LSPEnv: Location-based Service Provider for Environmental Data

Katarzyna Wac and Lemonia Ragia

Abstract. This paper presents an approach for forecasting environmental data for location based services. The environment becomes a very important issue and especially people with health problems need more information and support in their daily life. In this work we propose a system for making predictions for spatial-temporal variables using the Bayesian Network method as a machine learning. To handle the missing values in our data we use the Structural Expectation Maximization Algorithm. The architecture of our system is based on a three-tier architecture which assists the distribution of the evaluation process. The case study is based on real environmental data from the Swiss national network. The provided data represent different types of location, e.g. rural, urban, etc. and are taken in different time. The results can be presented on a mobile device, in Internet and to any mobile user.

Keywords: location based services, machine learning, environmental data, prediction

1 Introduction

Location based services known as LBS are developed the last two decades because of new technology. They are widely used for advertising via cell phones, for sending information when somebody visits a new place, etc. Especially with the usage of GPS the position of a person is known and new challenges are open. LBS has proven to be a new growth business. Location-based services blend information about a person's location with other useful content, providing *relevant, timely and local information* to consumers *when* and *where* they need it.

There are a lot of efforts from vendors and governmental institutions to provide such services, to improve the existing systems or to use efficiently the technology. Lots of products appeared in the market. One can use the services of a product after installing it or after payment of a fee. For example, navigation software connected to mobile phones makes navigation easier for mobile people. GIS companies try to share their information via cell phones and they spend a lot of money to improve their software, e.g. using new technology for visualization.

High bandwidth in the mobile nets (UMTS, GPRS, HSCSD), personal digital assistants (PDA), wireless connection via WLAN and the accurate position location via GPS afford new opportunities for location based services. Analysts expect that there will be a huge impact on this kind of business and the services will become more easily available for every user. A very critical factor in mobile application is security. It gives assurance to the users to use the mobile application.

In this paper we present an approach for forecasting environmental data for LBS. It is widely accepted that there is a link between the state of the environment

and human's health condition. Statistics show that many diseases can be caused by environmental pollutants [3].

Here is the description of two scenarios where people with health problems caused by the particular state of the environment can be helped in their daily life using our system.

Scenario 1: Asthma is a chronic disease in which a person experiences breathing problems. The occurrence of this disease is very much influenced by the air quality. Major factors influencing air pollution are related to transportation: gasoline and fuel fumes from cars, trains and planes. Industry also produces massive air pollutants. Smog is one of the results of this situation. Particularly, smog is produced by the existence of nitrogen dioxide in the air and occurs especially in big cities in the high traffic hours. Another factor influencing asthma is related to the house heating systems based on burning of fossil fuels. Concerning the state of the environment in a city, there are studies arguing the strong association of environmental causes of asthma and the health of inner city children [4]

Asthmatic people are interested to travel and visit new places. However, due to their condition, it would be of high importance for them to know in advance the forecasts for best and worst hours in a city center of the visited city, or generally to know which places in the city would be comfortable for them to visit in which hours and days. With such information at hand, asthmatics could be empowered to enhance their quality of experience and avoid health risks, by adapting their trip schedule in advance to their health condition and state of the environment.

Scenario 2: Allergies are chronic diseases in which a person reacts in a sensitive way to some substances in the air, food or water, having no comparable effect on the average individual. For air-related allergies, causes for such illness can be products of particular plants or other substances in the air. For example, the existence of pollen in an area with vegetation or some kind of dust at home, or air pollutants in the atmosphere can create allergic reactions. What is worse, some substances like tree pollen can be dispersed kilometres from the original area where the trees are located.

In general, allergic persons have to relocate themselves and avoid particular places in which the surrounding environment stimulates their disease. These people require (beforehand) detailed information about the surroundings environment in order to better schedule their trip and better manage their disease. Typically this information would be provided to them before they relocate, and it would contain a pollen forecast for a specific season in a given region or pollution and smog level in the city they want to visit.

In order to support these scenarios we propose in this paper the development of the Location-based Service Provider for Environmental data, denoted further on as LSPEnv. The LSPEnv provides forecasts for environment state in a given location and time over the wireless or wired access networks, i.e. to mobile or fixed service users. The information provided by the LSPEnv can be used further by people suffering from chronic diseases like asthma or allergies, to better manage the health risk taken while visiting unknown places.

2 Previous work

Location based services are "information services accessible with mobile devices through the mobile network and utilizing the ability to make use of the location of the mobile device" [5]. Similar definition is given also according to the international OpenGeospatial Consortium [6]. They are position specific information where the current position is given by a GPS. Is is also a communication means between different people when they share information. The messages are provided to the users by maps, or in textual form.

The domain of location based services is defined by the GSM Alliance Service Working Group [1] as: Asset management, Fleet management, Emergency Services, Person Tracking, Localized Advertising, Mobile Yellow Pages, Network Planning, Dynamic Network Control, Traffic Congestion Reporting, Routing to nearest enterprise, Roadside Assistance, Navigation, City Sightseeing.

Some examples are finding the next hospital or the next exit in a highway, or via SMS the most important tourist attraction in a city. Local authorities use LBS often for tourism purposes. In the beginning without GPS people could have information about a tourist attraction only when they are in a place which is covered by network based tracking. Now the tourists can inform others about interesting places or specific monument with their exact position. In traffic management it plays an important role to avoid accidents or traffic jam. For example, one driver can publish a geo-referenced message that informs about a street with heavy traffic. The same is for medical emergency information or for weather forecasting. In some cases, e.g. in the mountains, the skiers can get information for the current status of weather condition or special information about a ski region like number of mountains railways, ski rental etc. Regarding the location based services for environment there are only some concepts about web services which are related to spatial information exchange [1].

In location specific health information some examples are local disease rates including maps and guidelines, local health news, local weather, pollen and air quality alerts and maps (e.g. for allergic people), local health risks and hazards, addresses or local healthcare facilities, travelers' health information, local drugs, etc. Similar questions that have to do with the physical situation of a patient can be answered in a system where medical sensors are also connected wireless [19]. Finally the concept of location has different meanings depending on the application domain. Therefore it is impossible to cover all the different application with a unique representation.

Machine learning techniques have been used for environmental data where data mining approach has been developed for the analysis and mapping of spatially distributed data [16]. They are also used for air quality assessment which is estimated as a classification problem of real time air pollutants data [17]. In another paper three algorithms of machine learning are used to predict the daily peak concentration of an air pollutant for air quality control [18].

3 Architecture Framework

3.1 Requirements

To concentrate the operational functions that are needed for effective support in making decisions we have to define some restrictions that have to be taken into account. We consider that not only professional users but all people can be potential users of our system. Based on the given scenarios we consider the following *functional requirements* for the LSPEnv:

a. Environmental state information gathering. This requirement has to do with the available data, which can be provided by different sources and in different time frames. This category includes following characteristics:
   - From environmental sensors
   - In different locations
   - Continuously in time, but at least 1 sample per hour

b. Environmental state information processing. The modelling of the data is a key issue for the further analysis of the data. It involves:
   - Transformation of sensor data into a standard format as used by LSPEnv
   - Location attribute processing along the GPS coordination
   - Date attribute processing, derivation of day of week
   - Variables processing and assignment of ranges corresponding to levels of influence of a given environmental state variable on a person's health.

c. Environmental state information forecast. The main aims of this requirement are:
   - Prediction engine running continuously over the collected data
   - Possibility to forecast for now or any given time in a future
   - Possibility of forecast estimation for locations, times for which the data does not exist
   - Possibility of forecast variables, for which the data does not exist, based on the existing data; however that would require incorporation of specialized knowledge by the LSPEnv

d. Environmental state forecasts dissemination. The user graphical interface has to be considered in order to provide clear and understandable information. The results can be:
   - web-based user interface where user inserts a query criteria: location(s), time (s), variables he/she is interested in
   - Available for use on the mobile phone.

The users in our scenario are mainly interested in using the LSPEnv service disseminating the forecasts (function 4 above). Nevertheless, the other three functions (1-3) serve as a basis for the dissemination function. Particularly, the forecast function is a core function of the LSPEnv service provider, and this function has the following functional requirements:

a. Incremental learning upon historical data
b. Forecasting upon incomplete or uncertain historical data
c. Forecasting upon repeating/conflicting historical data

Moreover, the forecast function has the following non-functional requirements:

a. Scalability with number of data items
b. Performance in terms of prediction speed and accuracy
c. Minimal service usage cost (in case if forecasting service is used by mobile users)
d. Ensuring user's anonymity

3.2 Design

The architecture of the LSPEnv is build upon the functional requirements for LSPEnv and hence is as proposed in figure 1. The environmental state information gathering function is responsible for acquiring the sensors readings that act as an input for a environmental state information processing function, which transforms the data for prediction engine in the environmental state information forecast function. Finally, the forecast function uses the dissemination function for user's query processing and predictions provisioning to them accessing this service from a fixed or mobile Internet nodes.

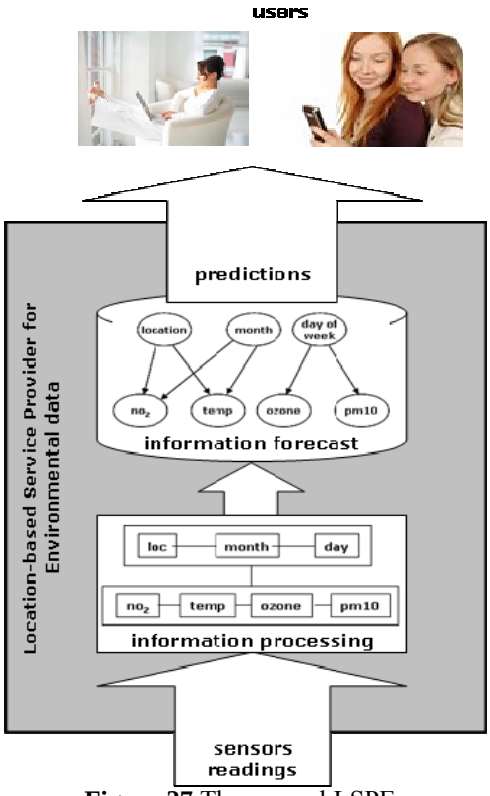


Figure 27 The general LSPEnv architecture

We propose a high-level architecture for the LSPEnv distributed system, based on a prototype application, e.g. for mobile users. The proposed architecture is very generic, since we do not want to limit ourselves to a given set of environmental data, or particular user-dependent technologies. The architecture is based on a widely-accepted distributed three-tier architecture, comprised of a client-, middle-, and resource-tier, that supports a distribution of computational tasks and separation of their concerns. Depending on the environment in which it the LSPEnv may operate, a number of specific requirements may exist, some of which may have a direct impact on the design of the LSPEnv architecture.

A *client-tier* may be composed of any kind of application; it can be

application for end-users or application acting on behalf of end-users to retrieve the necessary information from the LSPEnv. The example of the latter one can be a user-agent application or any application that requires environmental information for processing purposes to be able to fine-tune the parameters of related processes e.g. car traffic system regulation in the city. Applications can run on standard PCs or on mobile phones and Personal Digital Assistants (PDAs). The LSPEnv system may have a number of heterogeneous clients, thus, it is crucial that environmental information is accessible over one or several standardized interfaces.

The *middle-tier*, sometimes also called 'business-tier' or 'enterprise-tier', encapsulates all the logic of the LSPEnv system, and particularly the information forecast engine. This tier holds all the logic of the LSPEnv system, it acts as a layer between the client-tier that requests for environmental forecasts and the resource tier that holds the raw data. The forecast engine is accessed over a standardized interface, such as a web service. For example, when a client application requests for information on a particular environmental variable, the engine fetches the required metadata from the metadata repository, sends queries to the data sources specified for the variable, calculates the forecasted value and send the results back to the client.

The forecast engine handles all the tasks related to environmental variables forecast. The metadata repository holds all the data and metadata necessary for the forecast engine to operate. Particularly, it stores metadata such as environmental variables specifications. A rule engine allows handling rules associated to the forecast process. Moreover, the rule engine is a component that can monitor critical events and variables values, deliver alerts to users, and initiate other system actions if the LSPEnv is a part of a bigger system. For example, if the value of a particular environmental variable deviates from a predefined threshold, the rule engine may notify interested users automatically via e-mail or SMS, or may trigger a particular system action. The metadata repository and rule engine, when implemented, can be very specific for a given user/application.

Finally, the *resource-tier*, also called 'database-tier' or 'back-end', is composed of any system capable of providing environmental data to forecast engine. This data can be acquired directly from sensory readings, out of a system processing environmental data, out of (internal or external) historical databases or via invocation of specific remote procedures (Remote Method Invocation or Wireless Services) on a given environmental data source to obtain an environmental specific data.

4 The algorithm from ML

Due to the given functional requirements posed on the forecast function of the LSPEnv, we have chosen the Bayesian Networks (BN) as a machine learning method [7]. Bayesian Networks are useful models in representing and learning complex stochastic relationships between interacting variables and their probabilistic nature is capable of modeling the noise and handle missing values, as inherent in the environmental data. Moreover, that method allows for combining domain knowledge and historical data.

A Bayesian Network is a Directed Acyclic Graph (DAG) that consists of a) the structure or the directed edges that encode the causal relations and conditional independencies between the (mutually exclusive, collectively exhaustive) variables, b) the local parameters or the distribution function and parameters that encode the distribution of a child variable given its parents (CPDs). Bayesian Networks can include continuous and discrete variables [8]. Due to the nature of our problem, we have focused only on the discrete variables, with values in a finite value set. Therefore, for a given example parent B and child A, we denote their relation as:

$$P(A/B) = [P(B/A) * P(A)] / (P(B))$$

Where:

$P(A)$ – the prior probability of A

$P(B/A)$ – the conditional probability of B given A; also called the likelihood function

$P(B)$ – the marginal probability of B

$P(A/B)$ – the posterior probability of A given B

In our learning task, firstly we have focused on the inference of the graph structure from the data. A significant challenge to this task poses the fact that our data has missing values (at random). To tackle BN learning with missing values we use the Structural Expectation Maximization (SEM) Algorithm [9]. SEM searches in the joint space of (graph structure x parameters). It starts with a random structure and its parameters and estimates the probability distribution of missing variables with the SEM algorithm. Then it computes the expected score for each graph of the neighborhood and chooses the one which maximizes the score. It is presented in the following pseudo-code:

Loop for n=0,… until convergence
  compute the posterior P(A/B)
    E-step: for each $graph_n$, derive missing values for variables,
      based on knowledge on their distribution, then compute an expected score for this graph (via sum of log of its prior P(A) and log of its marginal probability P(B/A))
    M-step:choose neighborhood $graph_{n+1}$ that, for given values of variables maximizes the score of the $graph_n$
    if expected score of $graph_n$ == expected score of $graph_{n+1}$
  return $graph_n$

After several updates to the graph structure, we run SEM again to recalculate the missing values for the final graph. This final graph, with the assigned missing values is then used for forecasting service.

The used Bayesian Network has the following structure derived from data based on the SEM algorithm (1) implemented in Matlab Bayes Net Toolkit [10]

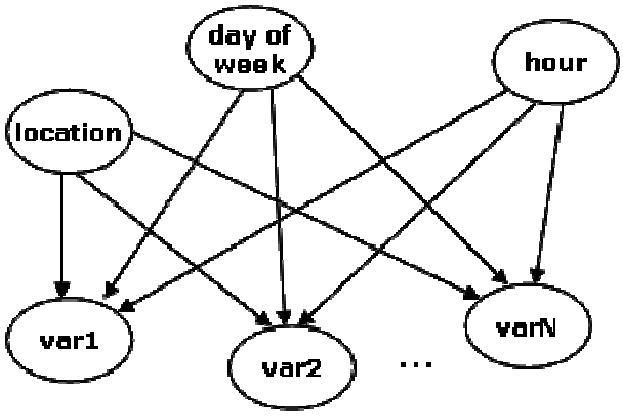


Fig. 2. Bayesian Network Graph Structure

5 Evaluation

The data are provided by the Swiss national network [12] which is representative for the whole country. The measurement stations represents the following locations:

a) urban with very heavy traffic,
b) urban areas with a population more than 100000 people
c) suburban areas
d) rural next to the highway
e) rural with an altitude higher than 1000 m a.s.l.
f) rural with an altitude lower than 1000 m a.s.l. and
g) areas in the high mountains

We have obtained data for 16 different locations in different time frames over Switzerland for temperature, ozone (O_3), nitrogen dioxide (NO_2) and particulate matter (PM10) as environmental variables. Among the air pollutants we choose some of the most important: the ozone, the nitrogen dioxide and the particulate matter because these are high associated with the human health [16]. The ozone is undoubtedly one of the basic causes for many health problems, it has short and long term effects on human health and plays an important role in the mortality rate [13]. Scientific results show that the nitrogen dioxide is a significant factor for the increase of a lot of significant allergic illnesses [15]. The particulate matter is a mixture of organic and inorganic substances and it causes the air pollution. There is scientific work that show the high responsibility of the particulate matter for health effects [14].

For ozone, the limit value is 120 ($\mu g/m^3$) as a daily maximum 8-hour mean, above which a person health may get affected. For nitrogen dioxide this limit is 80 $\mu g/m^3$ 1-hour mean, while for particulate matter its is 50 $\mu g/m^3$. Temperature does

not have critical levels of values in the range which have been measured in Switzerland.

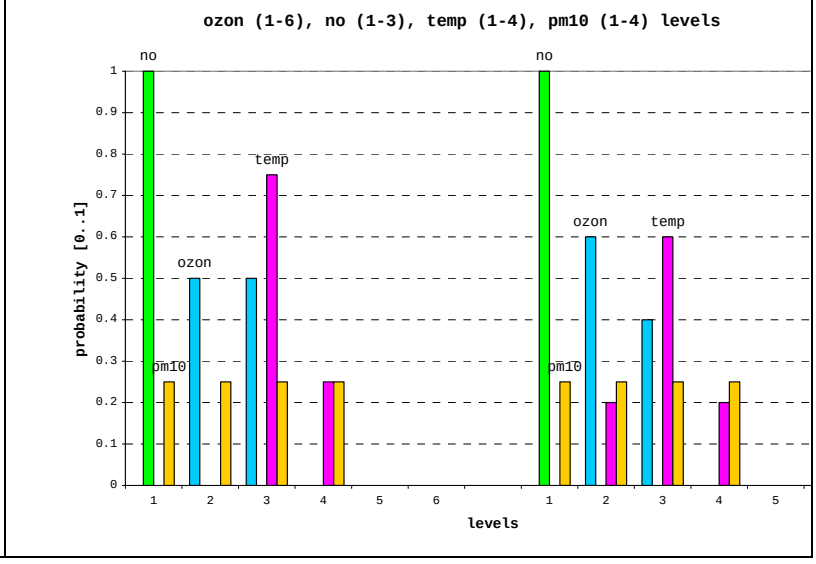
Tab. 1. Values and levels of different environmental state variables

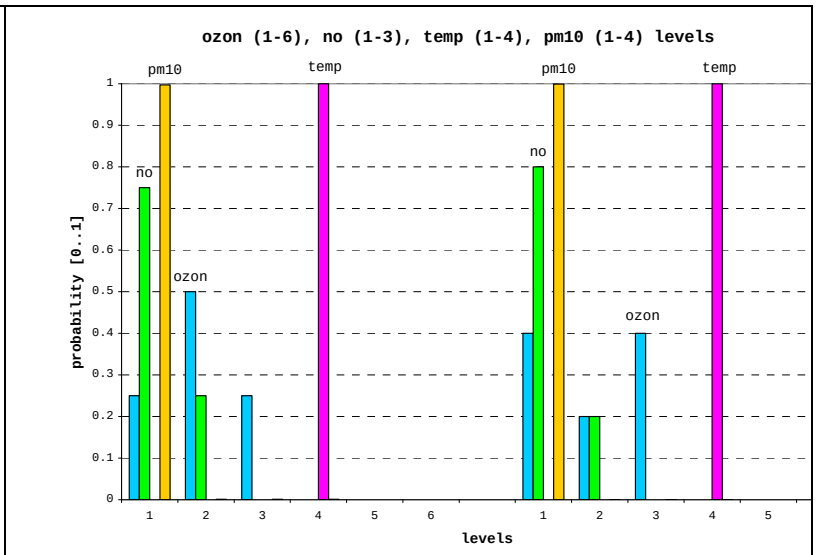
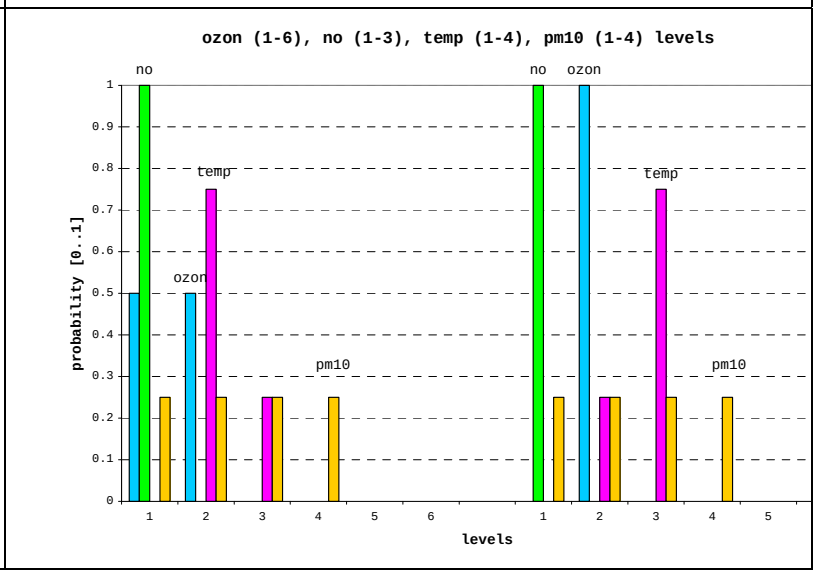
	min			levels of values						max
level	**1**		**2**	**3**		**4**	**5**			**6**
ozon	2	70		100	120		160	240		270
No_2	1		40		80		200			
temp	1		7		12		17		27	
pm10	1		50		75		100		116	

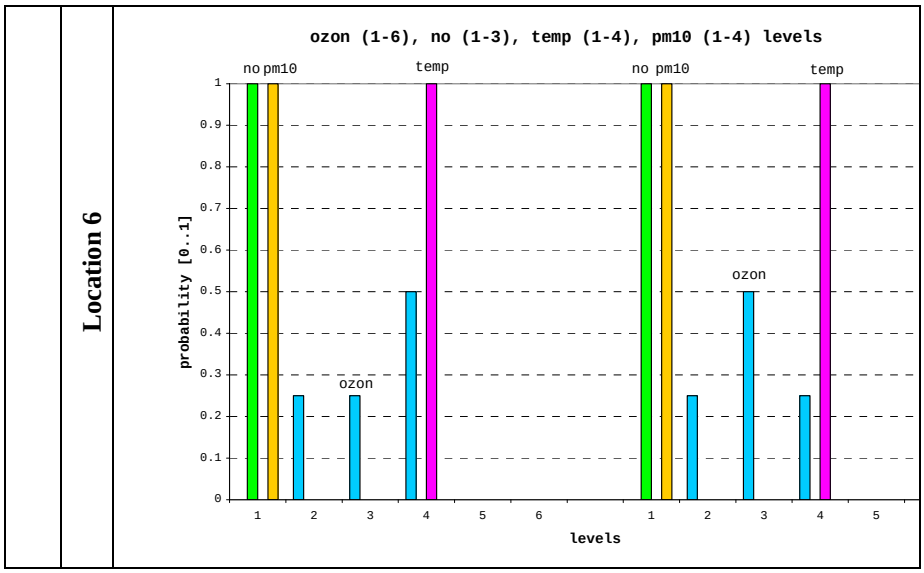
5.1 Evaluation scenario

For the collected data we have evaluated scenario in which an asthmatic user sends the query to the system regarding the environmental state variables in two different cities in Switzerland (5 and 6) for a weekend (Saturday-Sunday) in summer months of July and August. The user plans a weekend trip to one of these cities, and may make a decision based on prediction for environmental state. Prediction results are presented in table 2.

Tab. 2. Probability distribution for different levels of environmental state variables for two different cities for a summer weekend.

		Saturday	Sunday
July	**Location 5**		

	Location 6	
August	**Location 5**	



As we could see from the prediction, the location 6 has higher temperature and level of ozone and nitrogen dioxide (only in July) higher than location 5, hence the person may make a decision upon visiting location 5 rather than 6.

5.2 A prototype LSPEnv application – technical aspects

In order to be able to evaluate the LSPEnv system, at least on a small scale, we propose a prototype application which builds upon the proposed system architecture (Figure 2). The prototype application builds upon a simplified version of our architecture which features only its core elements for the storage of data for and prediction of limited number of environmental variables. Other elements, as those related to metadata repository or enhanced rule engine, have been discarded for the sake of simplicity.

From a technical point of view, our prototype application is based on the architecture proposal presented in previous section (Sec. 3.2). It is composed of a client application that runs on a mobile phone, a environmental variables prediction engine, and a simple database system which holds the raw data for measurements. The use cases for the prototype system are as follows:

1. store set of historical data for set of environmental variables,
2. list available set of environmental variables,
3. provide predictions for given location, month and day of week.

Our prototype application can be based on a distributed, three-tiered architecture, as presented in Figure 3. The client-tier component can run on a mobile phone. The forecast engine of the middle-tier can run on a Java EE platform and can use a database system as metadata repository. Finally, the resource-tier is represented by a database system. A rule engine can be part of our prototype

application, but it not yet implemented. Figure 3 shows the overall architecture of the prototype application.

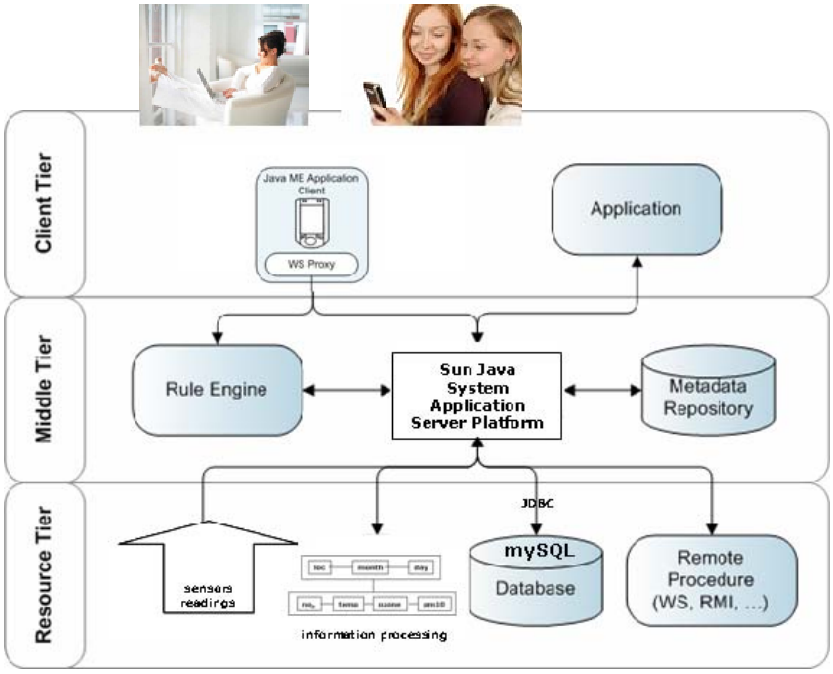


Fig. 3. The implemented LSPEnv system architecture

The application client runs on a Java Platform, Micro Edition (Java ME) which provides an environment for applications running on mobile devices. It allows to display a prediction results to a mobile user, where the screen can be updated automatically every few seconds by a timer task or upon the notification of the prediction change triggered from the LSPEnv system. Furthermore, the application client should allow to list all available environmental variables, and to display detailed information on each one of them on a mobile device. The example prediction for environmental variables can be displayed to the mobile user in a simple graph form, as shown for an example query in Table 2.

Since nearly all information displayed by the application client is provided by the middle-tier, an application may requests data whenever needed. For this purpose, the client can call the web service interface of the middle-tier component. Web services are use open, XML-based standards and transport protocols to exchange data with clients. The Simple Object Access Protocol (SOAP) can be used to exchange data. SOAP defines the envelope structure, encoding rules, and conventions for representing web service invocations and responses. The requests and responses are transmitted over the Hypertext Transfer Protocol (HTTP). The WS Proxy, which represents the remote web service on the application client, is

generated by the Java API for XML Web Services (JAX-WS), based on the Web Services Description Language (WSDL) file of the web service. Whenever the application client requires data, it simply invokes the methods on the WS Proxy.

The LSPEnv prediction engine can run on a Sun Java System Application Server Platform, a compliant implementation of the Java EE 5 platform. The prediction algorithms is implemented in Matlab, as indicated in previous sections, and can be called upon the client's prediction request. The metadata repository could be implemented in a simple file-system holding specifications of environmental variables.

Finally, the resource-tier can be represented by a standard mySQL database which stores all the 'raw' data. It is on this database that the prediction engine executes the queries that are specified by the variables stored in the metadata repository. The connection between the prediction engine and the database is established through the Java Database Connectivity (JDBC), which provides methods for querying and updating data in a database.

6 Conclusions and discussion

We propose an open system for forecasting environmenatal issues by integrating different spatial-temporal data and representing them in a graphical user interface. We presented an approach for location based services that takes into account the air quality monitoring and supports people with health problems. The approach is based on the Bayesian Networks method as the prediction function and the Structural Expectation Maximization Algorithm is included to avoid problems with missing values in raw data. We use the three-tier architecture, composed of a client-, middle-, and resource-tier.

We have used real environmental data with four variables from Switzerland which represent lot of different types of the locations. The results are the predictions for these variables for a specific date, or month in the future. The results can be shown in Internet via wireless connection or in a PDA in simple graph form.

We are investigating more functionalities for data retrieval and the visualization of the results. We try to use methods which combine textual and map information for better understanding of the results.

References

[1] GSM The GSM Alliance Services Working Group http://www.gsmworld.com

[2] Ragia L., El Isbihani A., Kiehle C., 2006: Web Service for Groundwater Vulnerability. International Conference, Protection and Restoration of the Environment, July 3-7, Chania, Greece.

[3] European Commission, 2002: "Health statistics - Key data on health 2002- Data 1970 - 2001", ISBN 92-894-3730-8

[4] Mortimer K. M., Tager I. B., Dockery D. W., Neas L. M., and Redline S. 2000: The Effect of Ozone on Inner-City Children with Asthma. In American Journal of Respiratory and Critical Care Medicine, Vol. 162, No. 5, pp. 1838-1845.
[5] Virrantaus K., Markkula J., Garmash A., Terziyan Y. V., 2001 : Developing GIS-Supported Location-Based Services. In: Proc. of WSIS'2001-First International Workshop on Web Geographical Information Systems, Kyoto, Japan, pp. 423-432.
[6] Open Geospatial Consortium (OGS), 2005. Open Location Services.
[7] Charniak, E. (1991). Bayesian networks without tears. *AI Magazine,* 1991. **12**(4): p. 50-63.
[8] Heckerman, D. (1995). A Tutorial on Learning with Bayesian Networks
[9] [Friedman, N. (1998) The Bayesian structural EM algorithm, in G. F. Cooper & S. Moral, eds., Proc. *14th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA.
[10] Murphy, K. (2001). The Bayes Net Toolbox for Matlab. *Computing Science and Statistics*, **33**(1)
[11] Murphy, K. (2001). The Bayes Net Toolbox for Matlab. *Computing Science and Statistics*, **33**(1))
[12] Federal Office for the Environment FOEN, Department of the Environment, Transport, Energy and Communication http://www.bafu.admin.ch/index.html?lang=en
[13] [European Commission (2002): Health statistics Key data on health 2002.
[14] [Dingenen R. V. et al. (2004): European Aerosol Phenomenology - 1 : physical characteristics of particulate matter at kerbside, urban, rural and background sites in Europe in *Atmospheric Environment* Vol. 38, Issue 16, pp. 2561-2577
[15] Jenkins H. S. et al. (1999) The Effect of Exposure to Ozone and Nitrogen Dioxide on the Airway Response of Atomic Asthmatics to Inhaled Allergen. In *American Journal of Respiratory and Critical Care Medicine*, Vol. 160, No. 1, pp. 33-39.
[16] Kanevski M. et al. (2004): Environmental data mining and modeling based on machine learning algorithms and geostatistics. In Environmental Modelling and Software, Vol. 19, Issue 9, pp. 845-855.
[17] Athanasiadis I. N. et al. (2003) Applying Machine Learning Techniques on Air Quality Data for Real-Time Decision Support. In First International Symposium on Information Technologies in Environmental Engineering, Gdansk, Poland, ICSC-NAISO Publishers.
[18] Kalapanidas E. and Avouris N. (1999) Applying Machine Learning Techniques in Air Quality Prediction. In Proc. Advanced Course on Artificial Intelligence, ACAI ´99, Chania, pp.58-64.
[19] Van Halteren A. et al. (2004) Mobile Patient Monitoring: The MobiHealth System. The Journal on Information Technology in Healthcare Vol. 2, Issue 5, pp. 365-373.