DT8852 evolves more smoothly that the Nokia measurements. This is due to the fact that conventional SLMs such as the one used here use time weighting rather than linear averaging (as applied by integration-averaging SLMs, as well as NoiseTube) for computing sound levels from instantaneously varying air pressure, which has the effect of smoothing out fluctuations. Because of this, and because measurements of the phone and the DT8852 have different timings, it is difficult to quantify accuracy levels at the level of individual measurements. Rather we should compare logarithmic averages for corresponding intervals. Over the entire itinerary, which lasted 81 min, the calibrated Nokia 5230 #0 measured an average sound level of 71.05 dB(A), while the DT8852 measured 70.90 dB(A). For the first 25 min the averages are respectively 70.01 and 69.82 dB(A). Over the interval between minute 4 and 8, in which the phone matches the SLM very well (at least visually), we find average sound levels of 70.57 and 72.94 dB(A) respectively, which is not as good as for the longer intervals but still very close. For the windy interval between minute 9 and 13 the phone overestimated sound levels substantially, measuring an average of 68.34 dB(A) compared to 58.90 dB(A) on the SLM. This indicates that the lack of wind protection can cause temporary systematic errors of up to 10 dB. Still, we notice that this effect all but disappears if we average the sound level over sufficiently long periods, as demonstrated by the negligible differences for the 25 and 81 min intervals. Considering that the microphone we deal with is not at all designed for this purpose, nor is the rest of the phone's hardware, this is a surprisingly good result.

The importance of large sample sets. The evidence presented supports the main punchline for this article, namely that the combination of calibration and statistical averaging over large datasets allows to counteract inaccuracies and imprecisions inherent to measurements taken by mobile phones. Indeed, the results above indicate that averaging is able to deal with random errors due to wind or user behaviour. *Heterogeneity* is the key concept here: the more measurements contained in a data set, the larger the probability that the errors vary, and thus that their average reduces this error. While calibration is a way to compensate for systematic errors (thereby increasing accuracy), averaging allows us to reduce random errors

[5] One should take into account that 3 dB is the smallest difference that can be detected by the human ear, so that these numbers indeed correspond to small errors.

(thereby increasing precision). While precision is certain to decrease in real-world settings, it is very difficult to obtain numerical values in such conditions, as measurements are not reproducible. However, as random errors are inherently unpredictable, one can only deal with them by working with large sample sets and thus ensuring statistical relevance. Moreover because the sound level in typical urban soundscapes varies constantly, averaging also allows us to get a representative value for different times or places. Finally, we note that heterogeneity is all the more easy to obtain in a setting where different users are involved, who carry different (calibrated) measurement devices, and by virtue of their mobility also introduce different trajectories, and thus orientations and incident wind speeds. Moreover, if we aim to compute average sound levels of particular times in the day, say a peak hour during a working week, we can consider measurements collected during different days of that week, which introduces still more contextual heterogeneity. As we shall see in Section 4, combining thorough calibration with the collection of massive amounts of data allows us to obtain noise maps which, while not city-wide, are of a quality comparable to those produced by EU member states today.

4. Participatory noise mapping

Having thoroughly improved the performance of individual phones as SLMs, the next step is to design a realistic participatory measurement campaign where we combine equipment with users, bring geographical factors into the picture and build up the lessons learned so as to make noise maps. We were fortunate in that we could collaborate with one of the best-known citizen-led action groups in Belgium, *Ademloos*. This non-profit association, initiated as a result of environmentally dubious government plans for reorganising traffic around Antwerp, has been labouring for the past ten years for a more sustainable solution to the city's congestion problems. Organisations such as Ademloos are the proverbial wet dream for a citizen science project such as this one, consisting of motivated, community-driven citizens who are concerned with their environment but who do not typically have any technical or scientific domain knowledge.[6]

We structured the actual campaign in a way that allowed us to maximise research outcomes. In particular as this was the first time NoiseTube was used in a coordinated measurement campaign, we decided to run the experiment in two phases. Phase 1 consisted of a smaller, more controlled experiment allowing us to test and fine-tune our procedure and guarantee that sample sets of a large enough size are built up. Phase 2 involved more volunteers and less constraints, thus corresponding more closely to participatory noise mapping where data accumulates as citizens use NoiseTube in the context of their daily lives. While the first phase ensures that complete maps produced for the target area in the second phase this is more difficult as there may be gaps in time as well as in space. However, we shall see below that the maps produced are still valuable, both at the level of noise pollution data as for learning how user behaviour changes in a setting with less constraints.

It should be noted that to represent the large amount of data gathered on a map we cannot rely on our standard NoiseTube web application, which produces one map per measurement track with coloured dots for each individual measurement. Instead we introduced a statistical component which enables us to produce one single noise map for a collection of measurement tracks. The basic procedure is this: divide the measured area into smaller areas, partition the set of measurements over those areas, make a statistical analysis per unit area, and finally, map the colour coded averages on each pertaining area. Clearly, this approach relies crucially on localisation to be able to classify the multitude of measurements that are gathered. For this reason it is also important to get an idea on the precision of GPS measurements. Since this is highly dependent on the urban layout (high buildings having an obvious effect) we achieved this by positioning ourselves at 7 locations in the target area, taking 100 measurements at each location, and then computing the spread on longitude and latitude values obtained. Note that this spread is computed with respect to the average longitudes and latitudes measured, rather than with respect to absolute geographical coordinates which are very hard to obtain (amongst others due to projection issues). We found the average spread on latitude to be 2.76 m, and on longitude to be 2.23 m. Maximal differences measured were 8.25 m for latitude and 7.94 m for longitude. These values are important when deciding upon the granularity of noise maps, as we shall see the below.

As per our suggestion, the secretary served as the main communication channel between us and Ademloos members, his first task being to line up a list of interested members located in that area. Meetings were set up at each stage of the experiment. First, we organised an introductory meeting with interested members, after which volunteers for the actual measurement campaign were recruited and their main constraints summarised. The action group had the main say in practical issues such as the focus area (Linkeroever, or "left bank") and measurement dates and times. Second, 1 h-training meetings were set up right before each measurement phase, to distribute phones and to specify how measurements where to be carried out, both in terms of using the software and good practices for improving data quality. These training meetings were crucial since about half of our volunteers had limited experience with mobile phones. For this reason we simplified our user interface to pure sound level measurements, excluding the tagging component and deactivating automatic communication with our website to upload measurement data. We also supplied a short document summarising the measurement protocol in simple terms, including lots of pictures on phone usage and troubleshooting. These meetings were also used to communicate preliminary outcomes and to allow opportunities for feedback. A final meeting in May 2011 presented the outcomes of both phases of the experiment and the resulting noise maps, including

6 In fact, some of our measurers had never owned a mobile phone.

a discussion round, ample time for questioning, and a survey polling volunteers on their experience with NoiseTube and participatory noise mapping. While the survey is too small to draw any kind of general conclusion, we mention some notable outcomes relevant to this particular experiment in Section 5.

Phase 1. In the first phase of our mapping experiment we wanted to control as many free parameters as possible in order to be able to rectify potential issues of our measurement architecture and ensure statistical relevance of the obtained data. At the same time we wanted the experiment to be defined in a way that corresponds closely to a realistic use case of NoiseTube: that of coordinated measurement campaigns. Indeed, one way that we envision participatory platforms such as NoiseTube to be useful is to tackle local issues that are typically hard to capture in simulated noise mapping, for example due to neighbourhood issues, roadworks and the like. These campaigns can be carried out from the bottom up, i.e. through grassroots actions carried out by citizen collectives such as is the case here, or top–down, i.e. by authorities to quickly map an area of concerns. The middle way between the above requirements led us to consider fixed trajectories. While fixed trajectories are only one way to coordinate measurement campaigns, they are representative of this use case while at the same time allowing us to gather large data sets. Concretely, four volunteers followed a pre-defined measurement track twice daily during a working week during a chosen off-peak hour (5–7 July 2010, 9:00–10:00 pm). The next week a peak hour was covered by four different volunteers (12–16 July 2010, 7:30–8:30 am). Phase 1 covered a total area of about 0.4 km $\times$ 0.4 km with a track of about 2 km. The track was chosen as a composite of typical urban soundscapes such as an important intersection, a park, and residential streets, all near the area where volunteers lived. At the same time its length was chosen so that four people measuring one hour a day would produce enough data so as to ensure statistical relevance. A quick theoretical calculation shows that in ideal conditions we would gather a total of about 36,000 measurements,[7] which gives an average of 180 measurements per 10 m covered—ample for statistical analysis.

Participatory noise maps for phase one of the Ademloos experiment are shown at the left and centre of Fig. 4; the chosen trajectory is clearly recognisable in each map. The map on the left shows average sound levels for the peak hour and is based on 30 977 measurements, while that in the centre shows average sound levels for the off-peak hour and is based on 36 394 measurements. For each week the number of measurements is very close to our theoretical calculation above. Both maps use the same colour scale as the official L_{den} map for road traffic in Antwerp, the corresponding detail of which is shown at the right of Fig. 4. Measurements were made during different days of the week, but are accumulated here to obtain averages for the chosen hours. We chose to organise the measured area into a grid with a unit size of 20 m $\times$ 20 m, excluding grid elements with less than 50 measurements for statistical significance. Note that there is always a trade-off between these values, as a smaller grid size requires more data to achieve statistically significant results. Next to this, grid size should never be smaller than localisation precision, for obvious reasons. Maximal GPS errors, mentioned above, suggest that one could decrease map granularity to cells of 10 m $\times$ 10 m, at least within the targeted area. However, in experimenting with values for grid and sample size we found the above combination to be best in terms of balancing clarity with statistical relevance. The peak-hour map contains 172 grid elements, has an average of 164 measurements per grid element, average sound level of 63.6 dB(A) over all grid elements, and an average standard error of 0.36 dB(A) over all grid elements.[8] The off-peak hour map contains 192 grid cells, has an average of 167 measurements per grid cell, average sound level of 60.8 dB(A) over all grid cells, and an average standard error of 0.34 dB(A) over all grid cells. For both maps the spread of average sound levels is between about 55 and 70 dB(A): we find spreads of 56.7 dB(A)–71.9 dB(A) and 55.3–69.7 dB(A) for peak and off-peak hours respectively. The difference of 2.8 dB in average sound level is reflected by a clear visual discrepancy between the maps for both time periods—e.g. there is much more yellow on the off-peak map. High noise levels along the main busy road, horizontally traversing the middle of the area, are clearly recognisable on each map. An interactive version of these maps, and others generated from data collected in phase 1 (e.g. with different grid cell sizes), can be consulted online.[9] On these maps users can click individual squares to obtain further statistics such as minimum, maximum and average sound levels, standard deviation, and size of the data sample.

Comparison with official noise maps is difficult because of the difference in approach and in quantities that are represented. Official maps for Antwerp represent either L_{den} or L_{night}, and are generated for road traffic, air traffic, train traffic and industry by simulation, as discussed in Section 2. Ideally we should compare our peak and off-peak hour maps with L_{day} and L_{evening} maps respectively, but these maps are not mandated by the END and are therefore not available. The most suitable map to compare with is then the L_{den} map for road traffic, shown at the right of Fig. 4. This is because first, the main noise source in this area is road traffic, there being no airport or railways in the neighbourhood and industry also being somewhat farther away. Second, because all our data was collected during the day and the evening comparison with the L_{night} map would be flawed. However, one should note that because L_{den} is a weighted average (see Eq. (2)) it is one step further away from the $L_{\text{A,eq}}$ values of our participatory maps. The simulated map shown dates from January 2010 but relies on traffic statistics going back to 2006. It represents the L_{den} noise level in an average 24 h day at a height of 4 m above the ground. Only traffic on the larger roads is considered, with propagation into neighbouring streets computed by the simulation. Our maps are based on measurements of all noise (or rather sound) sources in all streets, taken at a height of about 1–1.5 m.

7 $36,000 = 30$ measurements/min $\times$ 60 min $\times$ 5 days $\times$ 4 people.

8 We will come back to the notion of standard errors at the end of this section.

9 See http://www.brussense.be/experiments/linkeroever.

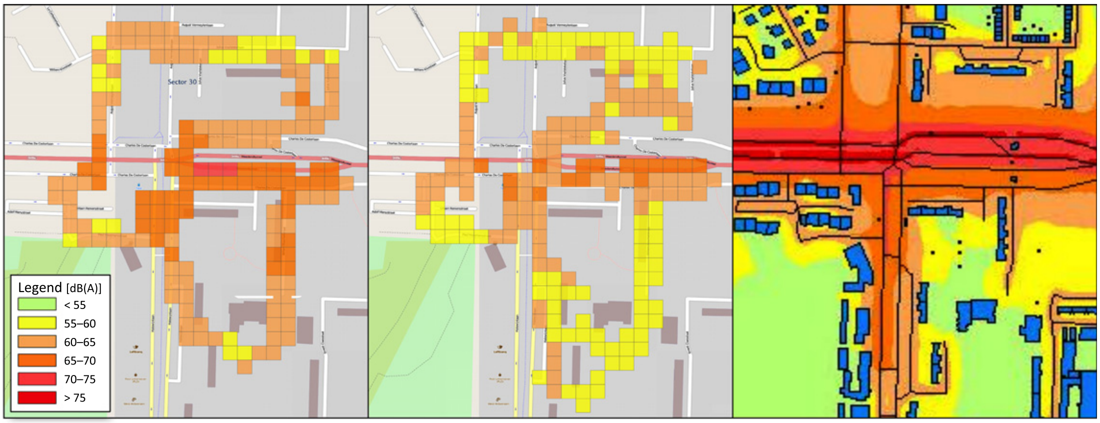

Fig. 4. Participatory noise maps: peak hour on the left, off-peak hour in the middle, simulated map for L_{den} on the right.

Despite the differences in approach and represented quantities there are a few noteworthy observations we can make by visual comparison. The first thing to notice is that our maps are rather similar to the official map in terms of overall sound level distribution. While this is not a proof of accuracy (see below), it is certainly reassuring, especially because the general patterns align well with expectations—indeed the main horizontal road and the intersection are the noisiest places. While there is little doubt that simulated maps can capture such general trends well, this has not been demonstrated before for participatory noise maps. Apart from the overall similarity there are also some marked differences between our participatory noise maps and the official one. One particularly interesting difference appears in the park area in the south-east quadrant, for which our maps indicate significantly higher sound levels than the official map. We argue that these higher values are partly due to an underestimation of simulated maps for this area, and an overestimation of our participatory maps. On the one hand, we note that official maps do not map this area directly, but rather estimate its sound levels in terms of propagation of traffic sounds from the nearby large roads. In this case the only road that is directly modelled is the main horizontal road, as is clearly witnessed by the decreasing sound levels as one moves away from it. In this model, sound has to travel from the main horizontal road, attenuating as it passes onto smaller roads and blocked by buildings along the way. While it may (or may not) be the case that the actual sound levels caused by the main road do not reach as far as the park, there is also considerable traffic on the main vertical road which could very well affect the park area directly. However, this behaviour is not included in the simulation model. On the other hand, participatory maps could overestimate sound levels due to the influence of the wind or occasional disturbances caused by the measurers themselves. Yet the influence of these factors is unlikely to have been constant during both weeks, and should thus be largely averaged out given the amount of data collected. Overestimation could also occur because of the lower measuring height, and because all sound sources are included. Likely sound sources which are not portrayed on the simulated map are other people in the park and measurers' footsteps. The sound level caused by the latter is a constant factor in all measurements, and may cause slight, yet systematic overestimations, in particular there where the ambient sound level is fairly quiet, as one could expect in this particular place. However, comparison of our two maps reveals that the average sound level measured in the park was significantly higher during the peak hour than in the off-peak hour. Because footsteps sound equally loud in the morning as they do in the evening there must be another factor. Given the time frames, this is in all likelihood related to a difference in traffic intensity on nearby roads. A final difference we would like to mention is that the average sound level our volunteers measured along the main busy road is actually somewhat lower that what is shown on the official map. This is however easily explained: while our participants walked on the sidewalk next to this particularly wide road, the simulated map also shows the sound level on the road itself. In fact if one looks more closely the values on the side of the road on the simulated maps are quite similar than the ones our participatory maps portray (in particular for the peak hour).

Phase 2. For the second phase of our experiment we imposed a less strict coordination: instead of specifying a fixed route and time frame we allowed participants to measure at will—albeit within a particular area and with a daily minimum, so as to ensure enough data was gathered. Loosening up space–time restrictions is reasonable because the results obtained during phase 1 indicate that the data gathered with NoiseTube Mobile is credible. This allows us to loosen up on space–time restrictions and focus instead on data collection patterns. The goal is to get an idea of how the freer movement of contributors and different size of the experiment affects the quality and completeness of noise maps.

Concretely, 10 volunteers were asked to measure for at least 1 (not necessarily continuous) hour a day during one working week in November 2010 in an area of about 1 km × 1 km, encompassing that of phase 1. This larger effort should in theory deliver at least 90 000 measurements,[10] which gives an average of just 36 measurements per grid cell over the whole week if we take a grid of 20 m × 20 m as above. Of course this number is just an indication as measurements are unlikely to

[10] =30 measurements/min × 60 min × 5 days × 10 people.

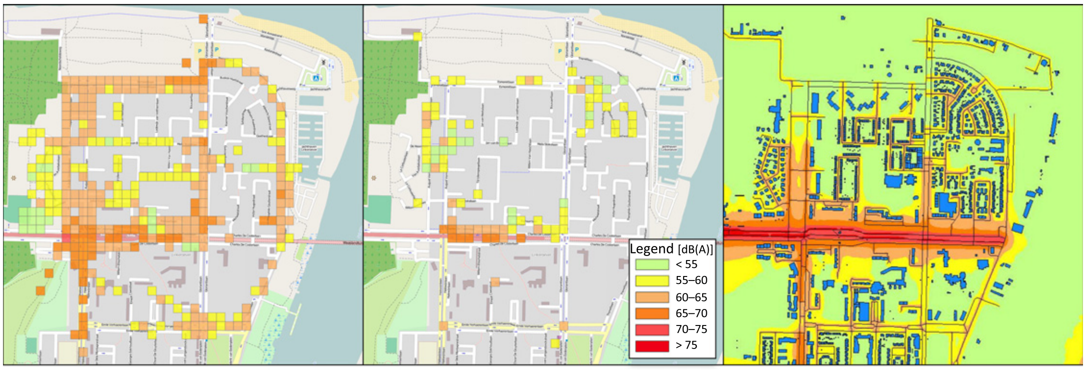

Fig. 5. Participatory noise maps: day period on the left, evening period in the middle, simulated map for L_{den} on the right.

be uniformly distributed over the area; still, it seems likely we may have to increase cell size to compensate for sample size. When the experiment was over a total of 84 309 measurements had been gathered, which is fairly close to what we predicted. A temporal analysis shows that almost all the data was collected during the day and the evening, with peaks at 11 am, 5 pm and 10 pm, and dips around 1 pm (lunchtime) and 7–8 pm (dinnertime). Unsurprisingly, very little measurements were made between midnight and 8 am.

As the total data set is too small to study shorter intervals while still maintaining significance, we focus on the day and evening periods, where the day period runs from 7 am–7 pm and the evening from 7–11 pm. Concretely, 62 853 measurements were gathered during the day, and 17 513 during the evening. With these numbers and the larger area to cover we found the best map representation to be one with grid cells of 40 m $\times$ 40 m and a minimum of 50 measurements per cell. Even so, the resulting map for the evening period, shown at the centre of Fig. 5, is very sparse. While many gaps remain, the map for the day period, shown at the left of Fig. 5, covers a much greater portion of the focus area. The day map contains of 329 grid cells, has an average of 77 measurements per grid cell, average sound level of 61.1 dB and an average standard error of 0.63 dB(A) over all grid cells. Sound levels vary between 44.0 dB(A) to 70.5 dB(A). The evening map contains 90 grid cells, has an average of 76 measurements per grid cell, and an average sound level of 58.0 dB(A) and standard error of 0.76 dB(A) over all cells. Here sound levels vary between 38.2 dB(A) and 68.8 dB(A), indicating, together with the difference in average sound level, that overall sound levels for the evening period are about 2 dB(A) lower—but more data is required to back up this statement properly. An interactive version of these maps may be found online at the same address as above, complemented by a number of alternative maps focusing on different users and time intervals. The liberty of choosing the latter is one big advantage of participatory maps with respect to simulated ones. However, a larger data set is required for a temporal analysis in terms of particular hours or days to become fruitful.

For completeness we once again juxtapose our participatory maps with the simulated map of L_{den} for road traffic of the target area, which is shown at the right of Fig. 5. The choice remains unchanged because almost no data was collected at night and – also in this larger area – road traffic is the main source of noise. For the same reasons as discussed above comparison of our maps with official ones is difficult. However it was further complicated due the fact that the data for phase 2 is much sparser, especially in the evening. Therefore we only attempt a visual comparison of our day period map. Again we see areas for which our map shows markedly higher noise levels than the official map does. The clearest example is the northernmost west-east road on the map. According to the official map road traffic noise along this road stays below 55 dB(A), but the data collected by our measurers strongly refutes that. The same holds, to a lesser extent, for the areas in the north-east and south-east corners of the map. By virtue of the outcomes of phase 1 of the experiment, which shows that we find the same results for large roads and hence that our approach is credible, we have confidence in the results of phase 2 pertaining to areas that are not adequately modelled in simulated maps. Indeed such areas are away from the busiest roads (which the simulations capture well) and unlikely to be reached by propagation of sound from the roads modelled. It is precisely in these areas that participatory mapping is particularly able to make a difference.

Another noteworthy aspect is the zone towards the west side of the map and just north of the main west–east road, for which our day period map shows sound levels that are suspiciously (given the road proximity) lower than on the official map. Inspection of the raw data revealed that the measurements in question were almost certainly taken at the home of one of our measurers. This case is a helpful reminder of the fact that unintentional user behaviour is a potential source of inaccuracies, or at least incompatibility with the rest of the data presented on the map. Again, heterogeneity is key: data coming from one user only is less credible than data coming from multiple users. A technological solution could be to assign automatic tags and/or a quality factor in terms of the number of users contributing and GPS behaviour.

Finally we should note that the gaps in the day map, and also in that for the evening, would likely have become much rarer if we had increased the number of participants and/or the duration of the experiment. However, without very strict coordination or truly massive numbers of participants, it may be difficult to achieve full coverage of the area. One way to fill gaps without measuring may be to apply spatial interpolation techniques. However, as noted in [4, p. 45] and confirmed by our own unsatisfying attempts, standard interpolation algorithms are not suitable for modelling sound propagation. A more suitable approach could be to fill gaps in participatory maps with data from simulated maps—which once again indicates that both noise mapping approaches can be complementary. For obvious reasons it may also be infeasible to create participatory maps for night noise, especially over large areas. Hence also for that purpose the simulation approach remains useful and necessary.

Notes on accuracy and precision. Through visual comparison of our maps for phase 1 and 2 with the official L_{den} map, we have found strong indications that support the validity (e.g. the capturing of expected trends) and value (e.g. the detection of traffic noise that is apparently underestimated by the official map) of our participatory approach. However, visual observations alone are not enough to draw firm conclusions about the quality of our maps—nor of official ones. A first issue is accuracy, which indicates these maps' correspondence to true sound level values. Actually accuracy is very hard to determine globally for both types of maps, as the only way in which to know what the true values are is by carrying out extensive field measurements with professional SLMs, which is exactly what the simulation and participatory noise mapping approaches aim to avoid. Instead accuracy can be estimated by making comparisons with actual sound level measurements at discrete points, e.g. where nodes of a professional sensor network are situated. In Brussels there are 19 such sensor nodes. With respect to these reference values official noise maps are accurate up to $\pm$ 2 dB [19, p. 11], but of course the accuracy may very well be different at other locations. For our participatory maps the only point of reference we have is the accuracy of $<$1 dB we found over the duration of our validation experiment discussed in Fig. 3. If we assume that the Antwerp maps (for which no accuracy information has been released) have similar error margins than the Brussels ones we at least operate in the same order of magnitude, though an exact comparison remains difficult as we have glossed over several factors such as the aggregation of data over different users and time spans. Obviously we should conduct more side-by-side experiments, with portable SLMs or existing official sensor nodes, to be able to make better claims about accuracy. We stress once more that using official maps as a reference is not an immediate option, as these maps present different quantities and have clear shortcomings of their own.

Another way to assess accuracy of our participatory maps is to apply descriptive statistics techniques, such as by computing confidence intervals over sample sets. The sample errors we have mentioned with the statistics of each of the maps above is a first step towards computing confidence intervals. Standard errors estimate how close a sample average is likely to be to the true population mean, whereas standard deviation is the degree to which individuals within the sample differ from the sample mean. Because sound level is a highly variable physical parameter, standard deviations are typically high and in any case give no indication on the quality of the random sampling process. On the other hand, standard errors are typically used only for repeatable experiments, which is not the case here as we are also taking averages over time. The next step would be to compute confidence intervals from these standard errors. However, this issue requires further research, as our averaging over several dimensions together with the fact that variables averaged over are not necessarily independent nor normally distributed complicate matters substantially.

Besides the differences in represented quantities, another complicating factor is that, due to the colour bands of 5 dB, a lot of spatial variability (which may or may not reflect reality) remains hidden. Based on the data collected in phase 1 and 2 we could easily generate maps with narrower bands, or compare the aggregated data numerically instead. Yet for the official map that is not possible because the responsible authorities do not provide access to the data behind the map,[11] hence we are stuck with the end product.

5. Citizens as data gatherers

We consider it important also to evaluate NoiseTube, as a system for participative grassroots noise mapping campaigns, from the user perspective. Therefore we need an evaluation tool. The feedback gathered with this tool can be used to guide further development of NoiseTube and similar participatory sensing applications.

Considering the limited number of people who took part in our Antwerp experiments we could have taken an interpretative qualitative approach to evaluate the experience during the different phases. But since we plan to reuse the evaluation tool with a larger audience (e.g. students who tested NoiseTube or registered users of NoiseTube.net) we have chosen a methodology that can be scaled up more easily, namely that of a standardised questionnaire. Hence, the Ademloos members are not only a pilot group for testing the NoiseTube system but also for the evaluation tool. To design this questionnaire we built on dimensions found in literature on participative/mobile sensing, in combination with open issues regarding data representation [20,21]. We kept some open questions since we are still exploring the dimensions of the experience of sensing. Building on the results with the Ademloos group, and possibly other pilot groups, we should be able to revise the questionnaire to have closed categories for these dimensions as well. The questionnaire (which is in Dutch) is included in [12].

The questionnaire was filled out by the group at the start of the final feedback session in May 2011, during which we presented the results of the measuring campaigns. All 13 participants of the experiment completed the questionnaire—those who could not attend the meeting filled it out at home. These are 7 men and 6 women with an average age of 62 (standard deviation of 13 years). From a methodological point of view we learned that some questions should be reconsidered because

[11] In fact our contacts in the Antwerp city administration tells us that even they cannot easily access this data because the actual noise maps are produced by a subcontractor.

the answers do not show a lot of variation. For example a ranking question, instead of a Likert item scale, would have given more insight in prioritisation of the kind of information to be shown on noise maps, as well as on the motivation to take part in a mapping campaign. Although this was a first test of the questionnaire as an evaluation tool, besides methodological insights it has also provided valuable feedback concerning the NoiseTube system as it was tested. Below we summarise the most important indications. We must however stress that statistical significance is not possible with this limited and opportunistic sample of testers. Hence, while the feedback provided by this particular group is certainly helpful, it cannot be used to draw general conclusions.

Looking at the motivational factors for taking part in the campaign, most agreement was on a general concern about noise pollution (11 out of 13 considered it "very important"), followed by personal experience of noise pollution ($\frac{8}{13}$ indicated this was "very important"), and supporting the Ademloos activism group (also $\frac{8}{13}$ "very important"). Supporting scientific research was also seen as an important ($\frac{5}{13}$) or very important (also $\frac{5}{13}$) motivation. Factors which were less important drivers for this particular group were an interest in technology, or that the campaign was a fun or a useful pastime. When asked about possible concerns regarding privacy, only one participant said he was worried because "*there is a log of where I was at which moment*". A lack of general concern about privacy can be hypothetically attributed to the fact that participants were part of an existing group that know and trust each other, who consciously took part in a scientific experiment and who got time to get acquainted with the researchers is person.

With open questions we asked for the 3 most pleasant and 3 most annoying things they experienced taking part in the campaign. Some unexpected dimensions appeared. The immediate feedback provided by the app was seen as pleasant because it gave insight in sound (level) and in the problem of noise (relativity, locality). Several people mentioned that they liked the team spirit the campaign created in their group. Multiple participants enjoyed the fact that walking was an integral part of the sensing activity, which was seen as beneficial for physical fitness and general health. However one participant found it annoying that the campaign obliged him to walk through unhealthy areas polluted by heavy traffic. Less annoying experiences were mentioned than pleasant ones. Other annoying experiences were: stability of the app,[12] the need to constantly hold the phone, the dullness of fixed routes (in phase 1), and not being able to have conversations while measuring (one person even noted that she regretted having to pass by acquaintances, which indicates how committed she was to the experiment). Regarding the fixed hours in phase 1 and the freely chosen times in phase 2, opinions seemed to be mixed. For instance, one user who took part in both phases mentioned that the fixed time frame was easier to keep up (comparing phase 2–1), while another who also took part in both phases mentioned the fixed time frame as an annoying aspect (of phase 1).

We also asked for feedback on features of the NoiseTube system which these users had not been given access to. For instance, we left the social tagging feature out of the app used by the participants. In the questionnaire we asked whether they would have liked to be able to indicate sources of noise while measuring. The group was very positive about this ($\frac{10}{13}$ answered "Yes"). Also popular was the idea of being able to comment on the measurements made by their group through a website ($\frac{6}{13}$ "Yes", $\frac{3}{13}$ "Maybe"). The proposed possibility of commenting on measurements made by other NoiseTube users (outside of their group), was seen as much less interesting ($\frac{8}{13}$ "No", $\frac{1}{13}$ "Maybe").

Finally we asked questions about the information that is, or could be, presented on noise maps. We found two interesting things. On the one hand, sound level was preferred ($\frac{9}{13}$) to be displayed in categories (as on conventional noise maps) rather than as exact values. On the other, a slight majority ($\frac{7}{13}$) preferred maps to show peak values rather than averages, which may indicate a flawed understanding of the dynamics of sound perception.

6. Towards large-scale adoption

The main contribution of this article is that it shows that, when taken seriously (i.e. using calibrated phones, ensuring spatio-temporal density, etc.), participatory sensing can serve as an alternative or complementary approach to official END-mandated strategic noise mapping efforts—especially when the focus is on specific local issues occurring in reasonably-sized areas or time frames. Additional analysis and experiments are required before we can make more precise statements about the accuracy of these maps—a situation which is similar to that of official maps. Nevertheless, qualitative comparison with the official road traffic noise map of the area already provides strong indications that support the validity (e.g. the capturing of expected trends) and value (e.g. the detection of traffic noise that is underestimated by the official map) of the participatory approach, as well as its complementarity with the conventional simulation-based method. There are certainly several other projects which, like NoiseTube, focus on applying participatory sensing to the assessment of environmental noise, with a current acceleration of activity in that area that goes hand in hand with the growing popularity of citizen science and open data initiatives [10,11,22–27]. However, to the best of our knowledge none of our peers have conducted experiments to validate participatory noise mapping at this level of scale and realism as we did.

The goal of this research was to first and foremost tackle the prevalent scepticism with respect to the quality of gathered data. We have shown that this scepticism is not justified, at least in the particular setting of our experiments. The question

12 Related, for the most part, to the automatic restarts necessitated by a memory leak in the JavaME OS, which were confusing and sometimes failed. This is no longer an issue in the Android version of NoiseTube.

remains then how we can extrapolate our efforts towards large-scale adoption in a realistic way. Immediate issues that come to mind are calibration, rules for defining measurement protocols such that certain quality standards are achieved, increasing the scale of experiments so as to move on to operational (as in, non-experimental) deployments.

The calibration efforts carried out for this particular noise mapping experiment are extensive and of course one cannot expect all participatory mapping actions to follow the same procedure in general. However, most of our users never have to deal with calibration experiments, as NoiseTube Mobile comes with built-in settings for a number of models and automatically download additional ones as they accumulate online. All data that is gathered is labelled with the calibration method used, so that we can always derive the accuracy of contributed data after the facts. We should note that calibration settings for phone models are not only added by ourselves, but also at times by some of our more experienced users. On top of this, when we are contacted to guide a noise mapping campaign, it is often the case that we lend our equipment to the party concerned. In our experience there is high potential for such a "floating" sensing system/organisation, tailored to small-size campaigns and moving around a country or even a European-sized area, to boost the adoption of participatory techniques, especially when moving towards other parameters for which more specialised equipment is required (e.g. air pollution).

The next issue deals with defining proper measurement protocols. In this, one should of course always first determine what the requirements of the community in question are. Once this is clarified one can rely on some rules of thumb, though it must be said that each campaign is different and we are glossing over a number of aspects here. Depending on a number of parameters such as the size of the area, the time interval (or intervals) of interest and the number of participants, one can make estimates of the effort required to achieve significant sample sets, as we have done in the above. Given an area of width W and height H (both in metres), mapped for a time span T (in minutes) and D days, at a rate of R measurements per minute[13] and with grid cells of size $r \times r$ (where r stands for resolution, also in metres), n volunteers would deliver a number of measurements M per grid cell given by

$$M = \frac{R \times n \times T \times D \times r^2}{W \times H}. \tag{3}$$

Using this formula, one can determine one parameter, such as the number of volunteers or days required, in function of the others, keeping in mind that M should be large enough to ensure statistical significance. A more precise number can be obtained in terms the length of all walkable streets in the area, and of course the actual conditions of a campaign should always be checked once it is finished. Note that we used this formula to predict that phase 2 of our experiment would deliver relatively small sample sets, effectively inhibiting temporal analysis (as this would divide these sample sets in even smaller constituent parts). However, the formula also indicates that completing phase 2 maps would be just a matter of increasing effort. Of course campaigns may be defined in much different terms than the ones we have reported on here, for example on the basis of type of location rather than area, e.g. someone may be interested to compare airport areas all over the world. A formula such as the one above would need to be adapted in such a setting.

This brings us to the final issue, which deals with scaling up the approach presented so as to move towards operational deployments. First and foremost we note that scaling up our approach can occur by increasing the size of campaigns, but certainly also by increasing their frequency. Indeed, the experiment reported here already corresponds to a real-world use case, as it was our volunteers' explicit request to assess the noise present in the area of the given size. Since this experiment we have repeated analogous campaigns in two places, with two more planned in the near future. Each time the goal is to assess local noise variations in an area of relatively small size through a grassroots campaign, and each time we were able to apply a similar approach with similar results. Motivating users is not an issue in this context, as citizens involved are already concerned and willing to take up action. Actually this should come as no surprise, since this is precisely an area in which participatory mapping nicely complements official maps, which are inherently incapable of capturing fine-grained spatio-temporal variations. Nevertheless one can and should also consider increasing not only the frequency of grassroots campaigns but also their size, e.g. to map larger areas or focus on larger time spans. While it is certainly the case that district- or even city-wide campaigns are more difficult to achieve, we have several indications that our architecture and mapping algorithm scales up to these grander settings. Indeed the NoiseTube website has been running several years now, so that for some cities significant amounts of data has been gathered. For example for the Brussels Region, which is roughly a region of 4 km × 4 km, 93 users have contributed a total of 603 847 measurements. In the context of recent work on a privacy-preserving extension of our map-making algorithm [28], we have created noise maps for this much larger dataset relying on the same grid-based approach as the ones presented in this article. Of course the quality of this map is a whole different matter, since it is a composite of data contributed over several years and by a multitude of users and phones, so that it contains gaps and the accuracy of the measurements contributed varies widely. Another challenge is motivating enough participants to contribute data which is focused and accurate enough so that relevant maps can be produced. City administrations are definitely beginning to take an interest in this novel approach for noise mapping, as witnessed by our experience in Belgium but also in the context of several EU-funded research projects (e.g. i-Scope). We are convinced that through city-regulated incentive schemes – which carefully balance altruism with a form of direct or indirect remuneration

[13] The rate of the NoiseTube version for Android is 60 measurements per minute, while that of the (older) Java ME version used here is 30 measurements per minute.

(possibly but not necessarily of a monetary nature) – it is possible to mobilise citizens at such a scale, which in turn raises data accuracy issues. On top of this incentives need to take the issue of dark spots into account, i.e. they have to be organised such that citizens are also motivated to map areas which correspond to gaps in the collective participatory maps. Further research is required before these matters may be clarified; in the meantime, small-scale campaigns are definitely gaining ground.

Acknowledgments

Our special thanks goes out to Prof. Patrick Guillaume, without whose guidance, expertise and equipment we would never have been able to carry out the calibration experiments reported on in this paper with such professionalism and thoroughness. And of course to all volunteers from Ademloos, without whom this research would not only not exist, but also would not make any sense. Thank you Guido Verbeke, Ivo Delrue, Els Meelker, Staf Waelbers, Rudy Crombé, Raymond Leunis, Jos Aerts, Luc De Belie, Chantal Flament, Liliane Van Handenhoven, Marjo Leenknecht, Magdalena Maes and Peter Nijssen!

This research was funded by the Flemish Fund for Scientific Research (FWO), the Brussels Institute for Research and Innovation (Innoviris), and the EU CIP/ICT PSP Project i-Scope (interoperable Smart City services through an Open Platform for urban Ecosystems).

References

[1] Allison Woodruff, Jennifer Mankoff, Environmental sustainability, IEEE Pervasive Computing 8 (2009) 18–21.
[2] Jeffrey A. Brke, Deborah Estrin, Mark Hansen, Andrew Parker, Nithya Ramanathan, Sasank Reddy, Mani B. Srivastava, Participatory sensing, in: WSW'06: Workshop on World-Sensor-Web, October 31, 2006, Held at ACM SenSys'06, Boulder, CO, USA, October 31–November 3, 2006.
[3] Eric Paulos, Citizen science: enabling participatory urbanism, in: M. Foth (Ed.), Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City, Information Science Reference, IGI Global, 2009, pp. 414–436.
[4] Working Group—Assessment of Exposure to Noise (WG-AEN), Good practice guide for strategic noise mapping and the production of associated data on noise exposure, version 2, Position Paper, European Commission, 2007.
[5] United Nations Environment Programme (UNEP), Rio declaration on environment and development, 1992, in: Proclaimed at the United Nations Conference on Environment and Development, Rio de Janeiro, Brazil, 3–14 June 1992.
[6] Eiman Kanjo, Jean Bacon, David Roberts, Peter Landshoff, MobSens: making smart phones smarter, IEEE Pervasive Computing 8 (2009) 50–57.
[7] Nicolas Maisonneuve, Matthias Stevens, Bartek Ochab, Participatory noise pollution monitoring using mobile phones, Information Polity 15 (2010) 51–71.
[8] European parliament and council, directive 2002/49/EC of 25 June 2002 relating to the assessment and management of environmental noise, Official Journal of the European Communities L189 (2002) 12–26.
[9] WHO Regional Office for Europe/European Commission Joint Research Centre, Burden of disease from environmental noise, quantification of healthy life years lost in Europe, Technical Report, 2011.
[10] Rajib Kumar Rana, Chun Tung Chou, Salil S Kanhere, Nirupama Bulusu, Wen Hu, Ear-phone: an end-to-end participatory urban noise mapping system, in: IPSN 2010: Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks (April 12–16, 2010, Stockholm, Sweden), ACM, New York, NY, USA, 2010, pp. 105–116.
[11] Eiman Kanjo, NoiseSPY: a real-time mobile phone platform for urban noise monitoring and mapping, Mobile Networks and Applications 15 (2010) 562–574.
[12] Matthias Stevens, Community memories for sustainable societies: the case of environmental noise, Ph.D. Thesis, Vrije Universiteit Brussel, Brussels, Belgium, 2012.
[13] Kenneth Kaliski, Eddie Duncan, James Cowan, Community and regional noise mapping in the united states, Sound and Vibration (2007) 14–17.
[14] International Organization for Standardization, Acoustics—Description, Measurement and assessment of environmental noise—part 2: determination of environmental noise levels, ISO 1996-2:2007, 2007.
[15] European commission, commission recommendation of 6 August 2003 concerning the guidelines on the revised interim computation methods for industrial noise, aircraft noise, road traffic noise and railway noise, and related emission data, Official Journal of the European Union L212 (2003) 49–64. Annex II of Directive 2002/49/EC.
[16] International Electrotechnical Commission, Sound level meters, IEC 60651:2001, 1979–2001. Withdrawn, replaced by IEC 61672-1:2002 and IEC 61672-2:2003.
[17] Silvia Santini, Benedikt Ostermaier, Robert Adelmann, On the use of sensor nodes and mobile phones for the assessment of noise pollution levels in urban environments, in: INSS2009: Sixth International Conference on Networked Sensing Systems (17–19 June, 2009, Pittsburgh, PA, USA), IEEE, 2009, pp. 1–8.
[18] International Electrotechnical Commission, Electroacoustics—sound level meters—part 2: pattern evaluation tests, IEC 61672-2:2003, 2003.
[19] Brussels Instituut voor Milieubeheer, Geluidshinder door het verkeer—Strategische kaart voor het Brussels Hoofdstedelijk Gewest/Bruit des transports—Cartographie stratégique en Région de Bruxelles-Capitale, Technical Report, 2010 (in Dutch and French).
[20] Paul M Aoki, R.J. Honicky, Alan Mainwaring, Chris Myers, Eric Paulos, Sushmita Subramanian, Allison Woodruff, A vehicle for research: using street sweepers to explore the landscape of environmental community action, in: CHI'09: Proceedings of the 27th International Conference on Human Factors in Computing Systems (April 4–9, 2009, Boston, MA, USA), ACM, New York, NY, USA, 2009, pp. 375–384.
[21] S. Consolvo, D.W. McDonald, T. Toscos, M.Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, I. Smith, J.A. Landay, Activity sensing in the wild: a field trial of UbiFit garden, in: CHI'08: Proceeding of the Twenty-Sixth annual SIGCHI Conference on Human Factors in Computing Systems (Florence, Italy, April 5–10, 2008), ACM, New York, NY, USA, 2008, pp. 1797–1806.
[22] S.B. Eisenman, E. Miluzzo, N.D. Lane, R.A. Peterson, G.-S. Ahn, A.T. Campbell, BikeNet: a mobile sensing system for cyclist experience mapping, ACM Transactions on Sensor Networks (TOSN) 6 (2010) 6:1–6:39.
[23] Theodor Foerster, Simon Jirka, Christoph Stasch, Benjamin Pross, Thomas Everding, Arne Bröring, Eike Juerrens, Integrating human observations and sensor observations—the example of a noise mapping community, in: DE-2010—Towards Digital Earth: Search, Discover and Share Geospatial Data 2010. Workshop Held at Future Internet Symposium, Volume 640 of CEUR Workshop Proceedings, Berlin, Germany, September 20, 2010.
[24] Immanuel Schweizer, Roman Bärtl, Axel Schulz, Forian Probst, Max Mühläuser, Nois map-real-time participatory noise maps, in: PhoneSense 2011—Second International Workshop on Sensing Applications on Mobile Phones, November 1, 2011, Held at ACM SenSys 2011, Seattle, WA, USA, November 1–4, 2011.
[25] HabitatMap, Lunar logic polska, Aircasting, 2011.
[26] European Environment Agency (EEA), Eye on earth, 2008–2012.
[27] WideTag, Inc., WideNoise, 2009.
[28] George Drosatos, Pavlos S Efraimidis, Ioannis N Athanasiadis, Ellie D'Hondt, Matthias Stevens, A privacy-preserving cloud computing system for creating participatory noise maps, in: COMPSAC 2012.