The Extraction and Fusion of Meteorological and Air Quality Information for Orchestrated Services

Lasse Johansson
The Finnish Meteorological Institute,
Dept. of Atmospheric composition
Erik Palmenin aukio 1
00101, Helsinki, Finland
lasse.johansson@fmi.fi

Victor Epitropou
and Kostas Karatzas
Aristotle University of Thessaloniki,
Dept. of Mechanical Engineering,
54124 Thessaloniki, Greece

Leo Wanner
Catalan Institute for Research and Advanced Studies,
Dept. of Information and Communication Technologies,
Pompeu Fabra University, Barcelona, Spain

Ari Karppinen
and Jaakko Kukkonen
The Finnish Meteorological Institute,
Dept. of Atmospheric composition

Stefanos Vrochidis
and Ioannis Kompatsiaris
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece

ABSTRACT

The PESCaDO system (Personal Environmental Service Configuration and Delivery Orchestration) aims at providing accurate and timely information about local air quality and weather conditions in Europe. The system receives environment related queries from end users, discovers reliable environmental multimedia data in the web from different providers and processes these data in order to convert them into information and knowledge. Finally, the system uses the produced information to provide the end user a personalized response. In this paper, we present the general architecture of the above mentioned system, focusing on the extraction and fusion of multimedia environmental data. The main research contribution of the proposed system is a novel information fusion method based on statistical regression modelling that uses as input data land use and population density masks, historic track-record of data providers as well as an array of atmospheric measurements at various locations. An implementation of this fusion model has been successfully tested against two selected datasets on air pollutant concentrations and ambient air temperatures.

1. INTRODUCTION

Recently, the emergence of social media, personalized web services and the increased public awareness of environmental conditions that impact the quality of life have resulted in the demand for easier access to environmental information tailored to personal requirements. In particular, in case of the atmospheric environment, there is a need for an integrated assessment of the impact of air pollution, allergens and extreme meteorological conditions on public health [9], [8]. In addition, this information has to be disseminated to citizens in an easily accessible form [7].

In: S. Vrochidis, K. Karatzas, A. Karpinnen, A. Joly (eds.): Proceedings of the International Workshop on Environmental Multimedia Retrieval (EMR2014), Glasgow, UK, April 1, 2014, published at http://ceur-ws.org

Getting a direct answer to a seemingly simple question such as "How will the air quality be tomorrow in Glasgow?" involves extensive manual search and expert interpretation of the often contradictory and heterogeneous information found on various web sites. Furthermore, a significant portion of air quality and meteorological information is published on the Internet only in the form of colour-mapped, geo-referenced images [1]. Also the quality of information might vary significantly in reliability and relevance with respect to the queried location and time. On the other hand, even biased and inaccurate information about air quality could be utilized effectively by data fusion methods in order to provide reliable information. The success of fusing multiple model results is evident in the case of models with no major deviation of forecasting performance, and has been demonstrated in many related studies [22].

In this context, in [5] it has been presented an approach to provide air quality information for any location within a large geographical domain, by fusing air quality data from multiple sources, by using a statistical air pollution model (RIO). In a review of land use regression (LUR) models it has been stated that LUR-models have been very successful in predicting annual mean concentrations of NO_2 and $PM_{2.5}$ in urban environments [4]. However, these state-of-the-art LUR models are difficult to utilize for the accurate prediction of hourly concentration of air pollutants – a more dynamic approach is needed. Another complication is the extremely heterogeneous nature of input data which may contain model forecasts and observations, both with varying reliability, time of validity and location. Spatial and temporal gaps are also a matter of concern; there are only a finite number of measurement stations, and forecasting models also have a finite spatial and temporal resolution. These considerations lead to the need to use some form of data interpolation either in space or time, or both.

In this paper, we aim to describe the general architecture of the PESCaDO system, focusing especially on the fusion of extracted information [20], [21]. First, we discover environmental nodes (i.e. web resources that include environmental measurements), which are relevant to the area of interest. Then, a specific service called AirMerge is presented, which is capable of performing extraction and fusion of information from a wide range of online Chemical Weather (CW) forecasting systems. The online fusion service is then presented; this is a general method for the fusion of

processed meteorological and air quality data, and is also the main topic of this paper. There are many definitions of data fusion, as it is a method that is applied to various scientific domains, such as remote sensing, meteorological forecasting, sensor networks, etc. [19]. We use the term "fusion" to describe the process of integration of multiple data and knowledge into a consistent, accurate, and useful representation. An evaluation of the performance of this fusion system is presented for two selected cases: i) the fusion of atmospheric temperature forecasts and (ii) the fusion of measured NO_2 concentrations.

2. FRAMEWORK

We present here an overview of the general architecture of the PESCaDO system. For a more detailed description, the reader is referred to [20], [21].

2.1 An overview of the PESCaDO system

The purpose of the PESCaDO system is to address the need for timely personalized environmental information (see www.pescado-project.eu for more information). It first processes user queries, based on the personal information on the user, formulated in terms of a user profile. For instance, health conditions such as asthma may affect the displayed warnings and recommendations while the user group (e.g. citizen or administrative expert) affects the level of detail and technicality of the response.

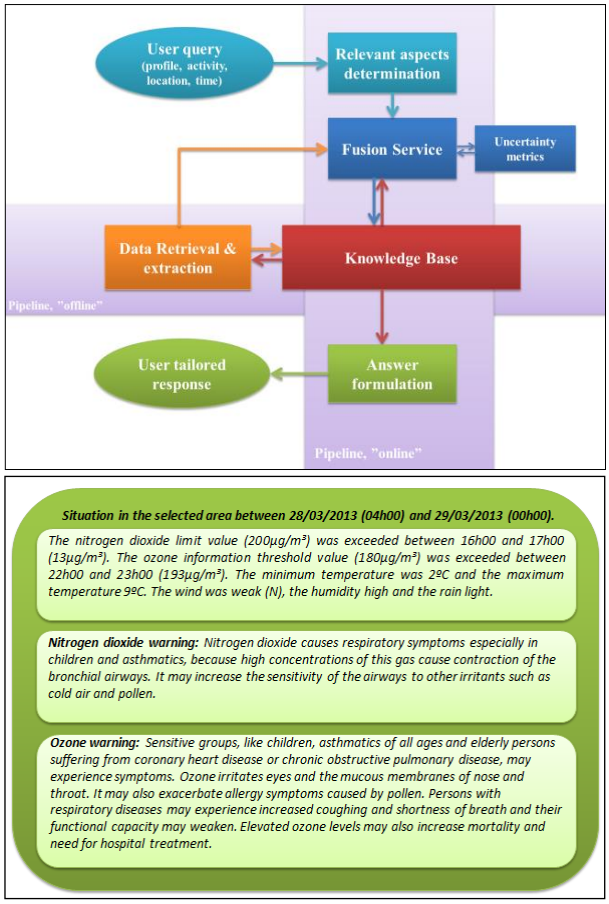

Figure 1a-b: A simplified schematic diagram of the PESCaDO system, starting from the user defined query and ending at the delivery of response (a). An example response for the user is presented in figure (b).

The queries are formulated in terms of PESCaDO's Problem Description Language via an interactive web interface. First, the system discovers environmental nodes that contain measurements for the areas of interest. Then, for each query, (i) relevant environmental data sources are orchestrated, (ii) data from textual and image formats in the sources are identified, extracted, fused and reasoned over to assess the relevance of the data for the user, and (iii) his query and the outcome are presented in terms of a bulletin in the language of the preference of the user.

Figure 1a illustrates the information flow of PESCaDO from the viewpoint of the Fusion Service, which is the backbone of the system. The system includes two uncoupled process chains, called here as pipelines, that operate in offline and online modes. In the offline pipeline, environmental websites that cover the region targeted by the user are searched for in the web and data are extracted from the identified sites and fed into the database of the system. We use the term 'offline' here since at the time of user query the data used by the pipeline has already been retrieved, processed and stored into a local database. In the online pipeline user queries are processed and answered. The online pipeline starts from the specification of personal information and query by the user. With this information, the system first determines which aspects of environmental and contextual knowledge (e.g. temperature, CO_2 concentration, etc.) are relevant to the user and his query (cf. Fig, 1, Relevant aspects determination). Next, the Fusion Service (FS) is given a request to produce fused information about the identified relevant aspects. At this stage, the system retrieves information from the database and starts to process it. The 'relevant aspects' could be, for instance, "NO_2 concentration and ambient air temperature, tomorrow between 12:00 and 18:00 in a specified region in Helsinki, given the reported traffic density". Furthermore, the user profile (administration personnel vs. citizen; healthy individual vs. allergic, etc.) affects the way the response is ultimately presented to the user (relevant aspects determination).

The Data Retrieval Service (DRS) serves as an interface, through which other PESCADO services can retrieve information (i.e. environmental measurements) from the database. The Fusion Service queries the DRS to receive environmental data available for the requested geographic areas and time periods for all related environmental aspects. After the FS fuses the data retrieved from the DRS these are inserted to the PESCaDO Knowledge Base (KB).

The PESCaDO's KB contains, manages and provides information represented with the PESCaDO ontology to other services [13]. This KB also provides the Fusion Service with supporting information needed in the fusion process. This includes source identification and fixed coordinates if available, and source reliability. Furthermore, the PESCaDO ontology helps to translate verbal ratings into numeric form if needed. For instance, the expression "heavy rain" can be converted into mm/h numeric value with the help of the concept definitions in the ontology. More specifically, the KB is queried about the upper and lower limit for "heavy rain" in the specified region and then the average value of the returned limits can be taken to represent the input in numeric format - an approach related to the use of fuzzy logic methods in air quality problems [6].

Once all input data are in numeric form, the FS fuses the data by one variable (e.g. temperature, wind speed, NO_2 or O_3) at a time, utilizing available uncertainty metrics for each information source given by the Uncertainty Metrics tool (UMT). Fused data are stored in the KB and then the tasks, including the selection, structuring and presentation of the information resulting from the fusion to the user can be carried on. In parallel, the retrieved

information, which can be used for performance evaluation later on, is passed to UMT and stored. Using this stored information, UMT evaluates measured values against forecasts autonomously and produces updated source node uncertainty metrics.

2.2 Discovery of environmental nodes

As described in the previous section, the first step realized by the PESCaDO framework is the discovery of environmental nodes. The huge number of the nodes, their diversity both in purpose and content, as well as, their widely varying and a priori unknown quality, set several challenges for the discovery and the orchestration of these services [21].

The PESCaDO discovery framework combines the main two methodologies of internet domain specific search: (a) the use of existing search engines for the submission of domain-specific automatically generated queries, and (b) focused crawling of predetermined websites [23]. To support domain-specific search using a general purpose search engine [12], two types of domain specific queries are being formulated: the basic and the extended. Basic queries are produced by combining environmental related keywords (e.g. weather, temperature) with geographical data (e.g. city names). Extended queries are generated by enhancing the basic queries with additional domain-specific keywords, which are produced using the keyword spice technique [14]. Both types of queries are then submitted to Yahoo BOSS API search engine.

In parallel, a focused crawler is employed, built upon the Apache Nutch -crawler and is based on [18]. This implementation attempts to classify sites by using hyperlink and text information (i.e. anchor text and text around the link) with the aid of a supervised classifier. This approach is new in comparison to a previously presented method for web-based information identification and retrieval with the aid of a domain vocabulary and web-crawling tools [2].

The output of both techniques is post-processed in order to improve the precision of the results by separating relevant from irrelevant nodes and categorizing and further filtering the relevant nodes with respect to the types of environmental data they provide (air quality, pollen, weather, etc.). The determination of the relevance of the nodes and their categorization is done using a supervised classification method based on Support Vector Machines (SVM). The SVM classifiers are trained with manually annotated websites and textual and visual features extracted from the environmental nodes. The textual features are key phrases and concepts extracted from the metadata and content of the webpages using KX [15] and the vector representation is based on the bag of words model. The visual features (MPEG-7, [17]) are extracted from the images included in the discovered websites in order to identify heatmaps that are usually present in air quality forecast websites.

2.3 Orchestration of environmental nodes and data extraction

Once the environmental nodes have been detected and indexed, they are available as data sources or as active data consuming services (if they require external data and are accessible via a web service API).

To distil data from text, advanced natural language parsing techniques are applied, while to transform semi-structured web content into structured data, regular expressions and HTML trees are used. Data extraction from images focused on heatmap analysis using the AirMerge system, described in the following section.

2.4 AirMerge subsystem

A significant portion of Air Quality (AQ) related information (in particular, Chemical Weather forecasts) is published on the Internet only in the form of colour-mapped geo-referenced images. Such image-based information is impossible to be parsed via usual text-mining and screen-scraping techniques used in web mash-up-like services. It was thus important to provide PESCaDO with a specialized service that allows accessing and using CW forecast images as another source of data to use during the Orchestration and Fusion phases. Such a system, called Air Merge, has already been developed and described in [3], [1].

AirMerge is an open access system, which is currently dedicated to the whole European continent (the coverage of different territories is possible, accessing a wide number of environmental nodes containing CW information, and can automatically extract data from various data sources). These images commonly have geographical spatial resolutions ranging from 1x1 km to 20x20 km, and temporal resolutions from a minimum of one hour to an entire day [10]. The reported values usually are maximum or average air pollution concentration values for the selected integration time.

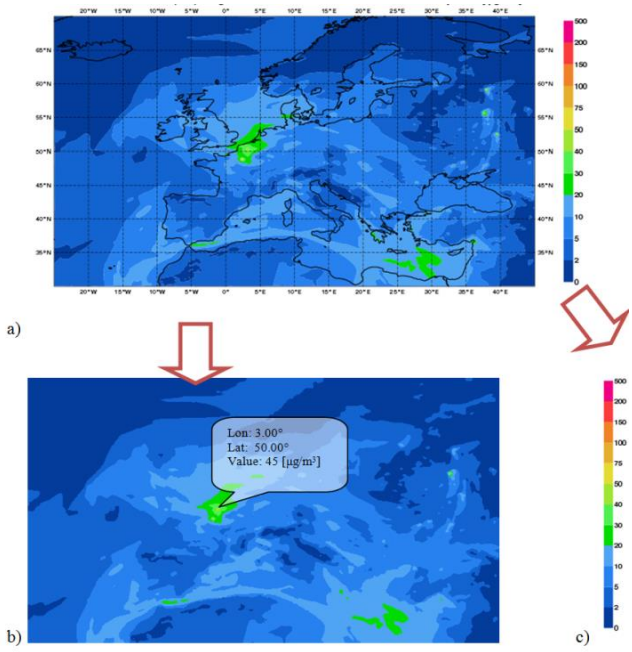

Figure 2: Example of a $PM_{2.5}$ forecast (produced by MACC) conversion process using AirMerge. Bitmap data (a) is transformed into numerical form by using the colour scale c). The heatmap a) has been reproduced in b) using the converted numeric grid.

In the context of PESCaDO, AirMerge apart from performing image extraction, it acts as an autonomous web-crawling, parsing and database-storage mechanism for CW forecasts, using its own means and processes which are distinct from those of PESCaDO, having been developed independently. The harvested data cover most of Europe for a time period going back to August 2010 when it first became operational. Time resolutions range from one hour to a day, depending on the capabilities of the sources used.

A typical set of CW models and the resulting images can be found in the European Open-access Chemical Weather Forecasting Portal described by [1], that has been developed in

the frame of COST Action ES0602 (www.chemicalweather.eu). AirMerge is able to convert such image-based concentration maps into numerical, geographically referenced data, accounting for geographical projections, missing data, noise and the differences in publishing formats between different model providers. The result is the effective conversion of image data back into numerical data, which is then made directly available for a number of numerical processing applications.

It should be clarified that in the proposed system AirMerge has two roles: a) it performs image data extraction and b) it is an additional environmental node that provides environmental data encoded in images.

3. FUSION OF EXTRACTED INFORMATION

The fusion of information in an orchestrated service such as PESCaDO, offers several advantages to the user. First, the output of the system includes only one set of values instead of an extensive collection of pieces of information that may not agree with each other. Secondly, the fusion result will be of a better quality with respect to the individual sources. Third, small geographic and temporal gaps in the input data can be extrapolated.

The above mentioned services for environmental node discovery and data retrieval guarantee a large amount of relevant input data which need to be fused with respect to the user defined query. However individual competing pieces of information from different nodes can seldom be regarded as equally relevant and thus a general measure for information relevance and quality is needed for data fusion.

In the fusion process, all pieces of meteorological and air quality data correspond to a certain time and place. These pieces of information can be regarded as statistical estimators $\theta_i(\mathbf{r_i}, t_i)$ or θ_i in short, in which $\mathbf{r_i}$ is distance and t_i is time, for the conditions governing the area and time of interest for the user:

$$\theta_0(\mathbf{r}_0, t_0) = \theta_i(\mathbf{r_i}, t_i) + \varepsilon_i \quad (1)$$

where $\mathbf{r}_0$ / $\mathbf{r_i}$ is the coordinate vector for the location of interest / location associated with the estimator, t_0 / t_i is the time of interest / estimator time and ε_i is the estimator error. For sensors the estimator time is simply the time of measurement. The algorithm that is used in calculating the fused value requires information about the statistical properties of ε_i, namely the expected variance of ε_i. Thus, a detailed description of the evaluation of $VAR[\varepsilon_i]$ is given. The fusion service estimates an aggregate statistical variance measure for each ε_i and these variance measures are then used for the assignment of averaging weights to each $\theta_i(\mathbf{r_i}, t_i)$. Essentially a large estimated aggregate variance causes the assigned weight to decrease, while the data from the more accurate and relevant sources are assigned larger weights and gain more emphasis in the fusion.

3.1 Variance estimation

The variance of ε_i, $VAR[\varepsilon_i]$, is affected by the information source's capability to properly assess the phenomenon of interest. In addition, information about air pollutant concentrations and weather conditions loses accuracy rapidly as a function of the temporal interval between the measurement time and the time of interest defined by the user. Furthermore, a data point near $\mathbf{r_0}$ should always get a larger weight in the fusion in contrast to other data points that describes the conditions in more remote locations. Thus, we assume that the variance related to ε_i is the sum of these three individual (independent and thus summable) components, given by

$$VAR[\varepsilon_i] \approx f(d) + g(\tau) + VAR[\theta_i(\mathbf{r_0}, t_0)] \quad (2)$$

where $f(d)$ is the variance component as function of d, $g(\tau)$ is the temporal variance component as a function of τ, in which

$$d = ||\mathbf{r}_0 - \mathbf{r}_i|| \quad (3a)$$

$$\tau = |t_0 - t_i| \quad (3b)$$

$VAR[\theta_i(\mathbf{r_0}, t_0)]$ in Eq. 2 describes the information source's inherent quality in terms of variance, i.e., the capability to estimate $\theta_0(\mathbf{r}_0, t_0)$ at point-blank range when d and τ are equal to zero. For the evaluation of $VAR[\theta_i(\mathbf{r_0}, t_0)]$, stored information about the source's prediction accuracy in past can be used, evaluated by the Uncertainty Metrics Tool (see Fig 1). More specifically, measurements and model forecasts are paired together if they represent the same time and location and the statistical variance is then calculated for the population of evaluation pairs.

In the presented PESCaDO framework, the location $\mathbf{r_i}$ for the estimator $\theta_i(\mathbf{r_i}, t_i)$ may not have been defined exactly; this is usually the case, for instance, with extracted weather forecasts for cities. In these cases $\mathbf{r_i}$ actually pinpoints the center of city while information represents the conditions through-out the city. In such cases the coordinates are flagged as approximations and set $d_i = min\{r_c, ||\mathbf{r}_0 - \mathbf{r}_i||\}$, where r_c is the radius of the city.

The variance models $f(d)$ and $g(\tau)$ can be formulated with statistical methods. In the fusion service these have been formulated individually for each air pollutant species using regression analysis with historical measurement data. For the pilot application of the method, these data represent 6 to 43 measurement stations across Finland, depending on the measured values. More specifically, the following simple regression models are employed:

$$g(\tau) = a_3\,\tau^3 + a_2\,\tau^2 + a_1\,\tau + a \quad (4a)$$

$$f(d) = b_1 d \quad (4b)$$

where parameters $a_3 \ldots a$ and b_1 are defined with statistical regression techniques. More complex regression models were also studied but the added benefit for using more natural, logarithmic regression models was negligible; the achieved correlation of $g(\tau)$ polynomial models is generally very high for the temporal domain of interest ($\tau < 36h$). In the formulation of $f(d)$, the measurement station's capability to predict the measured phenomenon at a distance of d (covariance of the two time series) is evaluated.

3.2 Optimal weight calculation

Assuming all data sources to be independent and the estimators to be non-biased ($E[\varepsilon_i] = 0$), an optimal fused value $\theta_F(r_0, t_0)$ can be calculated according to [16] given by:

$$\theta_F(\mathbf{r_0}, t_0 = \sum_{i=1}^{n} w_i \theta_i(\mathbf{r_i}, t_i) \quad (5)$$

where individual weights w_i is given by

$$w_i = \frac{VAR[\varepsilon_i]^{-1}}{\sum_{i=1}^{n} VAR[\varepsilon_i]^{-1}} \quad (6)$$

To assure statistical independence of $\theta_1 .. \theta_n$, only the most relevant estimator θ_i per data source is selected for the fused value calculation in Eq. 5. If a collection of estimators $\{\theta_1(\mathbf{r_1}, t_1), \ldots \theta_k(\mathbf{r_k}, t_k)\}$ is available from the same source, the selected θ_i to represent the source is simply the one with the lowest $VAR[\varepsilon_i]$ from the collection. In the particular case for extracted time series from measurement stations, the estimator which has the smallest τ is selected to represent the source, as d and the base variance are the same for all $\theta_1 .. \theta_k$.

Theoretically, it can be shown that the fused value $\theta_F(r_0, t_0)$ is the optimal estimator in terms of mean squared error and that the prediction accuracy increases while the number of independent data sources (n) is increased [16]. More importantly, $\theta_F(r_0, t_0)$ does not suffer from low quality input data, as long as $VAR[\varepsilon_i]$ in Eq. 2 has been estimated reasonably well.

3.3 Bias correction

In the algorithm presented in section 3.2, it was assumed that each θ_i is an unbiased estimator for the conditions in $\mathbf{r_0}$ at the time t_0. Local air quality measurements from a different environment, however, are usually significantly biased estimators for the conditions in other nearby environments. Moreover, the hour of day may even contribute to the bias (consider a measurement station near a busy road during the morning traffic). Thus, in order to use Eq. 5 effectively, the fusion service utilizes a geographic profiling feature to detect and automatically remove this kind of structural bias from the estimators. The fusion service was incorporated with high-resolution land use and population density masks for Finland (the selected domain for the PESCaDO prototype). For land-use, a dataset from CORINE with a resolution of 50m x 50m is being used. For population density data (for 2010), the fusion service has the prototype domain covered with a resolution of 250m x 250m. These two data sources are used for profiling and comparing the differences between the environments in $\mathbf{r_i}$ and $\mathbf{r_0}$ and ultimately, $\theta_i(\mathbf{r_i}, t_i)$ is polished into a non-biased estimator for $\theta_0(\mathbf{r}_0, t_0)$. The profiling is done as follows:

- The surrounding land use (with evaluation radius of 200m) and population density (a wider evaluation radius of 6km) for both $\mathbf{r_i}$ and $\mathbf{r_0}$ is evaluated.
- The evaluated environment is expressed as a collection of selected land-use frequencies and population density. This collection is referred to as a profile in this paper (Fig 3).

After the evaluation of profiles, the difference between the expected values is evaluated. Let $p(\mathbf{r_i}, t_i)$ be the estimator profile and $p(\mathbf{r_0}, t_0)$ be the evaluated profile corresponding to the user defined location and time. Then, a bias corrected estimator $\theta_i'(\mathbf{r_i}, t_i)$ is given by

$$\theta_i'(\mathbf{r_i}, t_i) = \theta(\mathbf{r_i}, t_i) - (E[\theta_i(\mathbf{r_i}, t_i)] - E[\theta(\mathbf{r_0}, t_0)] \quad (7)$$

where $E[\theta_i(\mathbf{r_i}, t_i)]$ is the expected hourly concentration of the pollutant at the estimator's location at time t_i and $E[\theta(\mathbf{r_0}, t_0)]$ is the expected pollutant concentration in the user defined location at the time t_0.

The evaluation of Eq. 7 requires yet another statistical model (for each pollutant) to calculate the expected concentration as a function of time and key land-use frequencies. Such a set of statistical models has been implemented with the fusion service, using the archived measurement time series in Finland as calibration data: the environments around the stations were evaluated and multi-variable regression was applied. The regression was repeated with several different land-use and population evaluation radii; the best correlation was achieved with the abovementioned values (land-use with a 200m radius, population density with a 6km radius). Nevertheless, this mathematically intensive regression procedure is not discussed in this paper further although for the NO_2 pollutant, a demonstration of the profiling method and its capability to predict the expected hourly concentration is presented in section 4.1.

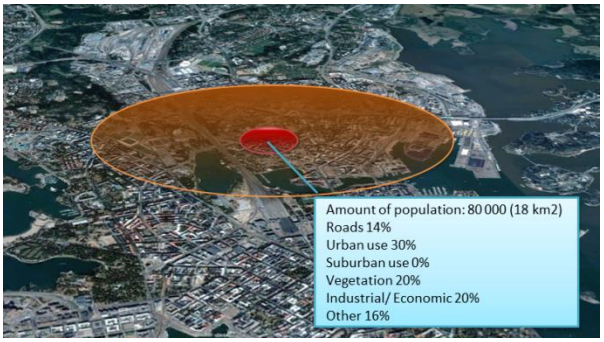

Figure 3: Profile evaluation with land use and population density maps. The larger circle represents the area for local population determination and the smaller red circle represents the area for land use determination. Satellite image provided by Google Earth.

As discussed in at the beginning of section 2.1 the fusion service stores measurements as evaluation material for individual service providers and models. Thus for another completely different region other than Finland, the regression parameters for profiling can be set without a fixed set of calibration material; the stored measurements that have flown through the PESCaDO system can be further exploited by setting up the regression parameters for profiling automatically as the number of measurements builds up over time. In this sense the profiling feature within the Fusion Service is adaptive.

The presented bias correction method offers yet another advantage: episodes that affect air quality on a major scale, such as forest fires, are automatically accounted for if the input data contains some measurements from the episode-driven locations. For instance, if a background station has measured an exceptionally high concentration of NO_2, then the expected NO_2 concentration at a nearby urban environment is going to be reflected on the episode-affected background concentration.

4. RESULTS

The performance of the presented environmental information fusion method was evaluated using temperature forecasts provided by four well known weather service providers (FMI, SMHI, Met Norway and Weather Underground). For 43 locations around Finland weather forecasts were extracted from respective online sites and stored during several months in 2012. Uncertainty metrics in terms of $VAR[\theta_i(\mathbf{r_0}, t_0)]$ for individual SPs were evaluated by comparing measured temperature values against individual stored forecasts for each SP; a total of 2500 forecasted versus measured temperature -pairs for each SP were gathered in order to get statistically meaningful $VAR[\theta_i(\mathbf{r_0}, t_0)]$ estimates as a function of forecasted period length. Then, fused forecasts (temperature of the next 3 days) for the locations in August 2012 were produced on a daily basis for each of these locations using the stored forecasts.

In Figure 4, the mean absolute error of temperature forecasts and the fused forecast is presented. According to the figure fused temperature forecasts have the lowest mean error with just four different SPs providing forecasts simultaneously. This result goes to show that the well-known benefits of forecast fusion can be exploited within web services such as PESCaDO when the performance of forecast providers is being monitored.

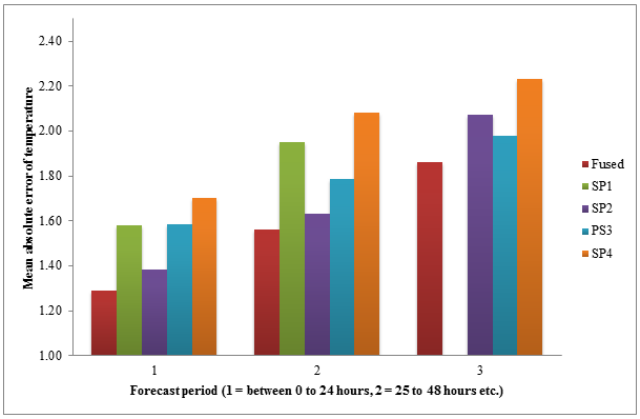

Figure 4: Mean absolute error of temperature (C) forecasts and the fused forecast for different forecast time spans. Forecasted and measured data for 43 different locations and time periods in august was used.

4.1 Performance of the environmental profiling feature

The environmental profiling feature of the fusion service was calibrated using measurement time series from Finland during 2010. To test the performance of this novel feature, 8 different NO_2 measurement stations with varying environments were selected in 2011, and the observed hourly concentrations were compared against the values predicted with the aid of the profiling feature. The profiling feature differentiates working days and weekends and for this test, the working days were selected.

It can be seen from the figures 5a-h that the profiling feature is able to predict the expected average NO_2 concentration well in various different environments. Background areas, urban and rural, fare better in the comparison while the traffic-intense environments are more difficult to predict. This is to be expected as the actual traffic volumes have to be derived using only the local population and road intensity. As a consequence, the profiling feature inevitably underestimates the expected concentration near large motorways that have a small surrounding population.

4.2 Comparison of measured and predicted NO_2 time series

The performance of the fusion of air quality measurements with the presented methodology was tested with NO_2 measurements in Southern Finland. Measurement time series for February 2011 from the available stations (n = 20) were used as input data and fused NO_2 concentrations were calculated for a remote location for which comparison time series was readily available. The domain for the test can be seen from Figure 6 which illustrates the fused concentration of NO_2 at one of the hours of interest.

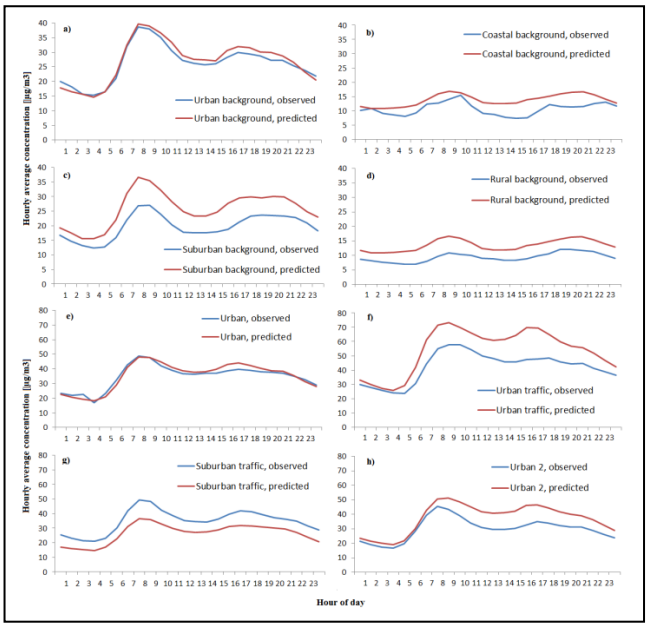

Figure 5a-h: Predicted and observed hourly average concentration of NO_2 during working days (Monday to Friday) in several measurement sites. Predicted values have been obtained by evaluating the station's environment with the aid of the profiling feature.

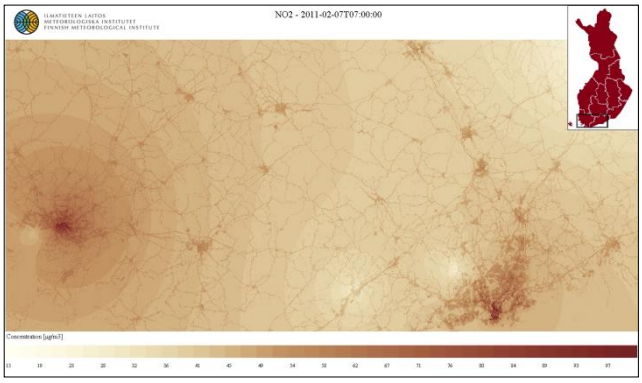

Figure 6: Fused NO_2 concentration in Southern Finland in 2011 at 07:00.

The highest concentration can be found at the centre of Helsinki, which resides in the bottom-right corner of the figure. The remote test area is a small city centre (Lohja), located approximately 70 kilometres to the right of Helsinki – 50 kilometres away from the nearest measurement station. The fused values were compared against the on-site measurements in the test area and results are shown in Figure 7.

The comparison between fused and measured NO_2 concentration at the test site (Figure 7) shows that the pollutant concentration has been estimated fairly accurately with the presented method.

During the study period the mean absolute error between predicted and measured NO_2 hourly concentration was of the order of 7 $\mu g/m_3$ (mean = 12µg, Var = 107 $\mu_2 g_2$). This error is significantly less than the achieved mean error when a conventional geographical extrapolation method would be used:

using inverse distance weighting (IWD), [11] the resulting mean absolute error would be 14 µg/m_3.

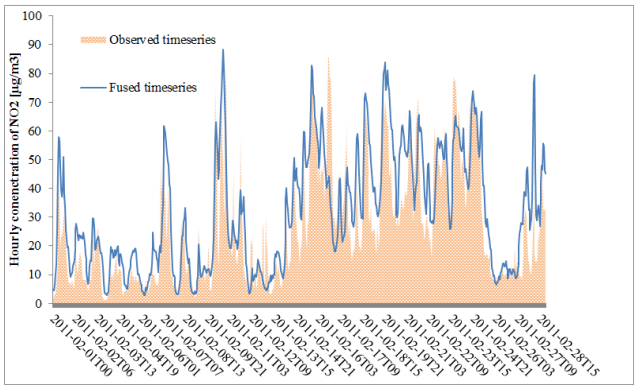

Figure 7: The observed and predicted NO_2 concentration during February 2011 at the test site, the centre of Lohja city.

Figure 8 illustrates a collection of mean absolute prediction errors from calculations similar to the one presented in Fig 7. One by one, the measurement stations were removed from the input data and the removed time series was compared against the fused time series which was produced using the remaining data. According to Fig 8 if the locations for near-by measurements represents similar environment than the location for IWD extrapolation (Laune station, Tikkurila station of Fig 8), then the IWD extrapolation may be able to predict the hourly concentration fairly well. Otherwise, the IWD-method without bias correction capabilities produces generally poor estimates in terms of mean absolute error whereas the fusion service performs well regardless of the collection of estimators used as input. Indeed, Luukki station, a rural NO_2 background measurement station is an example of this; there are several urban measurement stations nearby and thus the hourly concentration of NO_2 in Luukki cannot be extrapolated with conventional methods.

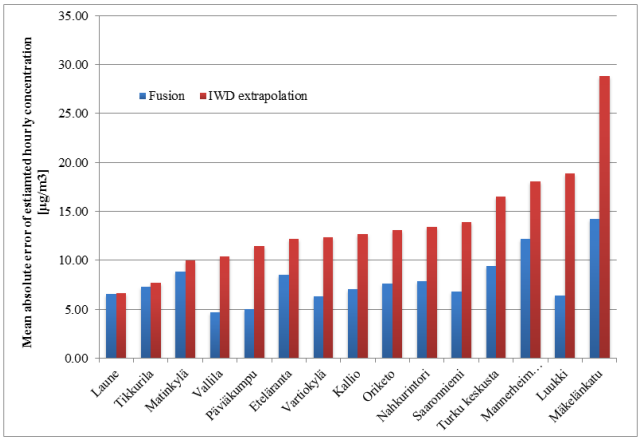

Figure 8: Comparison of IWD extrapolation and the presented fusion method in terms of standard deviation. Observed average describes the average hourly NO_2 concentration at measurement site.

5. CONCLUSION

To provide timely meteorological and air quality related information to citizens and administrative user alike, a prototype service PESCaDO was developed. By combining the data discovery, extraction and fusion methods, described in this paper, it possible to produce accurate and personalized information to the users. Unlike several search engines, the user is not confused by the sheer amount of presented data and suggestions; instead, the user is provided with a single, understandable yet precise answer. This is also what separates PESCaDO from a conventional, generic search engine. The self-maintaining design of PESCaDO system facilitates the discovery and indexing of new information sources. The source provider's performance can be evaluated and stored on a continuous basis and the stored performance data can be used to guide the fusion of information. Furthermore, the measured air quality and meteorological data that flows through the system can be used in the calibration of the fusion service's various statistical models effectively allowing the system to adapt into different regions.

The fusion method offers several advantages for the PESCaDO system. For instance, it is not necessary to discard any extracted information as the algorithm takes care that the irrelevant input is not over-emphasized. In this paper, a demonstration of the fusion of temperature forecasts was given. It was shown that the fused temperature forecast in fact had the lowest margin of error, which goes to show the benefits to be had in the fusion of information even if the amount of service providers is small.

It was shown that the presented profiling feature of the fusion service is able to predict hourly concentrations of NO_2 in different environments quite well. As a consequence, the fusion method was able to outperform a conventional extrapolation method (IWD). However, NO_2 is strongly affected by urbanization and road traffic and thus is an ideal phenomenon to be handled with the proposed fusion method. Other pollutants however, such as ozone and carbon monoxide are more difficult to handle with the presented profiling feature. In fact, the static environment based bias-removal needs to be more dynamic in the future. This could be achieved by introducing meteorology in the fusion process. For instance, the profile could be analysed from the wind's direction. Furthermore, the expected concentration could be a function of several meteorological parameters such as rain, sky conditions and wind speed. As a result, the PESCaDO system would be orchestrated in another new level, where the extracted meteorological data would be subject to fusion and used again in the fusion of air quality pollutants.

6. ACKNOWLEDGMENTS

This work was supported by the European Commission under the contract FP7-ICT-248594 (PESCaDO).

7. REFERENCES

[1] Balk, T., Kukkonen J., Karatzas, K., Bassoukos, A., and Epitropou, V., European Open Access Chemical Weather Forecasting Portal, Atmospheric Environment, 38(45), 6917–6922, 2011.

[2] Bassoukos A., Karatzas K., Kelemis A. (2005) Environmental Information portals, services, and retrieval systems, Proceedings of of "Informatics for Environmental Protection- Networking Environmental Information"-19th

International EnviroInfo Cenference, Brno, Czech Republic, pp. 151-155.

[3] Epitropou, V., Karatzas, K., Bassoukos, A., Kukkonen, J. and Balk, T., A new environmental image processing method for chemical weather forecasts in Europe, Proceedings of the 5th International Symposium on Information Technologies in Environmental Engineering. Poznan: Springer Series: Environmental Science and Engineering, 781–791, 2011.

[4] Hoek, G., Beelen, R., Hoogh, K., Viennau, D., Gulliver, J., Fischer, P. and Birggs, D. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment 42 (2008) 7561–7578, doi:10.1016/j.atmosenv.2008.05.057. 2008.

[5] Janssen, S., Gerwin, D., Fierens, F. and Mensink, C. Spatial interpolation of air pollution measurements using CORINE land cover data. Atmospheric Environment,Volume 42, Issue 20, June 2008, Pages 4884–4903, 2008.

[6] Karatzas K. A fuzzy logic approach in Urban Air Quality Management and Information Systems (UAQMIS), Proceedings of the 4th International Conference on Urban Air Quality Measurement, Modelling and Management (R. Sokhi and J. Brexhler eds), Charles University, Prague, Czech Republic, 25-27 March 2003, pp. 274-276, 2003

[7] Karatzas K. Informing the public about atmospheric quality: air pollution and pollen, Allergo Journal 18, Issue 3/09, pp 212-217, 2009

[8] Karatzas, K. and Kukkonen, J., COST Action ES0602: Quality of life information services towards a sustainable society for the atmospheric environment, ISBN: 978-960-6706-20-2, Thessaloniki: Sofia Publishers, 2009.

[9] Klein Th., Kukkonen J., Dahl Å., Bossioli E., Baklanov A., Fahre Vik A., Agnew P., Karatzas, K., and Sofiev, M., Interactions of physical, chemical and biological weather calling for an integrated assessment, forecasting and communication of air quality, AMBIO,, 41(8), pp. 851-864, 2012

[10] Kukkonen, J., Olsson, T., Schultz, D.M., Baklanov, A., Klein, T., Miranda, A. I., Monteiro, A., Hirtl, M., Tarvainen, V., Boy, M., Peuch, V.-H., Poupkou, A., Kioutsioukis, I., Finardi, S., Sofiev, M., Sokhi, R., Lehtinen, K. E. J., Karatzas, K., San José, R., Astitha, M., Kallos, G., Schaap, M., Reimer, E., Jakobs, H., and Eben, K., A review of operational, regional-scale, chemical weather forecasting models in Europe, Atmos. Chem. Phys (12), 1-87, doi:10.5194/acp-12-1-2012, 2012.

[11] Li, J. and Heap, A.D., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists. Geoscience Australia, Record 2008/23, 137 pp, ISBN 978 1 921498 30 5.

[12] Moumtzidou, A., Vrochidis, S., Tonelli, S., Kompatsiaris, I., & Pianta, E. (2012). Discovery of Environmental Nodes in the Web", Proceedings of the 5th IRF Conference, Vienna, Austria, 2012.

[13] Moßgraber, J., Rospocher, M. Ontology Management in a Service-oriented Architecture. Architecture of a Knowledge Base Access Service. Proceedings of the 23rd International Workshop on Database and Expert Systems Applications. 2012.

[14] Oyama, S., Kokubo, T., Ishida, T.: Domain-Specific Web Search with Keyword Spices Awareness in Urban Areas. J. IEEE Transactions on Knowledge and Data Engineering. 16 (1), 17—24, 2004

[15] Pianta, E., & Tonelli, S. KX: A Flexible System for Keyphrase Extraction. Proceedings of SemEval, 2010.

[16] Potempski, S. and Galmarini, S., Est modus in rebus: analytical properties of multi-model ensembles, Atmos. Chem. Phys., 9, 9471–9489, doi:10.5194/acp-9-9471-2009,2009,

[17] Sikora, T. The MPEG-7 visual standard for content description-an overview. IEEE Transactions on Circuits and Systems for Video Technology, 11(6), pp. 696-702, 2001

[18] Tang, T. T., Hawking, D., Craswell, N., & Sankaranarayana, R. S. Focused crawling in depression portal search: A feasibility study. Proceedings of the 9th Australasian Document Computing Symposium, Melbourne, Australia, 2004.

[19] Wald, L., Some terms of reference in data fusion, IEEE Transactions on Geosciences and Remote Sensing 37(3), pp. 1190-1193, 2001.

[20] Wanner, L., Vrochidis, S., Tonelli, S., Mossgraber, J., Bosch, H., Karppinen, A., Myllynen, M., Rospocher, M., Bouayad-Agha, N., Bügel, U., Casamayo,r G., Ertl, T., Kompatsiaris, I., Koskentalo, T., Mille, S., Moumtzidou, A., Pianta, E., Saggion, H., Serafini, L., and Tarvainen, V,. Building an Environmental Information System for Personalized Content Delivery. In (Hrebícek J., Schimak G., Denzer R. eds.): Environmental Software Systems. Frameworks of eEnvironment - 9th IFIP WG 5.11 International Symposium, Proceedings. IFIP Publications 359, Springer, ISBN 978-3-642-22284-9, pp. 169-176, 2011.

[21] Wanner L., Vrochidis S., Rospocher M., Moßgraber J., Bosch H., Karppinen A., Myllynen M., Tonelli S., Bouayad-Agha N., Casamayor G., Ertl Th., Hilbring D., Johansson L., Karatzas K., Kompatsiaris I., Koskentalo T., Mille S., Moumtzidou A., Pianta E., Serafini L. and Tarvainen V. Personalized Environmental Service Orchestration for Quality Life Improvement, 8th IFIP WG 12.5 International Conference, AIAI 2012 Workshops, IFIP AICT 382 (L. Iliadis et al., eds), Proceedings, Springer, pp.351-360., 2012

[22] Weigel, A.P, Liniger, M.A. and Appenzeller, C. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?. QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY. Q. J. R. Meteorol. Soc. 134: 241–260, 2008

[23] Wöber, K. Domain Specific Search Engines, In: Fesenmaier, D. R., Werthner, H., Wöber, K. (eds.) Travel Destination Recommendation Systems: Behavioral Foundations and Applications, 205—226. Cambridge, MA: CAB International, 2006.