2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing

Exploring Urban Dynamics based on Pervasive Sensing: Correlation Analysis of Traffic Density and Air Quality

Wenzhu Zhang*, Bing Zhu*, Lin Zhang*, Jian Yuan* and Ilsun You†
*Department of Electronic Engineering, Tsinghua University
Email: {zhwz, bingzhazha, linzhang, jyuan}@tsinghua.edu.cn
†School of Information Science, Korean Bible University
Email: isyou@bible.ac.kr

***Abstract*—Modern cities, with large population and complicated infrastructures, are complex entities with non-linear and dynamic properties that challenge the city management. Therefore, as the first step towards the goal of thorough understanding of the phenomena, pervasive urban sensing have become a cornerstone of future smart city that enhance the interplay between the cyber space and the physical world. We introduce a taxi-based pervasive urban sensing system and its key algorithm, aiming at the quantitative study of the correlation between human activities and environmental changes. Our contributions are twofold. First, we propose an urban crowd-sourcing framework that take automobiles as participatory mobile agents to the sensing tasks, and implemented a prototype in Beijing. Second, we design a Spatial-Temporal Manifold Learning (STML) algorithm to analyse the correlation between physical processes. Based on noisy and partially labelled dataset that are collected by pervasive urban sensor networks, we evaluate STML's performance by analysing correlation between the traffic density and the air quality. The results show great potential of STML for future urban sensing applications.**

I. Introduction

The nowadays information system is enabled by advanced sensor technologies with sophisticated capability to achieve information of surrounding environments. In everyday life, we are essentially surrounded by a great variety of sensing devices, from the CCD camera in the mobile phones to the collision sensors in our cars. Crosss-domain data-mining of the sensor data that reveals hidden connections in real-time may change the paradigm of urban management by providing revolutionary improvement in data accessibility, information provision and policy-making transparency.

One of the major driving forces of pervasive sensing is the need for better, greener and smarter urban live. A city can be reasonably considered as an inherently human-driven self-organizing structure, whose dynamics is driven by the social behavior of its residents, which is known to be strongly related to the environmental changes. Exploring the correlation between the human activities and the environmental changes, as well as the resource consumption, will be beneficial in many aspects in city management including the city plan, the resource utility optimization, the convenience improvement, etc.

A great shift of methodologies occurs in urban dynamic study, from the model-driven paradigm to the data-driven paradigm. In 1960s, regional modelling was developed to present a geographical system, which account for the exchange of population, goods, capital in the area. Ecological models deal with several qualitatively different types of relationships between a small number of components, aiming at understanding the most general laws of urban dynamics. With the development of the general theory of complex systems, cellular automata models and multi-agent models became popular to describe the macro-processes resulting from collective behaviour at the micro-level of land plots and migrating city individuals[1]. Nowadays, however, "straight forward" data mining and pattern learning methods become alternatives of model-based prediction, especially when we consider a metropolitan area with complex activities and interactions among millions of people, where a deterministic model is hard to formulate.

Moreover, former researchers are usually civil engineering experts who study urban phenomena as separated physical processes, based on their own domain knowledge. Now, what is more interesting is the correlation between different physical processes. Relationship among different types of processes and events, such like the traffic, the air quality, the energy consumption, could be studied as a whole dynamic system, so as to better predict the dynamics of the urban system.

To achieve this goal, at least two aspects of technical challenges should be considered. One is how to collect meaningful multi-dimensional dataset in a city scale? And the other is how to perform efficient analysis for the oceans of data (if available) to obtain deterministic and informative knowledge and prediction model.

For address the first challenge, we propose a pervasive urban sensing framework which takes automobiles as mobile agents to perform crowd sensing on the city scale. Nowadays crowd sensing is not a new concept already, especially when smart phones equipped with GPS receivers, compass modules and accelerometers become popular so that tasks, such as location-aware micro-blogging[2], personalized estimates of environmental impact and exposure[3], road/driving con-

978-0-7695-4684-1/12 $26.00 © 2012 IEEE
DOI 10.1109/IMIS.2012.137

dition profiling[4] , and live traffic information sharing[**?**], can be performed on a participatory basis. However, in a smart-phone based participatory sensing system, issues of sensor scalability, data consistency and user privacy arise as obsticles. Therefore, an alternative solution was proposed to embody sensors to the vehicle fleets, which are more consistent, powerful and less concerned about privacy. CarTel[5] and MobEye[6] are two piloting projects.

Great challenge remains, however, as there are many characters not considered. For instance, the uncertainty of the communication channel, the sampling density and the vehicle mobility patterns will become more complex as the number of sensing devices scales. In our design, the major concern is to protect the sensing performance against the sparse distribution and the diversified mobility patterns of vehicles that challenge every aspect of design.

For the second question, we propose a novel Spatial-Temporal Manifold Learning (STML) algorithm to analyse the correlation between arbitrary pair of physical processes. STML aims to solve the essential challenges in the machine learning of urban sensing data:

1) The unavoidable noise or imprecision in training data adds uncertainty to the reconstruction process.
2) The sparsity of data obtained from crowd urban sensing causes incompleteness and heterogeneity of dataset both in space and time.
3) Quantitative analysis among different physical processes in different measurement is difficult. Semantic abstraction are required to gain meaningful information.

In this paper, we report our work progress in the pervasive urban sensing and city dynamics study. In section II, we briefly review our work foundations in three aspects with regard to three inter-related technical planes: a) pervasive urban sensing, b) opportunistic communication and networking, c) city dynamics and behaviour study. In section III, we propose our prototype that performs urban environmental monitoring and the deployment on vehicular networks. In section IV, we propose a novel machine learning framework named Spatial-Temporal Manifold Learning(STML) algorithm. In section V, we share the preliminary results on the spatial-temporal distribution of traffic and air quality, which reveals the unseen information and potential cross-domain usage of sensor data. Finally, we conclude this paper and propose several further directions in section VI.

II. Work Foundations in Urban Area Crowd Sensing and City Dynamics Study

In this section, we briefly review our efforts towards better understanding of city dynamics in the context of pervasive sensing. Basically, we can decompose our work onto three inter-related technical planes, namely, the pervasive urban sensing, the opportunistic communication and networking, and the city dynamics behaviour study. In this section, we introduce our work foundations and perspectives for each technical plane.

A. Pervasive urban sensing

Pervasive urban sensing provides the ground truth to the urban dynamics study. In this sub-section, we will discuss the data description of urban environmental monitoring, application of compressed sensing technique in vehicular sensor network and some open questions.

1) *Urban data description*: To model a pervasive urban sensing system, we firstly need to define the range of data description which is required by the further study. To be specific, we refer to the classification methodology in context-aware system, which study how to represent contexts in a computation form and how to support an operational life-cycle in using context-aware systems[7].
2) *Compressed sensing*: Another aspect of urban sensing study is on the sensing capability and the related signal processing techniques. [8] proposed a cooperative data sensing and compression approach with zero inter-sensor collaboration overhead based on sparse random projections. The spatial correlation of signals (temperature, humidity, gas emission, et al.) is found to be beneficial to compressed sensing and improve reconstruction accuracy with much smaller communication load.

B. Opportunistic Communication and Networking

Communication and networking in urban area is the enabling technology to aggregate data and upload to the fusion center or computing cloud. The main challenge of this part is about cost, which is the most influential factor when the scale (number of nodes) of network becomes larger. Opportunistic communication and networking approach has been studied as low-cost solution for the urban sensing.

1) *Urban short range communication*: [9] studied the urban environmental impacts over the performance of IEEE 802.15.4c, which is a low-bit-rate wireless communicate standard that enables global interconnectivity and inter-operability amongst Wireless Personal Area Network(WPAN) transceivers from different manufacturers. Several real urban scenarios are investigated, including car to car, car to infrastructure communication under heavy and light traffic conditions, with LOS(Line-Of-Sight) and /or NLOS(Non-Line-Of-Sight) paths. Results show the most important performance influential factor is the availability of LOS, while the speed of vehicles is not as impacting as expected.
2) *Delay Tolerant Networking (DTN)*: DTN is a typical opportunistic networking approach, which usually use moving vehicles to improve sensing coverage of the

city. [10] proposes a distributed information dissemination algorithm, namely DAWN(Density Adaptive routing With Node awareness), which enables the vehicle's awareness of its neighbour density, and the transmission behaviours are fine-tuned according to the node density to maximize the data delivery probability with the constraint of the local channel capacity and the maximum allowable delay.

C. City dynamics and behaviour study

As pervasive sensing becomes available with large volume of sensory data about the environmental changes and human activities on the city scale. The analysis of the data arises to be a challenge. There are still a lot of open questions in the urban behaviour study, the pattern learning and the knowledge discovery.

[11] proposed a context-based framework, namely, the Context-Aware Metropolitan Sensing (CAMS), to rise to the increasing challenges in the context acquisition, fidelity, dynamics and complexity. CAMS is a high level framework that focuses on knowledge discovery among distributed or mobile users by three stages: context acquisition (local data collection and context sharing), context management (filtering, composition and storage), and context utilization (discovery, adaptation and annotation).

III. Prototype and Implementation

Although many smart phones nowadays are computationally powerful, yet they cannot support environmental urban sensing due to lack of appropriate sensors, for instance, the weather sensor (temperature, humidity, etc.) or the air quality ($\mathbf{CO}, \mathbf{SO_2}, \mathbf{H_2S}$, particulate matters, etc.) because of the volume, weight and cost constraints. We design and develop a prototype that embodies specific sensors to perform environmental sensing, especially for the temperature, humidity and carbon monoxide information.

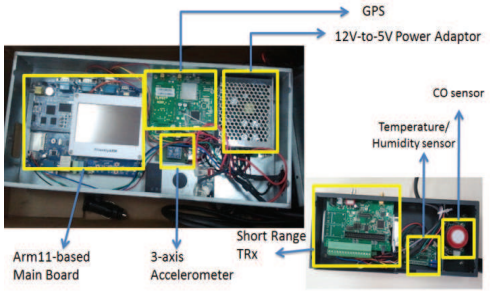

Figure 1. Pervasive Urban Sensing Prototype (Left: Sensing platform, size 15cm×40cm, Right: Environmental sensing module, size 10cm×15cm)

The prototype is designed to be equipped on vehicles, with each device including two inter-connected parts (Fig.1). One is environmental sensing module, which is installed on the roof of buses, and the sensors of temperature, humidity and carbon monoxide sensors, as well as a short range RF transceiver are integrated in it. The other is the sensing platform, which is installed in the passenger cabin, built with an ARM-11 based main board, a GPS module, cellular module, a 3-axis accelerometers, a battery and DC adaptor that converts 12V DC from the vehicle to 5V DC that is compatible with the inside modules.

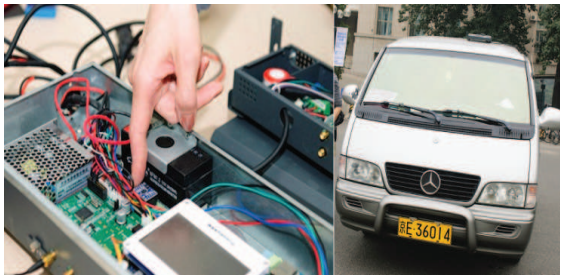

Figure 2. Deployment on Tour Buses

The processor is Samsung S3C6410 works at 533MHz, with a 256M RAM and 1G Flash. All the extended modules use serial ports to communicate with the main board. All the sensors are comodity-off-the-shelf products with factory calibration. An embedded version of Linux (OS version 2.6.36) is used as operation system on main board, with a Qt graphical system.

The sensing task is performed as a real-time application that schedules the working pattern of GPS, accelerometers and other sensors. An independent thread is kept for short range communication with close-by devices.

We have implemented this prototype on 15 tour buses in Beijing (Fig.2). The testbed has been running for three months. A rich information dataset is obtained as the first-hand information about the dynamics of the Beijing dynamics.

IV. A Manifold Learning Framework on City Dynamics Study

A. Classical Regularization Theory in Supervised Learning

Supervised learning are described as an inverse problem, in the sense that its formulation builds on knowledge obtained from examples of the corresponding direct problem, which involves underlying physical laws that are unknown[12]. To be specific, let the training set be described by

$$\text{Input signal:}\mathbf{x}_i \in R^m, i = 1, 2, \ldots N \quad (1)$$

$$\text{System response:}d_i \in R, i = 1, 2, \ldots N \quad (2)$$

The inputs are m dimensional vectors, while the output in our case is assumed to be one dimensional. The purpose of the learning process is to determine an approximation function to the unknown system, denoted by $F(\mathbf{x})$.

In most cases of engineering, inverse problems are ill-posed, according to Hadamard's definition. Tikhonov regularization theory was introduced to restrict the solution of the hypersurface reconstruction problem to compact subsets by minimizing the augmented cost function[13].

$$\Psi(F) = \Psi_s(F) + \lambda \Psi_c(F). \tag{3}$$

where, $\Psi_s(F)$ denotes the empirical cost function. In the least-square estimator case,

$$\Psi_s(F) = \frac{1}{2} \sum_{i=1}^{N} \big(d_i - F(x_i)\big)^2. \tag{4}$$

$\Psi_c(F)$ denotes the regularizer, which relies on certain geometric properties of the approximation function $F(x_i)$.

$$\Psi_c(F) = \frac{1}{2} \|\mathbf{D}F\|^2, \tag{5}$$

where $\mathbf{D}$ is a linear differential operator.

B. Generalized Regularization Theory and Manifold Learning

From a practical perspective, it is usually very hard to perform manual labelling of examples, especially in the urban sensing scenarios. In contrast, the collection of unlabelled examples is relatively inexpensive and much easier in the real-world deployment of pervasive sensing system. Given these practical realities, semi-supervised learning could be used to exploit the availability of both labelled and unlabelled examples in the learning process.

Classical regularization theory discussed above incorporates a single penalty function that reflects the ambient space, where the labelled examples are generated. Generalized regularization theory extends the classical theory by incorporating a second penalty function that reflects the intrinsic geometric structure of the input space[12]. It could be applied to the semi-supervised learning, based on labelled as well as unlabelled data.

To be specific, the input dataset $\{\mathbf{x}_i\}_{i=1}^{N}$ is divided into two subsets. One is a subset of data points denoted by $\{\mathbf{x}_i\}_{i=1}^{l}$, for which a corresponding set of labels denoted by $\{d_i\}_{i=1}^{l}$. The other subset denoted by $\{\mathbf{x}_i\}_{i=l+1}^{N}$, for which the labels are unknown.

Consider the labelled examples $(\mathbf{x}, d)$ generated in accordance to the joint distribution function $p_{\mathbf{x},D}(\mathbf{x}, d)$, and the unlabelled examples $\mathbf{x} \in X$ generated according to the marginal distribution function $p_{\mathbf{x}}(\mathbf{x})$. We suppose there is coherence between the two kind of data, i.e. if two input data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close to each other in the intrinsic geometry of the marginal distribution function $p_{\mathbf{x}}(\mathbf{x})$, then the conditional distribution function $p_{\mathbf{x}|D}(\mathbf{x} \mid d)$ evaluated at the data points $\mathbf{x} = \mathbf{x}_i$ and $\mathbf{x} = \mathbf{x}_j$ behaves similarly.

To this end, we verify the expression in Eq.(3) as

$$\Psi(F) = \Psi_s(F) + \frac{1}{2}\lambda_A \Psi_c(F) + \frac{1}{2}\lambda_I \Psi_I(F) \tag{6}$$

To find proper $\Psi_I(F)$, Belkin proposed manifold method which implies the intrinsic geometric structure of the input space[14]. Inspired by Belkin's work, we pursue the kernel approach based on manifold regularization. By manifold, we mean a k-dimensional topological space embedded in an n-dimensional Euclidean space where n is greater than k. Suppose we have a set of unlabelled examples denoted by $\mathbf{x}_1, \mathbf{x}_2, \ldots$, which are all n-dimensional. These examples can be represented as a set of data points in an n-dimensional Euclidean space. Most unsupervised-learning algorithms operate only on the ambient space, represented by the examples $\mathbf{x}_1, \mathbf{x}_2, \ldots$. Suppose, however, that we are able to construct a manifold of lower dimensionality than n, such that the true data may reside on or close to that manifold. Then it may be possible to design a more effective semi-supervised learning algorithm by exploiting the underlying geometric properties of the manifold in addition to those of the ambient space. This will provide a novel way of approaching problems of learning algorithms on manifolds that are revealed through sampled data points.

We use spectral graph theory to model a manifold[14]. Consider the training sample

$$X = \{\mathbf{x}_i\}_{i=1}^{N}, \tag{7}$$

which embodies N input data points, labelled as well as unlabelled. Given this training sample, we proceed by constructing a weighted undirected graph graph consisting of N vertices, one for each input data point, and a set of edges connecting adjacent vertices. We define any two nodes i and j are connected, provide that the Euclidean distance between their respective data points $\mathbf{x}_i$ and $\mathbf{x}_j$ is small enough to satisfy the condition

$$\|\mathbf{x}_i - \mathbf{x}_j\| < \epsilon \tag{8}$$

Let w_{ij} denote the weight of an undirected edge connecting nodes i and j. The weights in the graph as a whole are usually real numbers. Then the N-by-N weight matrix $\mathbf{W} = \{w_{ij}\}$ is a symmetric, nonnegative-definite matrix. Hereafter, we refer to the undirected graph, characterized by the weight matrix $\mathbf{W}$, as graph G. Let $\mathbf{T}$ denotes an N-by-N diagonal matrix whose ii-th element is defined by

$$t_{ii} = \sum_{j=1}^{N} w_{ij} \tag{9}$$

which is called the degree of node i. Intuitively, the larger the degree is, the more important the node i is.

We define the *Laplacian* of graph G as

$$\mathbf{L} = \mathbf{T} - \mathbf{W} \tag{10}$$

Assume that there are no self-loops, that is $w_{ii} = 0$ for all i, then for the ij-th element of the Laplacian $\mathbf{L}$, we have

$$l_{ij} = \begin{cases} t_{ii} & \text{for } i = j \\ -w_{ij} & \text{for adjacent nodes } i \text{ and } j \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Since the Laplacian $\mathbf{L}$ is a symmetric matrix, it has real eigenvalues. We adopt Rayleigh coefficient of a symmetric matrix to evaluate the variational characteristics of the eigenvalues of the Laplacian $\mathbf{L}$. To this end, let f denote an arbitrary vector-valued function of the input vector $\mathbf{x}$, which assigns a real value to each vertex of the graph G. The Rayleigh quotient of $\mathbf{L}$ is defined as below[12].

$$\lambda_{Rayleigh} = \frac{\mathbf{f}^T\mathbf{L}\mathbf{f}}{\mathbf{f}^T\mathbf{f}} \tag{12}$$

The N real-valued eigenvalues is shown by the set

$$\lambda_0 \leq \lambda_1 \leq \cdots \lambda_{N-1}$$

which is called the eigenspectrum of the Laplacian $\mathbf{L}$, or graph G.

Define the vector-valued function $\mathbf{f}$ in terms of the training sample X:

$$\mathbf{f} = [F(\mathbf{x}_1), F(\mathbf{x}_2), \ldots, F(\mathbf{x}_N)]^T \tag{13}$$

Hence, using Eqs. (11) and (13),

$$\mathbf{f}^T\mathbf{L}\mathbf{f} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\big(F(\mathbf{x}_i) - F(\mathbf{x}_j)\big)^2 \tag{14}$$

Then, we define the weight w_{ij} as a kernel function:

$$w_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \tag{15}$$

Here $k(\mathbf{x}_i, \mathbf{x}_j)$ is Gaussian kernel function,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}) \tag{16}$$

where $2\sigma^2$ is a tunable parameter which is assumed to be the same for all the kernels in the spectral graph.

Let $\Psi_I(F) = \mathbf{f}^T\mathbf{L}\mathbf{f}$ in equation 6, according to generalized representer theorem, which is proved by [14] , optimization of the cost function $\Psi(F)$ admits the form

$$F(\mathbf{x}) = \sum_{i=1}^{N} a_i k(\mathbf{x}, \mathbf{x}_i) \tag{17}$$

To compute $\mathbf{a} = [a_1, a_2, \cdots, a_N]^T$, we re-write equation (6) in matrix notations,

$$\Psi(\mathbf{a}) = \frac{1}{2}(\mathbf{d}-\mathbf{J}\mathbf{K}\mathbf{a})^T(\mathbf{d}-\mathbf{J}\mathbf{K}\mathbf{a}) + \frac{1}{2}\lambda_A\mathbf{a}^T\mathbf{K}\mathbf{a} + \frac{1}{2}\lambda_I\mathbf{a}^T\mathbf{K}\mathbf{L}\mathbf{K}\mathbf{a} \tag{18}$$

where $\mathbf{d}$ is $l - by - 1$ desired response vector: $\mathbf{d} = [d_1, d_2, \ldots, d_l]^T$, $\mathbf{J}$ is $N - by - N$ diagonal matrix, partially filled with l unity terms: $\mathbf{J} = diag[1, 1, \ldots, 1, 0, 0, \ldots, 0]$. $\mathbf{K}$ is $L - by - L$ Gram matrix: $\mathbf{K} = \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^{N}$. $\mathbf{L}$ is the Laplacian graph matrix. Differentiating this equation with respect to the vector $\mathbf{a}$, we get

$$\mathbf{a}^\star = (\mathbf{J}\mathbf{K} + \lambda_A\mathbf{I} + \lambda_I\mathbf{L}\mathbf{K})^{-1}\mathbf{J}^T\mathbf{d} \tag{19}$$

C. Spatial-Temporal Manifold Learning(STML) Framework

In this section, we merge two promising methods– manifold learning and spatial temporal correlation analysis, to solve the problem of correlation study of two indirectly related physical process, such as traffic density and air quality in urban area.

Obviously, we have to discretize the space and time at first. Suppose there are N adjacent but not intersect areas $A_1, A_2, \ldots, A_N$. Each A_i, $i \in \{1, 2, \ldots, N\}$, contains m blocks $B_1, B_2, \ldots, B_m$. For every A_i, there are m measurement(sensory data) coming from m blocks respectively, denoted by $X_i(t)$. Define the learning result(output) as scalar $d_i^X(t)$. So we have m-dimensional input $X_i(t)$ and one scalar output $d_i^X(t)$ at time-step t in area A_i.

Suppose we have only partial knowledge about the input-output mapping of $X_i(t)$ and $d_i^X(t)$. To be specific, for area $A_1(t) \sim A_l(t)$, we know the system output $d_1^X(t) \sim d_l^X(t)$, for area $A_{l+1}(t) \sim A_N(t)$, we know nothing about the output. That is standard semi-supervised learning problem, as we discussed above. Manifold learning method based on spectral graph theory is proposed to solve this problem.

Moreover, if there are two inter-related process $X_i(t)$ and $Y_i(t)$, $i \in \{1, 2, \ldots, N\}$. If we want to explore the implicit relationship between them, traditional methods adopt statistical methods such as canonical correlation analysis (CCA). However, it is hard to justify the meaning or significance of study results. Here we propose a new paradigm to perform correlation analysis. Firstly, for each dataset, we adopt manifold learning methods to reduce the dimension of data and obtain more "abstract" information, which we could be interpreted as the semantic level knowledge. Then the statistical methods, such as spatial temporal analysis, could be used for both semantic learning results. This approach is illustrated by Fig. 3.

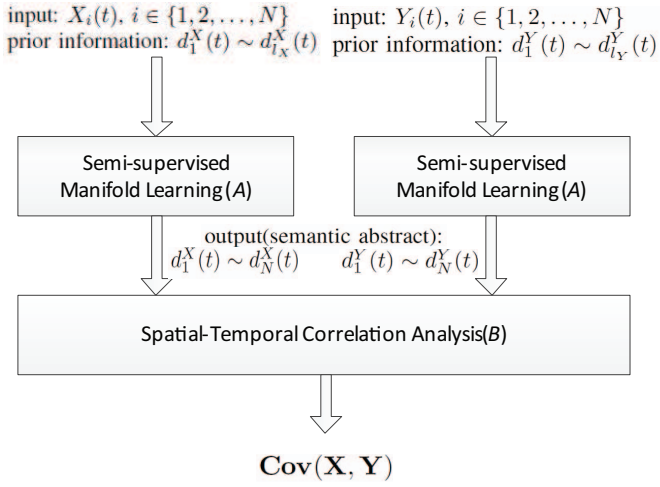

Figure 3. Spatial-Temporal Manifold Learning Framework

In the STML framework, we have two separated input

datasets: $X(t)$, $Y(t)$, each with partial prior knowledge about the output. The algorithm (A) in Fig. 3 is manifold learning method which performs semi-supervised learning and results in semantic abstract $d_X(t)$ and $d_Y(t)$. The algorithm (B) is spatial-temporal correlation algorihm that generate the correlation matrix $\mathbf{Cov}(\mathbf{X}, \mathbf{Y})$ with index i denoting area A_i and τ denoting the delay between two process $X(t)$ and $Y(t+\tau)$. To be specific, we conclude these two algorithms as follows.

Algorithm 1 Semi-supervised Manifold Learning

Input:

$\{X_i(t), d_i(t)\}_{i=1}^{l}$ and $\{X_i(t)\}_{i=l+1}^{N}$, which are respectively labeled and unlabeled.

Parameter: spectral graph parameters ϵ and σ^2, ambient regularization parameter λ_A and intrinsic regularization parameter λ_I.

Output:

$\{d_i(t)\}_{i=1}^{N}$ and approximating function $F(\mathbf{x})$.

1: Construct the weighted undirected graph G with N nodes, using:
Eq. (8) for identifying the adjacent nodes of the graph, and
Eqs. (15) and (16) for computing the edge weights.
2: Choose kernel function $k(\mathbf{x}, \cdot)$ and using the training sample, compute the Gram $\mathbf{K} = \{k(\mathbf{x_i}, \mathbf{x_j})\}_{i,j=1}^{N}$
3: Compute the Laplacian matrix $\mathbf{L}$ of the graph G, using Eqs. (9) and (11)
4: Compute the optimum coefficient vector $\mathbf{a}^\star$, using Eq. (19).
5: Use Eq. (17) to compute the optimized approximating function $F(\mathbf{X})$ and then the output $\{d_i(t)\}_{i=1}^{N}$.

Algorithm 2 Spatial-Temporal Correlation

Input:

Semantic abstraction $\{d_i^X(t)\}_{i=1}^{N}$ and $\{d_i^Y(t)\}_{i=1}^{N}$, suppose these two stochastic processes are jointly wide sense stationary.

Output:

Correlation Matrix $\mathbf{Cov}(\mathbf{X}, \mathbf{Y})\mid_{i=1,2,\ldots,N}$

1: For every area A_i, $i \in 1, 2, \ldots, N$, neglect the index i, compute

$$\rho_{XY}(\tau) = \frac{\mathbf{E}[(X(t) - \mu_X)(Y(t+\tau) - \mu_Y)]}{\sigma_X \sigma_Y}, \quad (20)$$

2: Correlation Matrix $\mathbf{Cov}(\mathbf{X}, \mathbf{Y})$ is a three dimensional matrix with $\rho_{XY}^i(\tau)$ for each index i.

V. Results Evaluation

Basically, there are two datasets that we used for analysis, within the range of 5th ring road of the Beijing city (E116.209 E116.544N39.76 N40.02). For population immigration, we use Beijing taxi dataset with involves totally more than 20,000 taxi trajectories in one month. The distribution and fluctuation of taxis could reveal people's moving pattern. For the environmental change, we used the dataset from our prototype system, which is a very sparse sampling result.

Fig.4 shows the density of vehicles in every 4 hours, where each small cell denotes 1km×1km area. We can see from Fig.4 that in the city center (inside the 3rd ring road), the population density is usually higher than that of other places, with the west regions higher than the east regions in the center. For temporal analysis, we can see two explicit peaks in 12am and 8pm, which reflects the most active status of people.

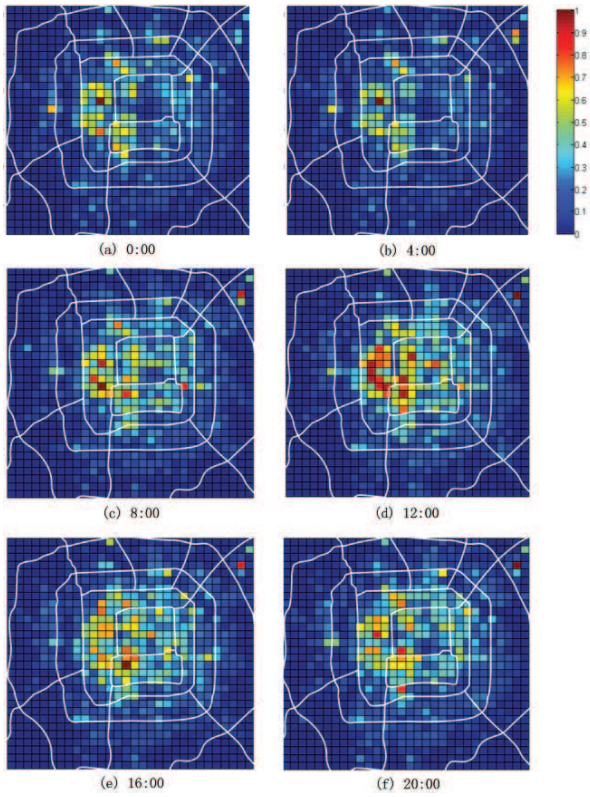

Figure 4. Traffic Density in Beijing in 24 hours

Fig.5 shows the distribution and fluctuation of carbon monoxide, which is an important index of air quality. In Fig.5 we can see an obvious hot zone, which indicates severe air pollution in that region. We find that there are several chemical factories in the south of Beijing city, which are reasonably responsible for the local air pollution.

The correlation between different indexes is of much value, if we could reveal the implicit relationship among different phenomena. We uniformly divide the urban area into 100 sub-area $A_1, A_2, \ldots, A_{100}$, each area $(3km \times 3km)$

owns 9 blocks.

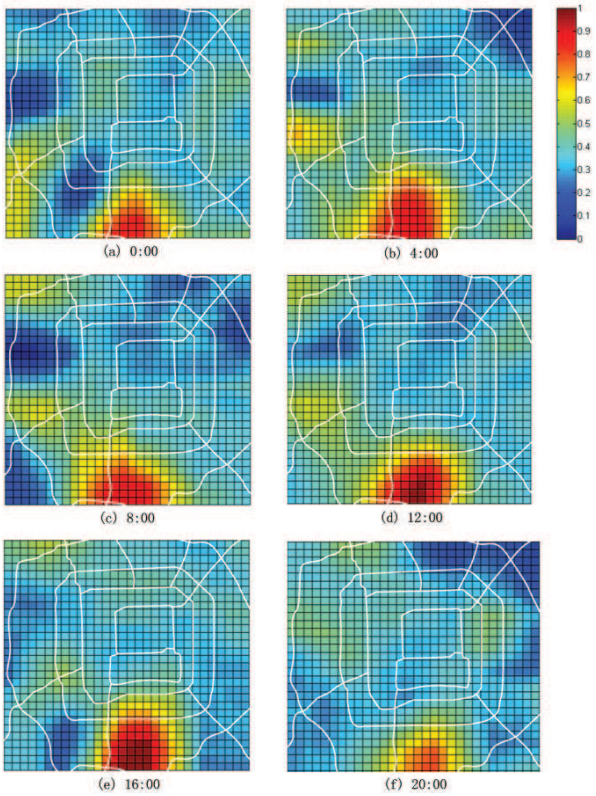

Figure 5. Carbon monoxide dynamics in Beijing in 24 hours

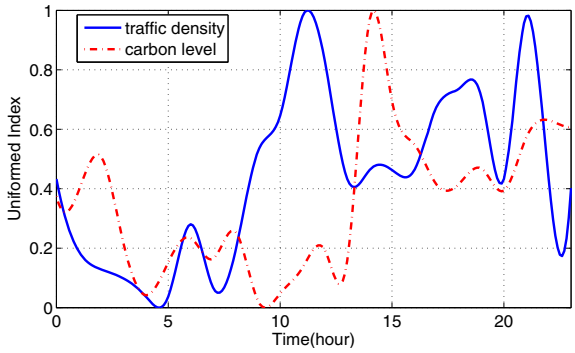

Figure 6. A Semantic Abstract of Traffic and Carbon in 24 hours

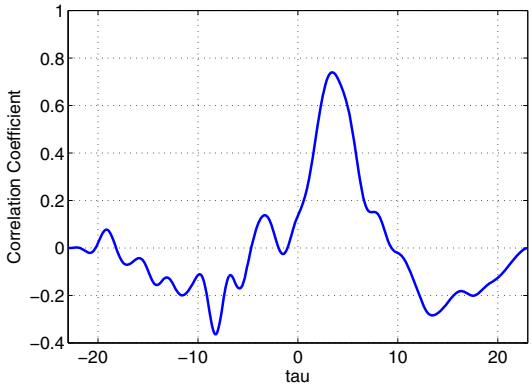

Figure 7. Correlation of Traffic Density and Air quality at Selected Area

In Fig.6 we show the learning results at selected area (Dong Tie Ying Bridge, a 9 km^2 region with center E116.43, N39.856). The blue real line denotes traffic density, while red dotted line denotes air quality. It is inferred that the air quality is probably influenced by population density. Fig. 7 shows the correlation between the two phenomena. A strong positive peak is monitored at $\tau = 3.4$, with correlation coefficient 0.74. The correlation peak means a delayed dependence of air quality to traffic density is reasonably justified by our methods. For the selected area, we can predict with confidence the air pollution peak will occur about three hours later after the rush hour.

VI. Conclusion

In this paper, we report our work progress on urban dynamics study. For crowd sensing in urban area, different aspects of technical improvements are discussed, such as compressed sensing and manifold learning, urban channel test, delay tolerant networking, context-aware framework, prototype and deployment, et al. The major contribution of this paper is Spatial-Temporal Manifold Learning(STML) algorithm, which is a novel framework to study the correlation of different urban physical processes. On one hand, STML reveals the intrinsic structure of dataset by spectral graph theory to achieve dimension reduction, while using regularization theory to perform noisy small dataset learning. On the other hand, STML makes it possible for the spatial-temporal correlation analysis of two urban phenomena relying not on the raw data, but the learning results (semantic information). The effectiveness of STML is justified by a case study of correlation analysis between the traffic density and the air quality. Also, other interesting applications and non-trivial results will be emerging to get better understanding of our cities.

Acknowledgment

This work is partially funded by the Robert Bosch Stiftung Urban Sensing of City Dynamics and Energy Use in the Beijing Metropolitan Area.

References

[1] I. Benenson, "Modeling population dynamics in the city: from a regional to a multi-agent approach," *Discrete Dynamics in Nature and Society*, vol. 3, pp. 149–170, 1999.

[2] S. Gaonkar, J. Li, R. R. Choudhury, L. Cox, and A. Schmidt, "Micro-blog: sharing and querying content through mobile phones and social participation," in *Proceedings of the 6th*

international conference on Mobile systems, applications, and services, ser. MobiSys '08. New York, NY, USA: ACM, 2008, pp. 174–186.

[3] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda, "Peir, the personal environmental impact report, as a platform for participatory sensing systems research," in *Proceedings of the 7th international conference on Mobile systems, applications, and services*, ser. MobiSys '09. New York, NY, USA: ACM, 2009, pp. 55–68.

[4] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: using mobile smartphones for rich monitoring of road and traffic conditions," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*, ser. SenSys '08. New York, NY, USA: ACM, 2008, pp. 357–358.

[5] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, and S. Madden, "Cartel: a distributed mobile sensor computing system," in *Proceedings of the 4th international conference on Embedded networked sensor systems*, ser. SenSys '06. New York, NY, USA: ACM, 2006, pp. 125–138.

[6] U. Lee, E. Magistretti, M. Gerla, P. Bellavista, and A. Corradi, "Dissemination and harvesting of urban data using vehicular sensing platforms," *Vehicular Technology, IEEE Transactions on*, vol. 58, no. 2, pp. 882 –901, feb. 2009.

[7] S. Poslad, *Ubiquitous Computing: Smart Devices, Environments and Interactions*. London: John Wiley and Sons Ltd, 2009.

[8] X. Yu, H. Zhao, L. Zhang, S. Wu, B. Krishnamachari, and V. O. K. Li, "Cooperative sensing and compression in vehicular sensor networks for urban monitoring," in *Communications (ICC), 2010 IEEE International Conference on*, may 2010, pp. 1 –5.

[9] B. Xia, Q. Fu, D. Li, and L. Zhang, "Performance evaluation and channel modeling of ieee 802.15.4c in urban scenarios," in *Communications (APCC), 2010 16th Asia-Pacific Conference on*, 31 2010-nov. 3 2010, pp. 497 –502.

[10] Q. Fu, W. Feng, Y. Zheng, and L. Zhang, "Dawn: A density adaptive routing algorithm for vehicular delay tolerant sensor networks," in *Forty-Ninth Annual Allerton Conference on Communication, Control, and Computing*, 2011.

[11] W. Zhang, L. Zhang, Y. Ding, T. Miyaki, D. Gordon, and M. Beigl, "Mobile sensing in metropolitan area: Case study in beijing," in *Mobile Sensing Workshop in 13th International Conference on Ubiquitous Computing (UbiComp'11)*, 2011.

[12] S. Haykin, *Neural Networks and Learning Machines, Third Edition*. Prentice Hall, 2009.

[13] A. N. Tikhonov, *Solutions of Ill-Posed Problems*. New York: Winston, 1977.

[14] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, December 2006.