An approach for the selection of optimal new AQ measurement locations in urban areas

HAQT: WP1

Lasse Johansson

Summary

In the Helsinki Air Quality Testbed project (HAQT) the existing HSY air quality monitoring network was augmented by adding new commercial AQ measurement instruments as additional AQ measurement data sources within the metropolitan area. The new instruments were tested in different locations and under different conditions to find optimal locations for them. The new AQ equipment used in HAQT includes, e.g., Vaisala Air Quality Transmitters AQT420 and AQT410, and Pegasor AQ Urban. Vaisala AQT420 is a compact, cost-effective solution to monitoring air quality and meteorological data. AQT420 is an air quality transmitter that measures up to four most common gaseous pollutants such as nitrogen dioxide (NO2), nitrogen monoxide (NO), sulphur dioxide (SO2), carbon monoxide (CO), and ozone (O3) and weather data, such as humidity, air pressure and temperature, plus particle mass concentration (PM2.5 and PM10) in the ambient air.

The air quality testbed area already contains a strong network of high quality measurement stations and therefore, to obtain maximum benefit from the additional AQ sensors the new measurement locations need to be selected carefully. For the selection of such optimal new measurement locations, an algorithm for the FMI-ENFUSER was developed. The optimal locations of new AQ instruments were assessed by defining a set of criterions that define the optimality mathematically. These criteria included a) the spatial coverage of the complete network taking into account individual measurement site quality and b) the heterogeneity of emission source signals with the measurement network. By using the optimality formulation it was possible to quantify the improvements that would occur in case a sensor was to be added in any location in the testbed area. In the produced algorithm tool a large number of candidate locations was assessed and the candidate location that showed the largest improvement was selected as a new measurement location; this iterative process was repeated until a measurement location was assessed for each AQ-sensor that was to be installed. The assessed final network of optimal sensor locations is shown in Fig 1. This result contains locations for 15 AQT sensors, 5 Pegasor AQ's and 5 additional backup locations (21.-25).

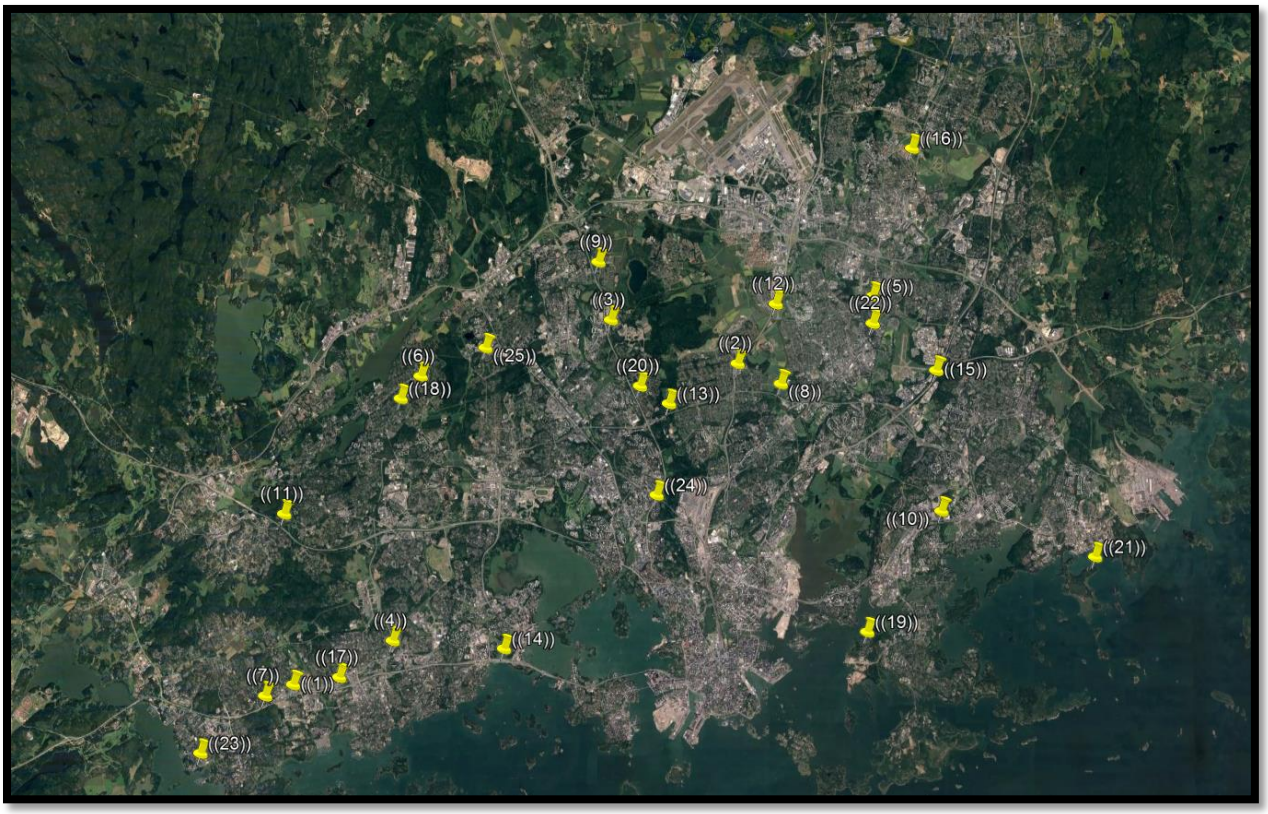

Figure 1: The sensor locations (up to 25) assessed with the optimal sensor location assessment tool in Helsinki region. The numbers shown in the map pins describe the installation order (priority). Satellite image privded by Google Earth.

It should be noted that the existing measurement network was different for each pollutant species and some of the installed measurement devices only measured PM2.5. Due to this, the developed tool was designed optimize the locations separately for each pollutant species and the user of the tool can assign preferences (weighting factors, relative importance) for these pollutant species. Further, since the location selection tool determines emission source signals (affected by time and meteorological factors) for measurement locations, it is also possible to define similar preferences in terms of emission categories. For example, by emphasizing PM2.5 and domestic household contribution for PM2.5 in the tool the assessed network of sensors emphasizes areas that are nearby small household residential areas.

Unfortunately, the development tool required significant amount of resources and by the time the tool was ready most of the AQT's were already installed. The developed tool can be used for any other region. The tool has been used for Nanjing Air Quality Testbed (NAQT) in a similar manner than it was used for HAQT. The tool can also be used in a reversed mode, so that it estimates which of the existing measurement sites and sensors add the least benefit for the measurement network and can be removed or relocated with least amount of information loss.

In order to use the tool FMI-ENFUSER modelling data for the area must exist for at least several weeks' time period. The limitation of the developed tool is that it does not yet consider legal issues,

permits and the availability of power, which impact significantly the final locations of sensor installations.

1. Methodology

In an attempt to find the optimal location for a new sensor the problem must be defined on mathematical terms. Here we define optimality as a function of two criterions: C1, C2 and C3 all of which can be computed using the FMI-ENFUSER model. The approach is to find an optimal location for each of these criteria separately, or more precisely, provide an improvement (optimality) field for the area using these criteria. An improvement value in this field indicates how much the criteria is improved when an additional sensor is installed in the location. The final selection can then be made by forming a linear combination from the multiple optimality fields.

It should be noted that while the optimal location with respect to a well-defined criteria can be computed, the final suggested sensor location is ultimately a compromise that is subject to user preferences. One task for the user is to define how much emphasis should be assigned to the individual criteria. In general, the criteria weightings can be different depending on the utilization of the measurement network. As an example, for modelling performance C3 is significantly more relevant than C2, but their relative importance can be reversed on local authorities' perspective. For similar reasons the user must also assign importance weights for a) different pollutant species and b) emission source categories (such as traffic, shipping, power plants).

In recent literature the topic of optimal measurement network design has been discussed in (Mofarrah et al 2010), (Hsieh et al 2015) and in (Yang et al 2013).

1.1 Area coverage with and without population density factoring (C1, C2)

In Johansson et al 2015 a data fusion algorithm has been presented. Given an existing measurement network of N measurement sites, for any selected location in the area one can use the presented algorithm to compute the expected standard deviation of the "fused" pollutant concentration when the measurements of the N stations are used as input. By performing this computation over the whole testbed area, one can produce a standard deviation (SD) field, such as has been presented in Fig 2a below. In general, the locations near existing measurement locations have lower SD and the areas that are farther away from measurement sites have larger SD. However, the individual measurement device quality is also taken into account in this assessment.

For FMI-ENFUSER it is beneficial to increase the coverage of measurement network (and therefore reduce SD in the area) for a number of reasons. The model uses measurement data to explain observed pollutant concentrations based on their origin (i.e., by emission categories) and one of these components is the background concentration and long-range transportation of pollutants. This task involves a combination of dispersion modelling and using a data fusion algorithm for modelled regional scale AQ data as well as measurement data. This process is not presented in this paper, but it can be said that that the quality of the process (especially for the background component) is improved when a geographically balanced measurement network covers the whole modelling area. We use this metric as a basis for criterion C1.

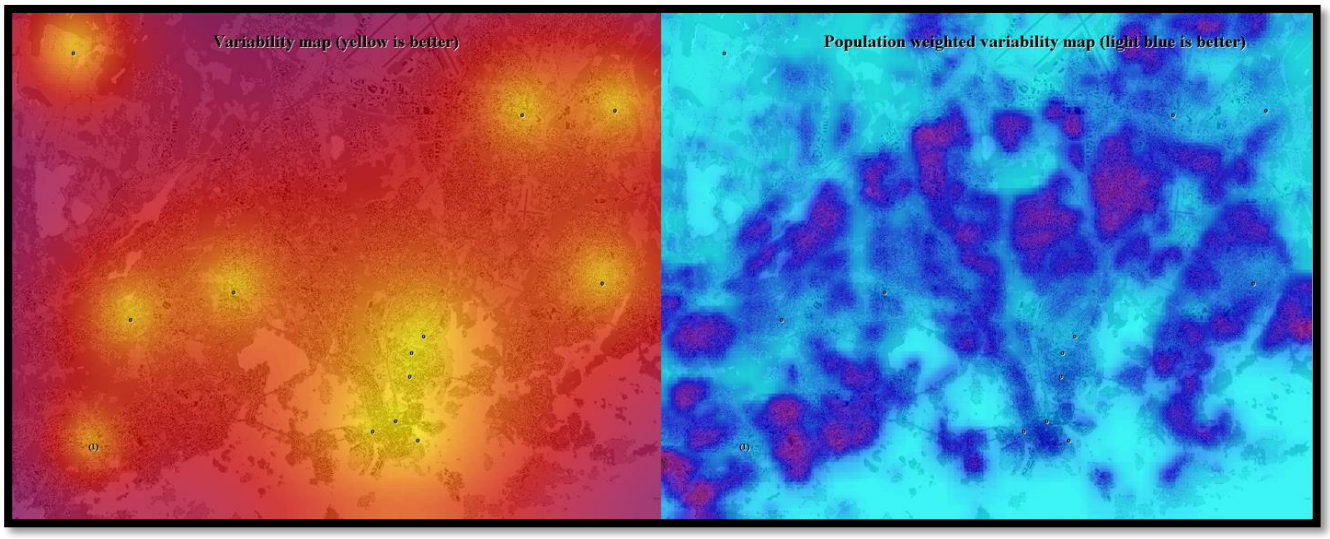

Figure 2a-b: Example of fused standard deviation heatmap (a) based on geographic distribution of existing measurement network with heterogeneous quality (lower SD is better, which is shown in yellow). In b) an example of population factored SD heatmap is shown (lower is better, which is shown in light blue).

For local authorities the geographical coverage of the network is a secondary priority, however. Based on expert opinions an improved coverage near populated areas is more favorable. Therefore, we combine the SD field with local population density by simply multiplying the SD-values with the local population density (see Fig 2b). In this combined product that is used to define C2, less is still better but the areas with little or no population become less relevant for improvements. **The optimization task for C2 is therefore, to find a location that causes the largest total reduction for population factored SD over the whole area of interest.**

The optimality of a location in terms of C2 is computed with respect to the current state when there are initially N measurement locations for the pollutant species in question. The modelling area is defined with H x W discrete evaluation points with a resolution of e.g., 500m. The current state for $C2_N$ is computed as follows:

$$C2_N = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} p_{hw} SD_{hw}$$
$$C1_N = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} SD_{hw} \qquad (1a\text{-}b)$$

Where p_hw is the population density in the grid cell (h,w) and tSD_hw is the estimated fusion variability in the grid cell. For each of these grid cell (h,w) a virtual candidate sensor is placed in the location, which causes reductions all across the evaluation area when the summation of Eq 1 is updated. The relative improvement for C1 after such a virtual addition somewhere in the area is

$$\Delta C2_N = \frac{C2_N - C2_{N+1}}{C2_N} \qquad (2)$$

The optimization task for C2 therefore is reduced into the optimal selection of (h,w) so that $\Delta C2$ is maximized and this can simply be performed in a brute-force iteration. It should be noted that the improvement is quantified in relative terms (percent) which makes it possible to combine C2 computations for a collection pollutant species. More specifically, an improvement map for C2(N+1) is computed for the whole discrete evaluation grid for each pollutant species S = {s1,s2,…sk} separately. The justification of using pollutant specific user defined weights is that commonly the end users of the provided AQ data prioritize certain pollutant species. As an example, local authorities in Helsinki prioritize PM2.5 and PM10 over O3. With a given preference weights for S (Ws = {w1,w2,…wk}, $\sum w_k = 1$) the combined optimality field can be computed as the weighted affine combination of the pollutant species –specific optimality fields.

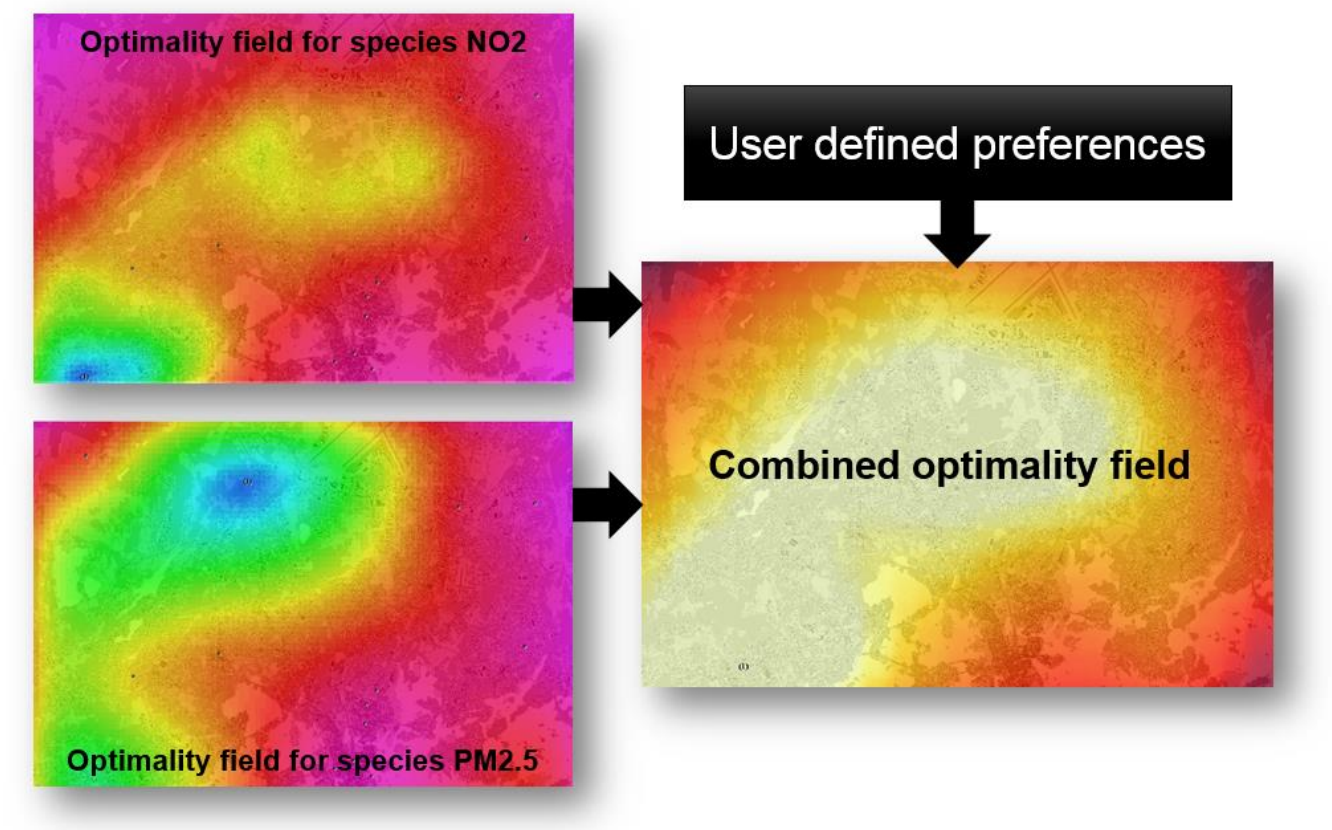

Figure 3: Example of optimality field combination for two selected pollutant species (PM2.5 and NO2) based on user preferences. In the species specific fields blue color corresponds high optimality whereas in the combined field bright yellow shows the areas with high combined optimality.

The optimal sensor location in terms of C1 and C2 is the location (h,w) that has the largest combined improvement. Note: the computation of optimal location with respect to C1 is all but identical, the exception being that the population density that was used in C2 is not taken into account.

2.2 Emission source signal variability (C3)

FMI-ENFUSER model uses measurement data to explain observed pollutant concentrations based on their origin (i.e., by emission categories). The dispersion modelling is performed separately for each emission source category and the most recent measurement evidence is utilized to a) fine tune emission outputs and b) gradually learn to model individual emission sources better over time. As such, on the model's perspective a measurement site offers a signal (measured total concentration) and as a function of geographic, temporal and meteorological factors this signal is processed into emission source specific components. This process is not described in greater detail in this paper; briefly it can be described that the model assesses an emission footprint area nearby each measurement location. The shape and alignment of this emission footprint is affected by meteorological conditions and utilize the Gaussian Plume solution (as referred in Stockie, 2019).

The quality of both the real time adaptation and the gradual learning is affected by the measurement network configuration. The data fusion algorithm of ENFUSER works better when the measurement network is expanded in way that the collection of measurements provide a rich heterogeneous collection of signals that yield information on the local emission sources – for every wind direction and the time of day. **Therefore, for criterion C3 the optimization task is to find a location that improves this collection of emission source signals in a wide range of different temporal- and meteorological conditions.**

The optimality for C3 is more difficult to quantify than for C1 and C2. A strong emission source signal is less valuable if the existing measurement network already provides strong signals for the same emission sources under similar conditions and temporal factors. Indeed, In Helsinki area the existing measurement network for PM10 was designed in a way that most of the measurement locations were nearby major traffic sources with little or no background information for ENFUSER's data fusion algorithm. In this respect, even the alignment of the nearby roads contributes to the general usability of measurement data in the data fusion. In light of these observations, the optimization problem has been defined as follows:

Let there be N stations. Using FMI-EFUSER the observed concentration C in each station (or sensor) is explained as the sum of K emission source sub-components, so that

$$C_n = \sum_{k=1}^{K} c_{kn} \qquad (3)$$

For a specific pollutant source type k, there are N estimates for their contribution in observed total concentrations, and a high variability for these values is favorable for the data fusion process. We calculate this variability of an emission source sub-component as follows[1]:

$$VAR(c_k) = \frac{1}{N}\sum_{i=1}^{N}(c_{kn} - \mu_k)^2 \qquad (4)$$

[1] In C3 the sensor quality has been ignored so far. In this equation the sensor quality (or the product of qualities for n and i) could be factored in.

Where μ_k is the mean value for all c_k and is equal to $\frac{1}{N}\sum c_{kn}$. It should be noted that $VAR(c_k)$ can change significantly depending on the time of day and the meteorological conditions for a fixed collection of measurement locations. To take this effect into account we use a large collection of M evaluation conditions that contain a variety of different wind directions and diurnal hours. Technically, ENFUSER model data including emission source category -specific output is being archived on an hourly basis; for the use of this tool a sufficiently large amount of this stored output data is loaded.

To include all emission source categories in the evaluation, we introduce user defined preference weights for emission source categories (W = {w1,w2,…wK}, $\sum w_k = 1$) for similar reasons as was done for pollutant type preferences; in the modelling area there can be a specific emission source category (e.g., private households or shipping) that is poorly known or of special interest for local authorities and the designed measurement network needs to take such considerations into account.

Finally, the tempo-meteorologically averaged total variability of emission source information for a selected pollutant species with N measurement stations is equal to:

$$C3_N = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} w_k VAR(c_k) \qquad (5)$$

The optimization task for C3 therefore is reduced into the optimal selection of a sensor location so that $\Delta C3$ is maximized via the changes in $VAR(c_k)$.The improvements for C3_N are search in a similar way that was done for C1 and C2: the area is represented with grid structure and a sensor installation in each grid cell is tested and the relative improvements in Eq. 5 are evaluated with

$$\Delta C3_N = \frac{C2_{N+1} - C3_N}{C2_N} \qquad (6)$$

Which is essentially a negated version of Eq. 2 since the objective is to increase C3 whereas the objective for C1 and C2 is to gain reductions. The computations for C3, however, are significantly heavier than for C1-2, but the optimal location can still be assessed with simple brute-force search algorithm once the search area is narrowed down. To achieve this, we search improvements for C3 only in the neighborhood of optimal locations with respect to C1 and C2 (e.g., we isolate the top 10% improvement area for the assessment of C3). The candidate locations in this limited area are represented in a discrete grid HxW with a resolution of 10m. Due to the coarse resolution and its use for area reduction purposes, the combination of C1 and C2 has been referred to (especially in the results section) as the "macro-scale" output of the tool. The finer scale combination where C3 has been taken into account has been called as the "micro-scale" output.

The combination of C3 for multiple pollutant species is performed in a similar way that was described for C1-2; the search iteration is repeated separately for each pollutant species and the resulting improvement fields are combined using the user defined preferences.

1.3 Use of multiple criteria and future additions

In previous sections it was mentioned that the user of the tool can define preference factors for pollutant species as well as for emission source categories. After all, this is ultimately a decision making problem where the user (local authority) preferences impact the definition of optimality and the outcome. After the estimations for optimal locations based on C1, C2 and C3 there is one more combination to be made. The user can define preferences how much C1 optimality field is emphasized over C2, for example.

As a future development, it is possible that the amount of criteria is increased. In particular, there are two potential additions:

1. **Remotely measured total column concentration heatmap**

Using Sentinel S5p TROPOMI satellite remote measurements for total column concentrations, one could create an averaged concentration map for the modelling area. S5p data is fortunately already available in the ENFUSER modelling system and it could add information on the areas that have high local total column concentrations but do not have known local emissions sources mapped in the areas. In such cases, there could be a notable unknown emission source nearby and the C2 criterion does not consider unknown emission sources. Adding a measurement sensor in the vicinity of such locations is naturally appealing.

2. **Local environment complexity**

As it was described above the model utilizes measurement data to gain improved understanding on the local emission sources and their behavior using emission source footprint assessment that uses the Gaussian Plume solution as a basis. In complex urban terrain the footprint assessment becomes more complicated and for this reason there are e.g., additional corrections that are applied in street canyons to simulate the effect of nearby buildings in the area[2]. However, the Helsinki Air Quality Testbed has showed that placing sensor in such complex urban areas can introduce new information that makes it possible to develop modelling capabilities. The model is capable in describing the complexity of the surrounding area in terms of elevation, building heights and vegetation and therefore it could be possible to form a criteria based on these measuring environment features.

3. **Unfeasible location masking**

Due to a number of reasons the given optimal location for the next sensor installation can be unfeasible. The reason may be legal or a practical one – there is no electricity available in the location. The tool already uses land use masking so that the areas within rivers, lakes and sea are filtered out in the assessment iteration. This feature could be improved in the future by introducing, e.g., a mask for electricity availability in case the added sensor cannot be powered by other means.

[2] This is a difficult task and a proper treatment of the urban environment in dispersion modelling would require, e.g., high resolution LES-modelling.

2. Using the tool

The location selection tool has been intended for manual infrequent use and is not connected to the operational ENFUSER model. The user can access this tool via the ENFUSER graphic user interface (Fig x).

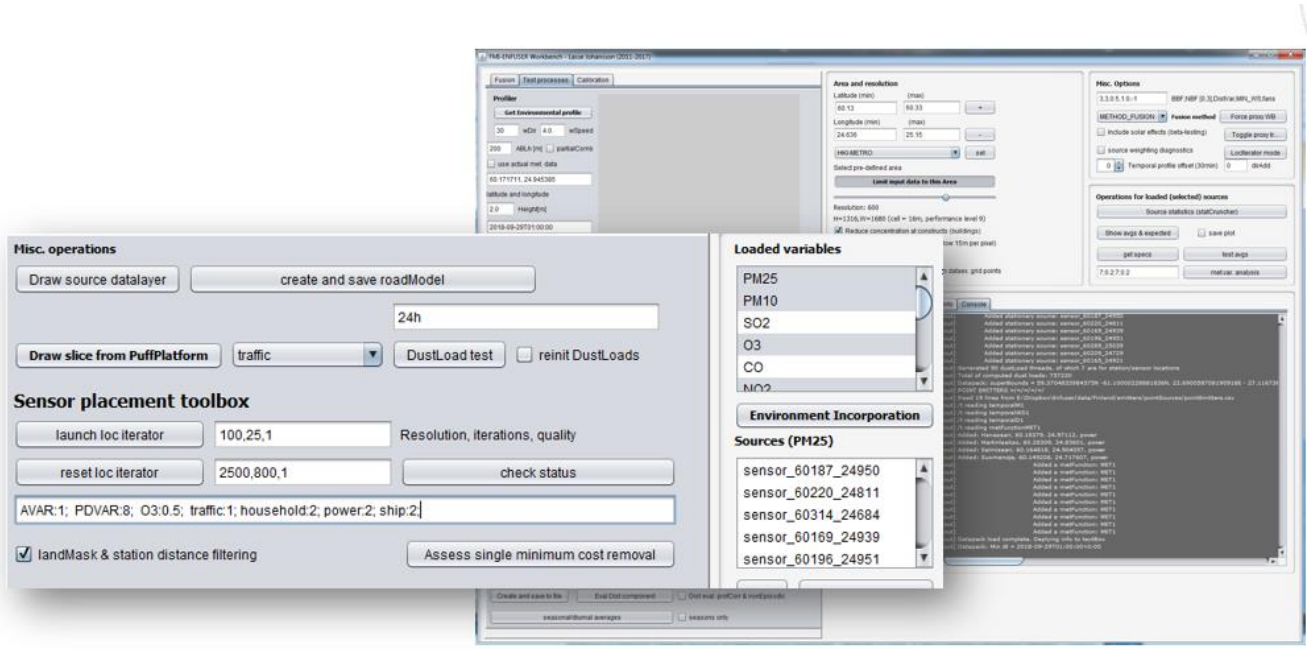

Figure 4: Graphic user interface of FMI-ENFUSER which allows manual processes such as the described sensor location assessment tool to be used.

Before the use of the tool, there a couple of mandatory preparation steps to be made:

- One must have a model installed for the area of interest, including an access to the online measurement data, which in turn defines the starting point for the assessment tool
- For duration of at least a couple of weeks one should use FMI-ENFUSER in an operational model to build-up an archive of modelled pollutant concentrations. This will provide the necessary input data for C3.

To use the tool, one needs to load archived pollutant concentration data from the archives. Then the analysis resolution and the amount of sensor additions (iterations) is selected. Also, the list of pollutant species the devices are able to measure is to be defined using the GUI and the quality rating for this measurement capability is required as input. Finally, the user defined preferences are set. In the example shown in Fig 4, the population factored weighting (C2, PDVAR) has been set to 8 whereas the value for C1 (AVAR) is 1. If a preference has not been specified it is assumed to be equal to 1, and the preference vector is normalized to sum up to 1 automatically.

Once these inputs have been inserted correctly once can launch a location selection iteration for any amount of sensors that are added in sequence. It is possible to, e.g., insert 10 sensors with certain characteristics and then add more sensors with different characteristics; the tool does not forget the

previously assessed locations until the assessment state has been manually reset. For each iteration the tool provides output in the form of figures (as in Figs. 2,3) and Google Earth pins (as in Fig 1).

3. Results

The existing measurement infrastructure in Helsinki Metropolitan Area (HMA) in September 2017 is taken as the basis for our analysis. In this area, depending on pollutant species (O3, PM2.5, NO2 and PM10) there are from 5 to 10 reference quality measurement stations. The stations are largely concentrated on the city center area and there are relatively few stations in the western part of the area. The locations of the existing measurement stations in the area can be seen in Figure 5c (marked with '**o**')

For simplicity the measurement quality for the reference stations in the area have been associated with a standard deviation rating of 1. The 25 sensors (of which 15 has been actually installed in the area) for which the optimal locations are assessed, have associated with a respective SD rating of 10^3. The user defined preferences has been set as shown in Fig 4. Many of the existing reference station are already located nearby road traffic and the emphasis on traffic emissions is slightly reduced. The pollutant species NO2, PM2.5 and PM10 have been associated with a higher priority than O3 in the Helsinki area.

The location assessment tool was launched for 25 sensor placement iterations. The location for the first added sensor is shown in Figure 5. The first sensor location is suggested for the western part of the area where there are now reference stations nearby, very close to a major highway.

[3] By the time of analysis there was no extensive information available that could've been used to estimate a more appropriate value for sensor standard deviation. The used value is likely to be an underestimation.

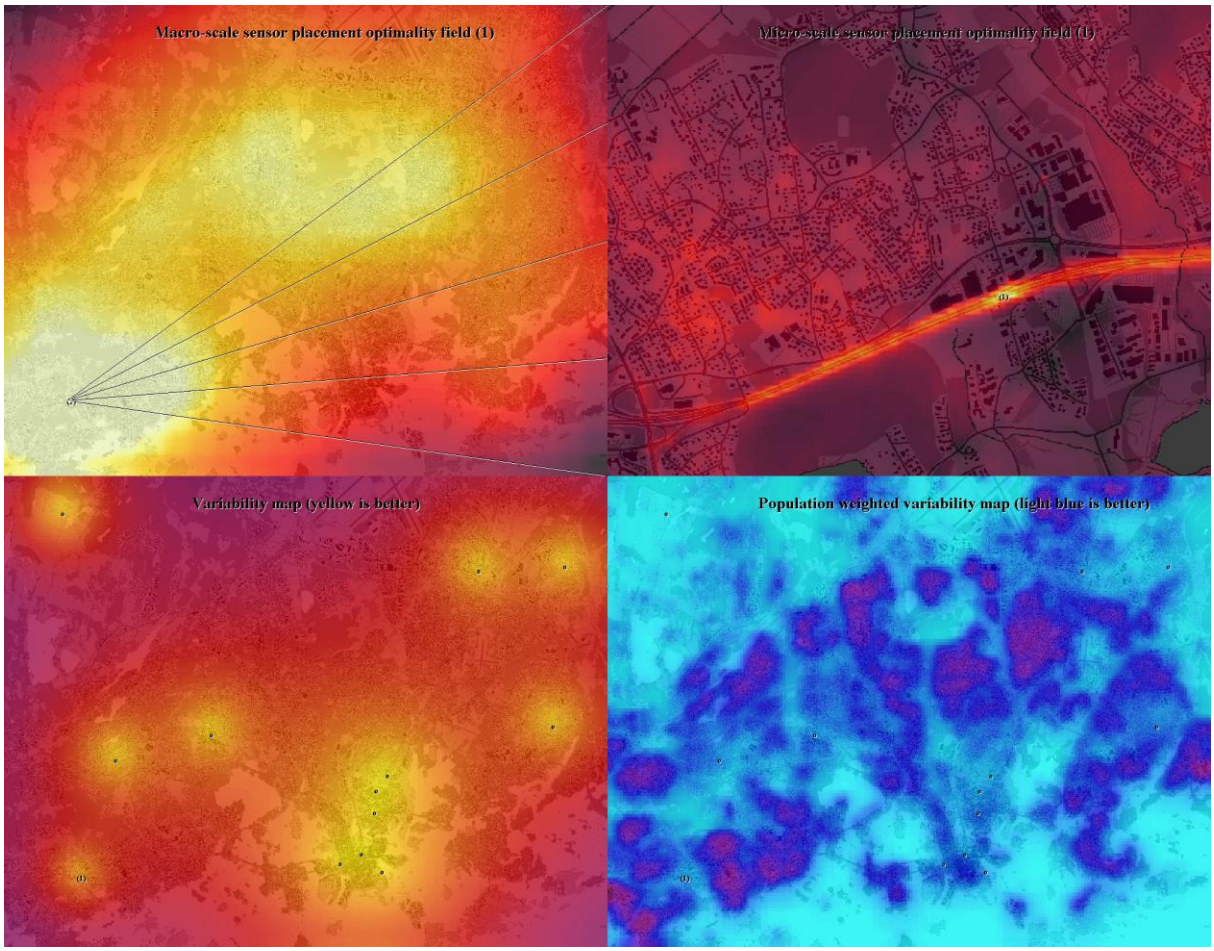

Figure 5a-d: The assessed optimal location for the first sensor. In upper left figure (a) the combination of C1 and C2 (macro-scale) is shown. In upper right figure (b) the micro-scale optimal location for a final combination of C1-C3 is shown. The lower figures show the state of C1 (c) and C2 (d) after the addition of the first sensor.

The next couple of sensors in the iteration also prioritize locations nearby traffic sources but the alignment of the nearby road with respect to the sensor location varies. During the 5th iteration the micro-scale evaluation a location nearby a residential area is prioritized over locations that have roads nearby (Fig. 6).

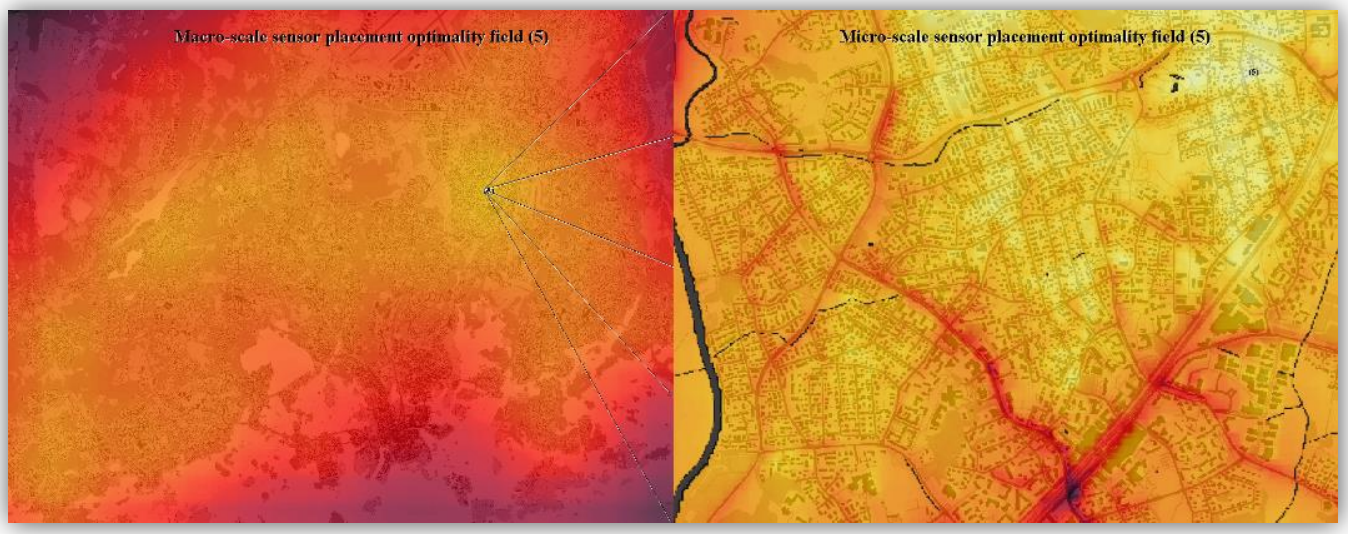

Figure 6: Optimal sensor location for the 5th iteration.

The final state of the suggested sensor network of 25 sensors is shown in Figure 7a-d. For the center of Helsinki there were no added sensor during the iteration. With the used parameters the 28th added sensor would be the first one to be located nearby the center. There are multiple sensors suggested for residential areas, targeting domestic household PM2.5 emission source signals. One sensor is suggested to be placed at the Vuosaari cargo shipping port. There are surprisingly many traffic measurement sites included with only a few background locations. This can be explained by the fact that depending on the wind direction many of the traffic location sensors work effectively as a source for background information.

It can be seen from Fig 7a, that the addition of the last sensor causes only marginal improvements for the combination of C1 and C2 criteria (Dark red colors illustrate marginal relative improvements and for example, in Fig 5a the addition of the first sensor improves the macro-scale criteria combination significantly more, which is illustrated with bright yellow colors). In fact, with each added sensor there are diminishing returns for the states of C1 and C2 given by Eqs. 1a-b. The progression of these criteria states during the iteration have been shown in Fig 8.

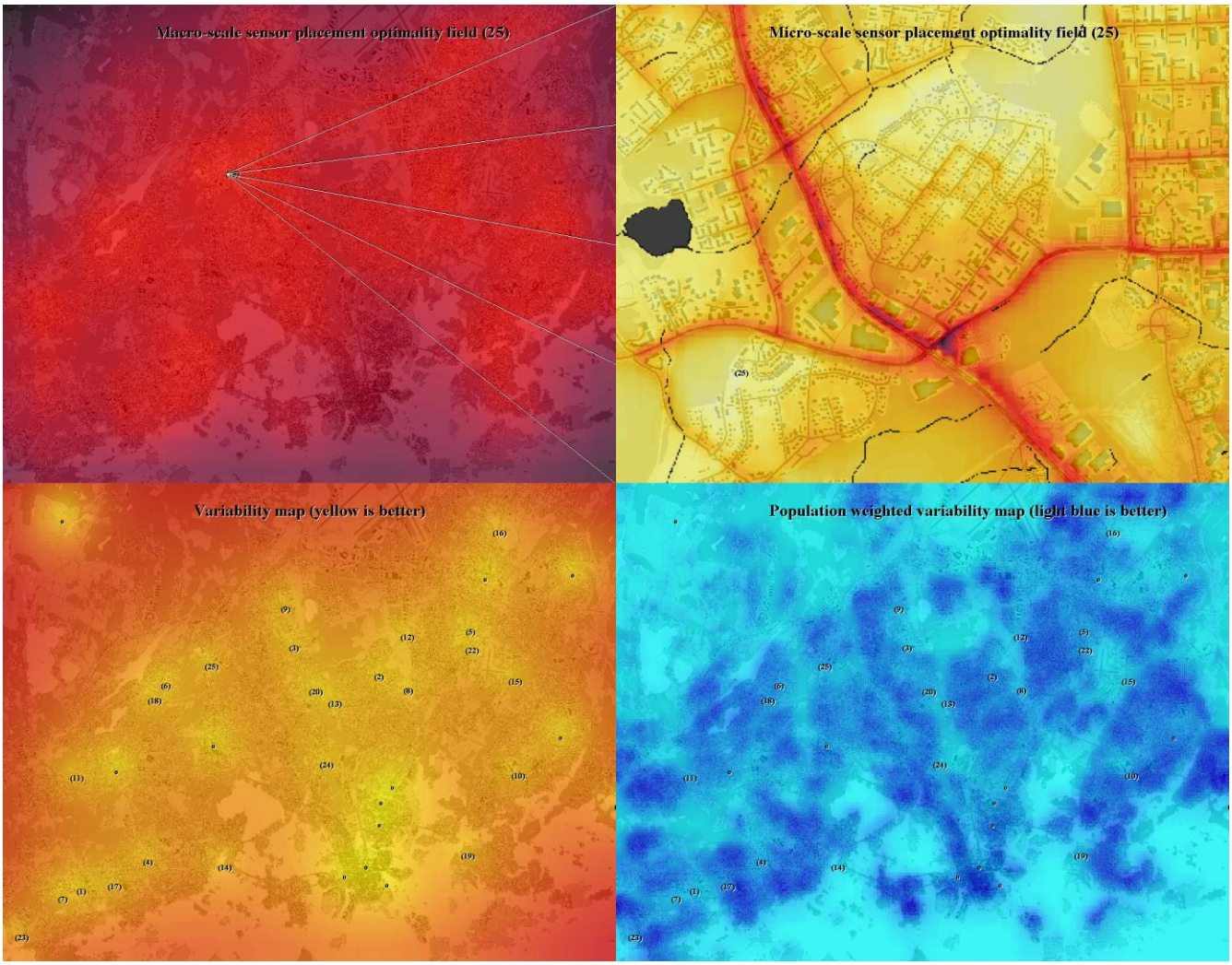

Figure 7:a-d: The assessed final optimal locations for the first sensors at the end of the iteration. In upper figures (a,b) the last iteration optimal location is shown. The lower figures show the state of C1 (c) and C2 (d) after the addition of the last sensor.

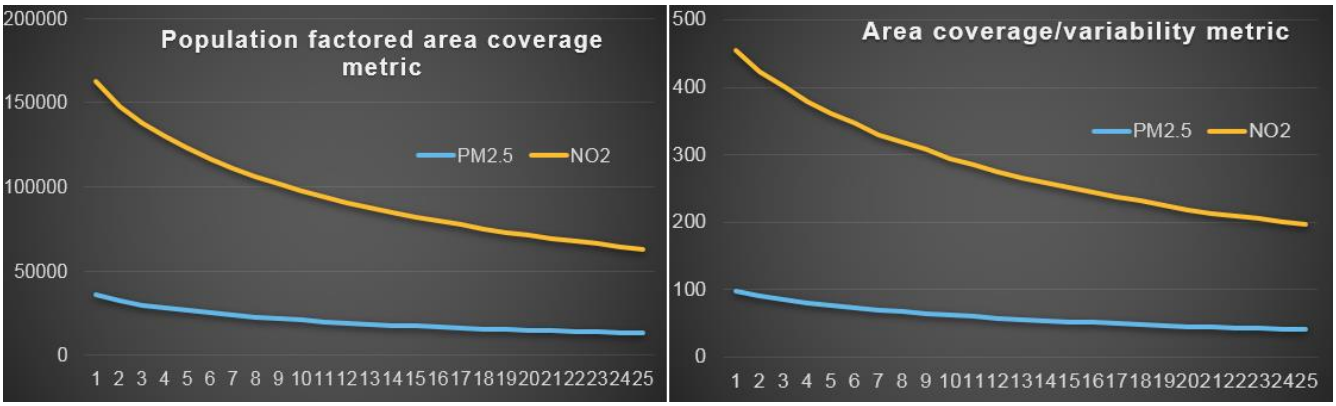

Figure 8: Progression of states for criteria C1 and C2 during the 25 iterations.

References

[1] Hsieh, H. P., Lin, S. D., & Zheng, Y. (2015, August). Inferring air quality for station location recommendation based on urban big data. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 437-446). ACM.

[2] Yang, C., Kaplan, L., Blasch, E., & Bakich, M. (2013). Optimal placement of heterogeneous sensors for targets with Gaussian priors. IEEE Transactions on Aerospace and Electronic Systems, 49(3), 1637-1653.

[3] Mofarrah, A., & Husain, T. (2010). A holistic approach for optimal design of air quality monitoring network expansion in an urban area. Atmospheric Environment, 44(3), 432-440.

[4] Johansson, L., Epitropou, V., Karatzas, K., Karppinen, K., Wanner, L., Vrochidis, S., Bassoukos, A., Kukkonen, J. and Kompatsiaris I. Fusion of meteorological and air quality data extracted from the web for personalized environmental information services. Environmental Modelling & Software, Elsevier, 64 (2015) 143-155.

[5] Stockie, J.M., The Mathematics of Atmospheric Dispersion Modeling, SIAM Rev., 53(2), 349–372. (24 pages), DOI:10.1137/10080991X, 2011.