An Extended Spatio-Temporal Granger Causality Model for Air Quality Estimation with Heterogeneous Urban Big Data

Julie Yixuan Zhu, Chenxi Sun, *Student Member, IEEE*, and Victor O.K. Li, *Fellow, IEEE*

Abstract—This paper deals with city-wide air quality estimation with limited air quality monitoring stations which are geographically sparse. Since air pollution is influenced by urban dynamics (e.g., meteorology and traffic) which are available throughout the city, we can infer the air quality in regions without monitoring stations based on such spatial-temporal (ST) heterogeneous urban big data. However, big data-enabled estimation poses three challenges. The first challenge is data diversity, i.e., there are many different categories of urban data, some of which may be useless for the estimation. To overcome this, we extend Granger causality to the ST space to analyze all the causality relations in a consistent manner. The second challenge is the computational complexity due to processing the massive volume of data. To overcome this, we introduce the non-causality test to rule out urban dynamics that do not "Granger" cause air pollution, and the region of influence (ROI), which enables us to only analyze data with the highest causality levels. The third challenge is to adapt our grid-based algorithm to non-grid-based applications. By developing a flexible grid-based estimation algorithm, we can decrease the inaccuracies due to grid-based algorithm while maintaining computation efficiency.

Index Terms—Granger causality, spatio-temporal (ST), heterogeneous, big data, air quality estimation

1 Introduction

1.1 Motivation

Air quality has deteriorated rapidly in China and Hong Kong, with NO2 and PM2.5 levels frequently exceeding WHO safety guidelines. While poor air quality has clear public health impacts, very few monitoring stations (e.g., only 15 monitoring stations in Hong Kong and 36 in Beijing) are available for measurements of major air pollutants, severely limiting evidence-based air quality decision-making, and leading to severe criticisms about the transparency and public relevance of the official Air Quality Index (AQI), or Air Quality Health Index (AQHI) in Hong Kong. Since air pollution is highly location-dependent, and monitoring stations are costly and bulky, a city-wide air quality monitoring system would be prohibitively expensive. Spatio-temporal (ST) heterogeneous urban big data may help fill this gap. By analyzing the temporal dependency and spatial correlation between urban dynamics data, such as meteorology and traffic, one can estimate air quality at locations not covered by monitoring stations [1].

ST heterogeneous urban big data refer to the datasets containing spatial, temporal, and category information $(s,\ t,\ c)$ which reflect the status of various urban dynamics. The basic assumption is that air quality is considerably influenced by these urban dynamics (e.g., wind, vehicular traffic, and point of interest (POI)). As shown in Fig. 1, air quality at time t_k and at a location without an AQI station could be influenced by urban dynamics of the surrounding locations at the previous timestamp t_{k-1} and the current timestamp t_k. However, these influences are tangled and complicated, some of them are useless, noisy, or spatially and temporally redundant. Thus it is necessary to find the most influential data for air quality estimations in terms of category c, space s, and time t.

To deal with the complexity, Zhu et al. [23] proposed a Granger causality [2] model which integrates different urban dynamics $(s,\ t,\ c)$ in a consistent manner and selects the most influential data for air quality estimation. The model is illustrated using data from Shenzhen, China. In this paper we extend [23] in two ways. First, we show the model can be migrated to other cities, using Hong Kong as an example. Second, we discuss the inaccuracies caused by the grid-based estimation [23] and further adapt it to non-grid-based scenarios by developing a flexible grid-based estimation algorithm.

The goal of this paper is to solve three challenges regarding city-wide air quality estimation with ST heterogeneous big data, i.e., unifying the data diversity, improving time efficiency, and dealing with inaccuracies caused by grid sizes. We first try to analyze the causalities between urban dynamics and the air quality. Here causalities are expressed based on Granger causality [2], which represents the causality between two time series in a regression manner and determines a "Granger" cause if one time series can successfully predict another. To analyze all the causalities among urban dynamics in a consistent manner, we extend the Granger causality model to the ST space. We also rule out the urban dynamics that do not "Granger" cause air pollution by implementing non-causality test. As for the time complexity,

• *The authors are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Pok Fu Lam, Hong Kong. E-mail: {yxzhu, cxsun, vli}@eee.hku.hk.*

Manuscript received 10 Aug. 2015; revised 28 Oct. 2016; accepted 25 Dec. 2016. Date of publication 15 Jan. 2017; date of current version 7 Sept. 2017. Recommended for acceptance by Q. Yang.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TBDATA.2017.2651898

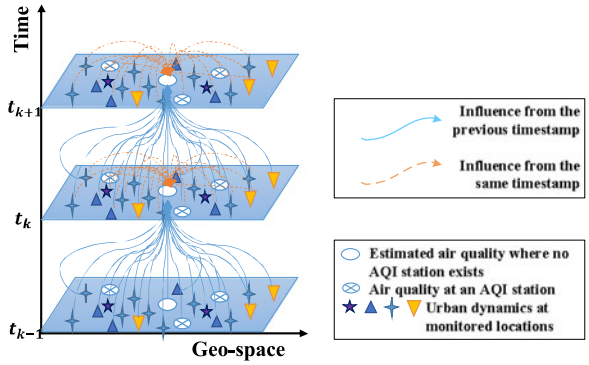


Fig. 1. Influences of spatio-temporal urban dynamics on air quality.

we propose an approach to discover the region of influence (ROI) by selecting data with the highest causality levels spatially and temporally. Results show that we achieve higher accuracy using "part" of the data than "all" of the data, supporting the approach for urban big data computing which just processes the most influential data. Finally, to overcome the inaccuracies caused by the grid size, or the spatial granularity of analysis, we further propose an algorithm enabling flexible grid-based air quality estimation. The causality model and the city-wide air quality map are verified and visualized using data from Shenzhen (a city in China, Jan 2013–May 2015) and Hong Kong (Jan 2014–May 2015).

1.2 Related Work

To estimate fine-grained, city-wide air quality with limited monitoring stations, there is already much literature from the following four research fields.

1.2.1 Spatial Interpolation

Spatial interpolation methods were first proposed by environmental and public health studies [10], [11] to examine the relations between ambient air quality and respiratory health effect. Approaches falling under this category, such as spatial averaging, nearest neighbor, IDW, and kriging, mainly base their assumptions on the spatial continuity of air pollutants distribution. Wong et al. [12] evaluated the performance of commonly used interpolation methods, and [13] showed IDW approach performs better than kriging in terms of errors. However, there are two limitations: 1) different interpolation approaches produced fairly different estimations, 2) the accuracy was not satisfactory since the original datasets are geographically sparse.

1.2.2 Ancillary Sensor Networks

A more trustworthy way to handle city-wide air quality monitoring is based on ancillary sensor networks (e.g., remote sensing, vehicular sensor networks (VSN) and crowdsourcing). Gupta et al. [14] estimated the global PM2.5 quality by means of aerosol optical thickness (AOT) retrieval from satellite imaging data. Yu et al. [15] proposed an urban environmental surveillance regime based on VSN, in which cooperative sampling is implemented by compressive sensing. Usually, the above-mentioned approaches require infrastructure deployment with relatively high cost. As for crowdsourcing, there is emerging interest focused on this topic. Li et al. [16] targeted at monitoring gas pollutants and particle matters with portable sensors and smartphones. However, the major challenge is in the sensing inaccuracy, since current portable sensors cannot achieve comparable precision to monitoring stations. Furthermore, for the scenario of sensing PM2.5, the device needs more than 1hr for data collection and averaging. Models are needed to reduce the error caused by mobility.

1.2.3 Emission Models

One type of approach belonging to this category relates to the emission factor. It estimates overall pollutant levels by detailed division of source sectors, such as fuel combustion, agriculture, transportation, etc. Various models and standards are proposed regarding each sector, elaborated by the guidebooks from US and Europe environmental agencies government [17], [18]. For example, the Motor Vehicle Emission Simulator (MOVES) model [19] acts as the new generation emission model for the mobile sector in the US, based on different driving modes. Another type of approach is based on the dispersion model [20], which predicts the pollutants concentration by characterizing how the meteorological processes disperse a pollutant emitted by a source. However, these models usually require many assumptions which may be difficult to satisfy and parameter settings which may be difficult to obtain.

1.2.4 Urban Computing

A new philosophy to handle this problem is based on urban computing [1], [28]. By analyzing the correlations and patterns from urban big data, one may infer and discover unknown knowledge. Zheng et al. [21] inferred the fine-grained noise pollution by using multiple data sources from the urban area. Shang et al. [22] estimated the gas consumption and pollutants emission based on GPS trajectory data. The causality-based approach proposed in this paper also belongs to this category. Ryan and LeMasters [25] reviewed the land use regression model which predicts air quality based on surrounding land use and traffic. Zheng et al. [26] proposed an urban data-driven method to forecast the air pollution in the next 48 hour.

1.3 Contributions and Outline

The merits of our work are three-fold. First, we propose a Granger causality model in the ST space to represent causalities among urban dynamics in a consistent manner. Two causality measurement factors are defined, for two urban dynamics and for multiple urban dynamics. The first factor is used for non-causality test, to rule out irrelevant categories of urban dynamics. The second factor can be visualized geographically on the map when quantified to ten causality levels, allowing us to observe how the urban dynamics influence air quality, rather than just getting the output (AQI estimations) from training the input (urban dynamics) without knowing what happens in the process. Second, we select the most relevant data spatially and temporally for AQI estimation based on causality analysis and ROI detection, thus enabling the reduction of "big data" into "small data". Third, we adapt our grid-based algorithm to non-grid-based applications. By developing a flexible grid-based estimation algorithm, we can decrease the inaccuracies due to the grid-based algorithm while maintaining computation efficiency.

TABLE 1
Basic Information of Shenzhen Urban Dynamics Data

Domain	Category	Urban dynamics	Tuple format	No. of tuples	Source
Air pollution	1	AQI	(VALUE, S,T)	133315	SLEN & SEMC
	2	PM2.5	(VALUE, S,T)	133315	
	3	PM10	(VALUE, S,T)	133315	
	4	NO2	(VALUE, S,T)	133315	
	5	CO	(VALUE, S,T)	133315	
	6	O3	(VALUE, S,T)	133315	
	7	SO2	(VALUE, S,T)	133315	
Meteorology	8	Pressure	(VALUE, S,T)	36796207	SZMB
	9	Humidity	(VALUE, S,T)	36796207	
	10	Temperature	(VALUE, S,T)	36796207	
	11	1 hour rain	(VALUE, S,T)	36796207	
	12	24 hour rain	(VALUE, S,T)	36796207	
	13	Wind	(VALUE, S,T)	36796207	
Traffic	14	Traffic speed	(VALUE, S,T)	7892434	TCSM
	15	Traffic index	(VALUE, S,T)	7892434	
Geography	16	POI	(VALUE, S)	153225	BAIDU Map
	17	Urban morphology	(VALUE, S)	500000	
	18	Roadmap	(VALUE, S)	500000	SECL

Note - Yellow: not completely open data, Green: open data

Shenzhen has 11 public AQI stations. We have additional air quality data from 3 non-public stations provided by Shenzhen Environmental Monitoring Center (SEMC) under non-disclosure agreement for research purposes. Data sources from public websites: Shenzhen Living Environment Network (SLEN), Shenzhen Environmental Monitoring Center (SEMC), Shenzhen Meteorological Bureau (SZMB), Transport Commission of Shenzhen Municipality (TCSM), Shenzhen Expressway Company Limited (SECL).

TABLE 2
Basic Information of Hong Kong Urban Dynamics Data

Domain	Category	Urban dynamics	Tuple format	No. of tuples	Source
Air pollution	1	AQHI	(VALUE, S,T)	262800	EPD HK
	2	PM2.5	(VALUE, S,T)	262800	
	3	PM10	(VALUE, S,T)	262800	
	4	NO2	(VALUE, S,T)	262800	
	5	O3	(VALUE, S,T)	262800	
	6	SO2	(VALUE, S,T)	262800	
Meteorology	7	Pressure	(VALUE, S,T)	7533600	HK Observatory
	8	Humidity	(VALUE, S,T)	7533600	
	9	Temperature	(VALUE, S,T)	7533600	
	10	Max wind speed	(VALUE, S,T)	7533600	
	11	Min wind speed	(VALUE, S,T)	7533600	
	12	Wind direction	(VALUE, S,T)	7533600	
Traffic	13	Traffic speed	(VALUE, S,T)	15992640	HK TD
	14	Road saturation	(VALUE, S,T)	15992640	
Geography	15	Urban morphology	(VALUE, S)	264000	BAIDU Map & HK TD
	16	Roadmap	(VALUE, S)	264000	

Green: open data

Hong Kong has 15 air quality monitoring stations, with all data available on the website of Environmental Protection Department (EPD). The meteorology data are collected from the HK Observatory website. Real-time traffic data and roadmap data are provided by the HK Transportation Department (TD). The urban morphology data are collected from Baidu Map.

The rest of this paper is organized as follows. Section 2 discusses how the causality among urban dynamics are defined. Section 3 analyzes the causality between urban dynamics and air quality, visualized using urban data from Shenzhen and Hong Kong. Section 4 describes the city-wide and fine-grained air quality estimation. Section 5 evaluates the estimation performance. Section 6 deals with the inaccuracies caused by grid sizes. Section 7 concludes with suggestions on future work.

2 How is Causality in the ST Space Defined?

2.1 Urban Dynamics

To detect and measure how urban dynamics cause air pollution, we need to represent ST heterogeneous urban dynamics in an understandable way. Tables 1 and 2 list 18 and 16 categories of urban dynamics for Shenzhen and Hong Kong respectively, divided into 4 domains: air pollution, meteorology, traffic and geography. We define the ST process for each urban dynamic as $X(s;t;c_i)$, where $s \in$ spatial domain, $t \in$ temporal domain and c_i is the ith urban dynamic category. Most urban dynamics (category 1-15 in Table 1) come in the form of time series at different locations throughout a city, obtained by crawling publicly accessible websites with data updated per hour. While for other dynamics (category 16-18 in Table 1) which just contain spatial information, we assume the values remain constant. Thus these static urban dynamics can be viewed as time invariant geographical snapshots at specific timestamps.

2.2 Causality Models

To know why and what will happen, "cause and effect" has long been discussed from three perspectives, i.e., the unit-level causality [3], the graphical causality [4], [5], and the predictive causality [2] (For more detailed discussion about causality models, please refer to Appendix A). The methodology in this article is based on the Granger causality [2] from the predictive causality perspective, where causality is loosely defined as one time series successfully predicting another, with its mathematical representation as:

$$Y_t = \Sigma_{k=1}^{L} a_k Y_{t-k} + \Sigma_{k=1}^{L} b_k X_{t-k} + \xi_t. \quad (1)$$

Here ξ_t are uncorrelated random variables with zero mean and variance σ^2, L is the number of timestamps, vectors $\boldsymbol{a} = \{a_k\}$ and $\boldsymbol{b} = \{b_k\}$ $(k = 1, 2, \ldots, L)$ are the correspondent weights for two processes X_t and Y_t. The null hypothesis that X_t does not cause Y_t is supported when $\boldsymbol{b} = 0$, reducing (1) to:

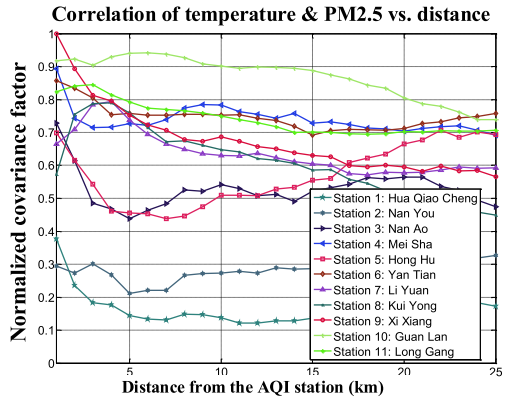
$$Y_t = \Sigma_{k=1}^{L} a_k Y_{t-k} + \tilde{\xi}_t. \quad (2)$$

This means process Y_t is caused by its own history, not by process X_t. For another extreme case where $\boldsymbol{a} = 0$ and $\boldsymbol{b} \neq 0$, process Y_t is only caused by process X_t, not by its own history.

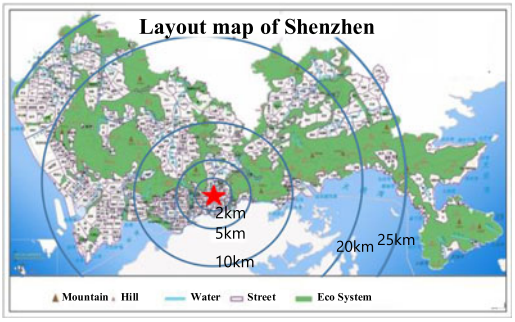
2.3 Extend Granger Causality to the ST Space

2.3.1 Unifying Different Urban Dynamics

For ST heterogeneous data, the "variety" matters with causality analysis. Fig. 2 shows the correlation of PM2.5 and temperature as a function of distance from different AQI stations. The temperature and PM2.5 are spatially correlated. But distance is another factor: PM2.5 in the current location may be less likely to be influenced by far-away environment (see Fig. 2a). For PM2.5 at a specific AQI station, e.g., Station 5 Hong Hu, the correlation decreases first with the distance, and then increases. The reason for this is explained by Fig. 2b. We observe that the urban morphology within the distance of 20-25 km from Station 5 tends to be similar to the region within 0-5 km, which covers more city area. Thus we need to represent the temporal and spatial causalities for two urban dynamics and also for multiple urban dynamics.



(a) Correlation of temperature and PM2.5 vs distance



(b) Urban morphology of Shenzhen

Fig. 2. An illustration to show the correlation of two urban dynamics may depend on multiple factors.

Another issue related to unification is the normalization of different urban dynamics, which come in different units and distributions. Simply normalizing the mean value of each time series to 1 will lead to computational bias, since those dynamics with low variance will hover just close to 1 while other dynamics with high variance may reach great values. Thus we use the normalized 1-hour difference (the 1-hour difference values divided by the standard deviation) as the input process. Observations show all the urban dynamics time series of 1-hour difference (category 1-15 in Table 1) obey the Gaussian distribution. Fig. 3 shows three representative histograms for the 1-hour difference of urban dynamics (left histogram) and the original distribution (right histogram), from the domain of air quality, meteorology, and traffic.

For the time series data in Shenzhen (air quality, meteorology, and traffic data in Table 1), we collect the data from publicly available websites, from 2013-01-01 to 2014-09-30. Shenzhen has 11 public air quality monitoring stations, and we obtained data monitored by 3 additional stations from Shenzhen Environmental Monitoring Center (SEMC) under non-disclosure agreement for research purposes. Thus we totally have 14 air quality monitoring stations in Shenzhen, with the monitored records updated every hour. The meteorological data are collected from 141 meteorology monitoring stations, updated every 3 minutes. The traffic data are collected from 150+ main roads in Shenzhen, updated every 5 minutes. We map the traffic speed and index to the grids, based on the road network information. To align the timestamps, we average the meteorological and traffic data for each hour. Thus the time lag in the Granger causality model corresponds to 1 hour. Since the meteorological and traffic have fine temporal granularity, their 1-hour averages have

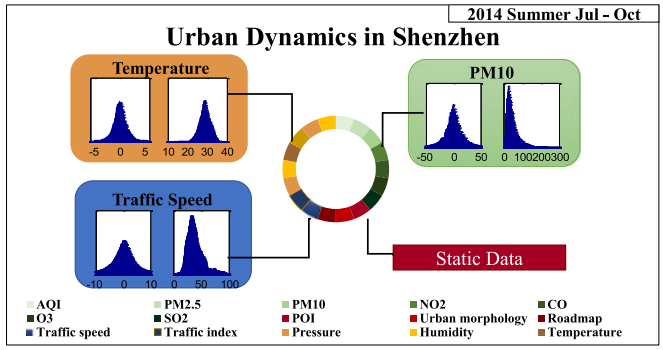


Fig. 3. Histogram for the 1-hour difference (normalized by the standard deviation) of urban dynamics and the original distribution of urban dynamics. We added D'Agostino-Pearson test [36], i.e., testing measures based on the skewness and kurtosis to verify Gaussian distribution, for the 1-hour differences of data (category 1-15) in Shenzhen. The statistics for the 15 categories of urban dynamics data all don't reject the null hypothesis and can be concluded as normally distributed.

very good quality and almost never suffer from missing data. For the air quality data, there are around 5 percent missing data due to either the failure of the government sensor or our crawlers. We fill the missing data for each time series using linear interpolation.

Note that there are three static urban dynamics (category 16-18 in Table 1), which are collected from public maps. We extract the color of 400 pixels in the format of $[0255] \times 3$ within each grid. Then we give different weights to each color, in order to represent different landforms (e.g., water, hill, wetland, etc.). After that, we normalize the summation of the 400 weighted pixel color levels within a grid to the range of [0, 1]. The weighting and scaling for map data target at representing each grid by summarizing different kinds of urban information within it. We collect the snapshot image from Baidu map (without black contents such as text and border) to simplify the processing, with different colors indicating different layouts (e.g., water, hill, wetland, etc.). This method is similar to another study in Hong Kong [31], where urban environmental quality are monitored using satellite data and by weighting and scaling multiple parameters. Since the time series data are processed as 1-hour differences, which obey the normal distribution, we scale the static values to the range [0, 1] which are close to other input values and thus help our model better converge. The raw datasets in Hong Kong are in similar formats as Shenzhen, the difference being that the time period for Hong Kong data is 2013-01-01 $\sim$ 2014-12-31.

2.3.2 Causality for Two Urban Dynamics

We divide the whole city into a set M of square grids, and s_x, s_y represent two regions $\in M$. For mathematical representation of Granger causality, we use X and Y to identify whether X Granger causes Y. Specifically, to measure how one urban dynamic $X(s_x; t; c_i)$ in grid s_x cause another $Y(s_y; t; c_j)$ in grid s_y, where c_i, c_j represent two categories of urban dynamics, t corresponds to the time, we need to extend Granger causality to the ST format based on (1):

$$Y(s_y; t; c_j) = \Sigma_{k=1}^{L_j} a_k Y(s_y; t-k; c_j) + \Sigma_{k=0}^{L_i} b_k X(s_x; t-k; c_i) + \Sigma_{k=0}^{L_i} r_k Z_{i,t-k} + \xi_t \quad (3)$$

This can be described as "single cause" to explain, for example, how the wind speed at grid s_x and time $[t - L_i, t]$ cause the PM2.5 concentration at grid s_y and time t. L_i, L_j are the numbers of timestamps for urban dynamics c_i, c_j. Vectors $\boldsymbol{a} = \{a_k\}$ $(k = 1, 2, \ldots, L_j)$, $\boldsymbol{b} = \{b_k\}$ $(k = 0, 1, 2, \ldots, L_i)$ are the weights for the two processes. Z_t is a Gaussian process with mean 0 and variance 1. It serves as a Gaussian noise to justify the dependencies among air pollutants and other urban dynamics, with vector $\boldsymbol{r} = \{r_k\}$ $(k = 0, 1, 2, \ldots, L_i)$ corresponding to its weights. The weights $\boldsymbol{r}$ are further utilized in a threshold function ϕ of the following definitions of causality measures (Equations (5) and (7)). ξ_t is the residual. The reason why vectors $\boldsymbol{b}$ and $\boldsymbol{r}$ start from b_0 and r_0 is that an urban dynamic is influenced by the historical dynamics and the current neighbourhood dynamics.

The two processes are generated by 1-hour difference of time series mentioned above. In this way, the null hypothesis can be written as $X(s_x; t; c_i)$ is not a Granger cause for $Y(s_y; t; c_j)$ when $b_k = 0$ for $k = 0, 1, 2, \ldots, L_i$. Here L_i, L_j are the time lags that minimize the expected residual ξ_t, which can be achieved by partial autocorrelation function (PACF) analysis for time series $Y(s_y; t; c_j)$ and $X(s_x; t; c_i)$ [6]. For example, considering a time series x_t, the partial autocorrelation between x_t and x_{t-L} is defined as the conditional correlation between x_t and x_{t-L}, conditioned on $x_{t-L+1}, \ldots, x_{t-1}$. We select the optimal number of lags L as the first value less than a 5 percent significance level. L_i, L_j are determined before the non-causality test, based on selecting the optimal time lags for the partial autocorrelation function. Thus they will not affect the behavior of the causality measurement factors. Vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{r}$ can be estimated by minimizing the error between the observed responses $Y(s_y; t; c_j)$ and the predicted responses $\hat{Y}(s_y; t; c_j)$ in Equation (3). When feeding the 1-hour differences for the time series and the scaled static data to Equation (3), we can obtain the weight vectors $\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{b}}$, and $\hat{\boldsymbol{r}}$ for the past air quality variable, the urban dynamics variables, and the Gaussian noise with 0 mean and variance 1. $\hat{\xi}_t$ corresponds the minimized error, i.e., the expected value of residuals ξ_t, represented as $\hat{\xi}_t = Expect(||Y(s_y; t; c_j) - \hat{Y}(s_y; t; c_j)||)$.

2.3.3 Causality for Multiple Urban Dynamics

To measure how $N(N \geq 2)$ urban dynamics $(s_x; t; c_1)$, $X(s_x; t; c_2), \ldots, X(s_x; t; c_i), \ldots, X(s_x; t; c_N)$ cause another urban dynamic $Y(s_y; t; c_j)$, we represent the Granger causality for multiple urban dynamics at two locations s_x and s_y in the following format:

$$Y(s_y; t; c_j) = \Sigma_{k=1}^{L_j} a_k Y(s_y; t-k; c_j) + \Sigma_{i=0}^{N} \Sigma_{k=0}^{L_i} b_{ik} X(s_x; t-k; c_i) + \Sigma_{k=0}^{L_i} r_k Z_{i,t-k} + \xi_t \quad (4)$$

This could be comprehended as "multiple causes", i.e., how all the urban dynamics in grid s_x cause the urban dynamic $Y(s_y; t; c_j)$ in grid s_y.

3 Causality Analysis

This section deals with ST causality analysis and its visualization with urban data from Shenzhen. Based on the ST extended causality model, we propose two causality measurement factors F_{test} and F_c for ST data.

TABLE 3
Non-Causality Test Between Urban Dynamics and Air Pollution in Shenzhen

Max(Lj,Li)	AQI	PM2.5	PM10	NO2	CO	O3	SO2	Pressure	Humidity	Temperature	1 hour rain	24 hour rain	Wind	Traffic speed	Traffic index
AQI	–	(1,5)	(1,0)	(1,3)	(1,0)	(1,3)	(1,4)	(1,0)	(1,0)	(1,0)	(1,0)	(1,0)	(1,0)	(1,1)	(1,1)
PM2.5	(4,2)	–	(4,0)	(4,1)	(4,0)	(4,0)	(4,0)	(4,0)	(4,2)	(4,0)	(4,1)	(4,0)	(4,2)	(4,1)	(4,1)
PM10	(5,0)	(5,0)	–	(5,2)	(5,0)	(5,0)	(5,0)	(5,0)	(5,0)	(5,0)	(5,1)	(5,0)	(5,2)	(5,2)	(5,2)
NO2	(5,0)	(5,0)	(5,0)	–	(5,0)	(5,5)	(5,2)	(5,0)	(5,1)	(5,1)	(5,0)	(5,0)	(5,0)	(5,1)	(5,1)
CO	(5,0)	(5,0)	(5,0)	(5,0)	–	(5,1)	(5,2)	(5,0)	(5,0)	(5,0)	(5,0)	(5,0)	(5,0)	(5,0)	(5,0)
O3	(12,4)	(12,0)	(12,0)	(12,1)	(12,0)	–	(12,2)	(12,0)	(12,1)	(12,1)	(12,0)	(12,0)	(12,0)	(12,1)	(12,1)
SO2	(8,4)	(8,0)	(8,4)	(8,5)	(8,0)	(8,4)	–	(8,0)	(8,0)	(8,0)	(8,1)	(8,0)	(8,0)	(8,2)	(8,2)
														Non-causality	

There are 15 columns of urban dynamics time series (X) and 7 rows of air pollutants time series (Y) for the Granger causality test based on their F_{test} measures. For each pair of relationship $Y \leftarrow X$, we first record the time delays L_j, L_i that maximize the corresponding F_{test}, and then mark the relationship where $F_{test} < 1$ to be red. $F_{test} < 1$ indicates X does not "Granger" cause Y. For example, in Shenzhen, Pressure does not "Granger" cause AQI. We thus rule out this relationship.

3.1 Non-Causality Test

We first define the causality measurement factor testing whether $X(s_x; t; c_i)$ "Granger" causes $Y(s_y; t; c_j)$ to be:

$$F_{test}(Y \leftarrow X | s_y; c_j; L_j; s_x; c_i; L_i) = \frac{\phi\left(||\hat{\boldsymbol{b}}|| - ||\hat{\boldsymbol{r}}||\right)}{||\hat{\boldsymbol{a}}|| \times \hat{\xi}_t} \quad (5)$$

$||\cdot||$ represents the ℓ_1-norm of vector $\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{b}}$, and $\hat{\boldsymbol{r}}$. L_i, L_j are calculated by the method mentioned in Section 2.3.2, $\hat{\xi}_t$ is the minimized error, and ϕ corresponds to a threshold function:

$$\phi\left(||\hat{\boldsymbol{b}}|| - ||\hat{\boldsymbol{r}}||\right) = \begin{cases} 0, & ||\hat{\boldsymbol{b}}|| - ||\hat{\boldsymbol{r}}|| < ||\hat{\boldsymbol{r}}|| \\ ||\hat{\boldsymbol{b}}|| - ||\hat{\boldsymbol{r}}||, & otherwise \end{cases} \quad (6)$$

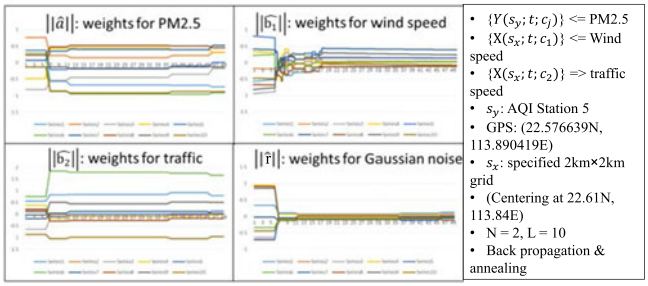
The threshold function equals 0 when the difference of the ℓ_1-norm of vector $||\hat{\boldsymbol{b}}||$ and $||\hat{\boldsymbol{r}}||$ is less than the ℓ_1-norm of weights vector $||\hat{\boldsymbol{r}}||$ for Gaussian (0, 1) process. We consider non-causality when $F_{test} = 0$ for every s_x, s_y. Table 3 illustrates the non-causality test between 18 urban dynamics and 7 air pollution dynamics for Shenzhen. Results rule out 22 percent of irrelevant bipartite relationships.

3.2 Causality Observation

Then, to represent the causality of multiple urban dynamics at the spatial level, we define the causality measurement factor for multiple ST processes $(s_x; t; c_1)$, $X(s_x; t; c_2), \ldots, X(s_x; t; c_i), \ldots, X(s_x; t; c_N)$, and $Y(s_y; t; c_j)$ as:

$$F_c(Y \leftarrow X | s_y; c_j; L_j; s_x; c_1; L_1; \ldots s_x; c_i; L_i; \ldots s_x; c_N; L_N) = \frac{\sum_{i=1}^{N}\left(||\hat{\boldsymbol{b}}_i|| - ||\hat{\boldsymbol{r}}||\right)}{N \times ||\hat{a}|| \times \hat{\xi}_t} \quad (7)$$

The area of Shenzhen is divided into $50 \times 25 = 1{,}250$ square grids, covering the longitude and latitude of (113.75-114.65E, 22.45-22.85N). Thus each grid is around $2\ \text{km} \times 2\ \text{km}$. We set the timestamps for urban dynamics (category 1-15) based on Table 2, with normalized 1-hour differences as the input time series. For static data (category



(a) 50 training iterations for the Granger causality weights (b) Parameter setting

Fig. 4. Training of Granger causality weights: How do wind and traffic affect PM2.5?

16-18), there are 1,250 grids and 400 pixels for each grid, so the total data volume for each geographical data is 500,000 bytes, of which we use constant values as input.

To better illustrate the causality measures, we give an example of the training process in Fig. 4. We would like to determine how the urban dynamics wind and traffic will affect PM2.5. We set the category of urban dynamics as $N = 2$, and the time lags in the causality model as $L = 10$. $Y(s_y;t;c_j)$ corresponds to PM2.5 at AQI station 5, located at (22.576639N, 113.890419E). $X(s_x;t;c_1)$ and $X(s_x;t;c_2)$ correspond to wind speed and traffic speed at the center of a $2\text{ km} \times 2\text{ km}$ grid, located at (22.61N, 113.84E). We use back propagation and annealing to train the parameters $\boldsymbol{a}, \boldsymbol{b_1}, \boldsymbol{b_2}, \boldsymbol{r}$ in Equation (7). The X-axis in Fig. 4a ranges from 1-50, representing 50 iterations. As shown in the figure, the vectors $\boldsymbol{a}, \boldsymbol{b_1}, \boldsymbol{b_2}, \boldsymbol{r}$ converge to different values. And after about 7 iterations, the weights for the Gaussian noise $\boldsymbol{r}$ converge to about 0. This indicates that the threshold function (6) in the causality measurement factor F_{test} and F_c would be very effective in differentiating "urban dynamics" from "Gaussian noise".

Fig. 5a illustrates the causality observation of different grids in Shenzhen. It aims to reveal the causalities between different grids. Factor F_c is quantified to discrete levels 1-10 for visualization. Ten colors are utilized to represent the ten levels, with red representing the highest causality and blue representing the lowest (see the color index in Fig. 5a). We can see that the grids that have higher influences to the target monitoring station (Station 5) are not only located at the neighborhood locations, but also at some farther locations. To show the effectiveness of the F_c factor, we present the corresponding histogram of causality levels based quantified F_c values in the left of Fig. 5a. On the right of Fig. 5b, we present the histogram of F-test measure, which is commonly used for non-causality test [37], or testing null hypothesis for statistical models [38]. The F-test measure in our experiment is defined as F-test $= \frac{\hat{\xi}'_t - \hat{\xi}_t}{\hat{\xi}_t}$, where $\hat{\xi}'_t$ is the expected regression error of the target urban dynamic $Y(s_y;t;c_j)$ regressed by its own history:

$$Y(s_y;t;c_j) = \Sigma_{k=1}^{L_j} a_k Y(s_y;t-k;c_j) + \xi'_t \tag{8}$$

Usually, the null hypothesis is rejected if the F-test calculated from the data is greater than a specific value where the F-distribution has a desired false-rejection probabiltiy (e.g., 0.05). However, the F-test measure could also reflect the information gain for the time series $Y(s_y;t;c_j)$ given the time series $(s_x;t;c_1)$, $X(s_x;t;c_2),\ldots,X(s_x;t;c_N)$. We thus

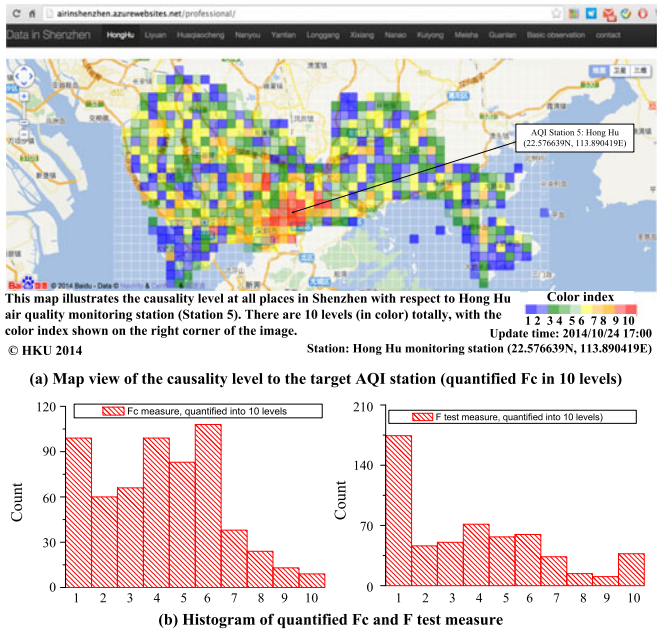


Fig. 5. Statistics of the causality model for Shenzhen data.

also quantize the F-test measure to 10 levels for comparison. As shown in Fig. 5b, the quantized F-test measure have many more grids with very high influences (level 10) to the target AQI stations, compared to quantified F_c measure. This may be explained by the definition of F-test. When there are not enough training samples in the regression Equation (4), $\hat{\xi}_t$ could have very small values, thus making the F-test value very large. While the F_c measure is dependent on both a threshold function ϕ and the regression error $\hat{\xi}_t$, which makes the measure more robust to Gaussian noises and helps select better region of influence (ROI) in the air quality map inference stage. We will introduce the concept of ROI in the next section.

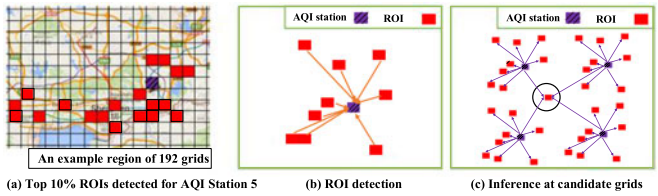
4 Air Quality Map Inference

The purpose of generating air quality map is to give a complete picture of air quality in Shenzhen and help people understand better about the urban dynamics. However, processing the massive volume of urban dynamics data poses a challenge. To overcome this, we propose a region of influence (ROI) approach to detect data with the highest causality levels spatially and temporally, thus allowing us to process "part" of the data instead of "all" of the data. In this section we shall elaborate on ROI detection as well as the algorithm for air quality map estimation.

4.1 Region of Influence (ROI)

ROI is defined as a set of grids which constitute the top ρ percent of all grids ranked by the causality factor F_c to a specified urban dynamic process at a specified grid.

Taking Shenzhen as an example, for PM2.5 at a grid, the ROIs for this grid are selected from other grids which constitute the top $\rho = 10$ percent of all grids ranked by F_c values. Fig. 6a shows an example, with a total 192 grids and the top 10 percent (or 19) ROIs (marked in red) for PM2.5 at the grid where Station 5 (Hong Hu, marked in



(a) Top 10% ROIs detected for AQI Station 5 (b) ROI detection (c) Inference at candidate grids

Fig. 6. Region of influence (ROI) detection.

purple and shaded) is located. In the beginning, we calculate the ROIs for the grids with AQI stations. Then, we infer the air quality in these detected ROI grids. Fig. 6b shows the ROIs for a specific AQI station. Fig. 6c shows the logic of ROI detection and AQI estimation. For example, we have 14 monitoring stations in Shenzhen, and for each station we identify the top 10 percent ROIs. Thus there must be some grids influenced by at least 4 grids with AQI values. We infer the AQI in these grids first, based on urban dynamics (category 1-18 in Table 1) at the n_{mig}= 4 most influencing grids based on backpropagation neural networks [9]. We deploy a neural network with a 20-node hidden layer based on C# Jeffheaton-book-code [32]. Now many grids will have AQI values at selected timestamps. Then we gradually fill the historical AQI values of other grids and timestamps, finally marking all the ROIs for each grid.

We set $\rho = 10$ percent since it generates the best estimation results in Shenzhen. We set $\rho = 10$ percent since this generates the best estimation results in Shenzhen. When ρ is too large, there will be too many grids selected and some of them may be useless or redundant for the estimation task. When ρ is too small, there may be no candidate grids (the overlapped grids covered by multiple monitoring stations) as in Fig. 6c, indicating the air pollution inference iterations may stop prematurely. Since Shenzhen and Hong Kong are geographically close and have similar layouts, we also set $\rho = 10$ percent when migrating the ROI detection and AQI estimation procedure to Hong Kong, which has 15 AQI monitoring stations. Thus there will be n_{mig}= 5 influencing grids for grids without monitoring stations.

4.2 Algorithm Description

Fig. 7 shows the work flow of ROI detection and AQI estimation, separated into 5 stages. The first stage interacts with input data flow, from both online and historical data. We also display the output data flow of causality model observation and air quality map for visualization.

Stage 2 deals with non-causality detection, which is based on the ST extended Granger causality model as well as the pre-defined F_{test} factor. We rule out the urban dynamics that do not "Granger" cause air pollution, and in the meanwhile record the maximum timestamps L_i, L_j for urban dynamics. Thus N categories of urban dynamics within the time window $[t - L_i, t]$, $i \in 1, 2, \ldots, N$ are selected for further estimation.

Stages 3 and 4 are integrated and iterative. The training stage (Stage 3) first calculates the causality measurement factor F_c for each grid with AQI values, and trains the parameters of a backpropagation neural network using the AQI values in a grid as output and the urban dynamics

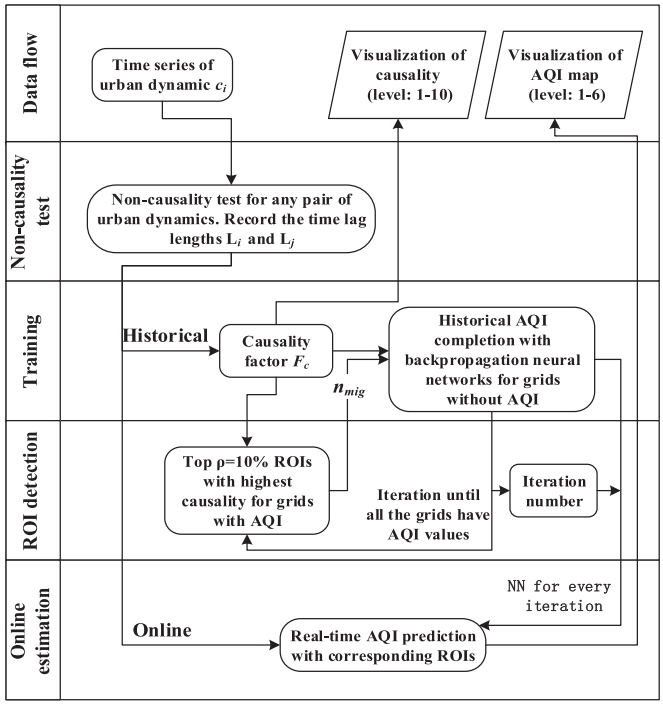
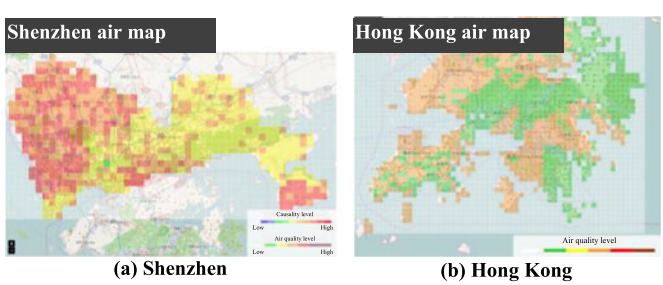


Fig. 7. ROI detection and AQI estimation work flow.

at its n_{mig} ($n_{mig} = 4$ *or* 5 in Shenzhen or Hong Kong) most influencing grids as input. Then in the ROI detection phase (Stage 4), we mark the top $\rho = 10$ percent ROIs for each grid with AQI values. Afterwards, Stage 3 again selects new grids without AQI values influenced by at least 4 grids with AQI values and estimate AQIs with the trained neural network previously. This phase combines ROI detection with historical air quality estimation, gradually training the parameters of the neural network for each iteration until all the grids are filled with AQI values. The neural networks and order of grids for estimation are recorded for Stage 5.

The last stage, i.e., the online estimation stage, is based on the neural networks trained in Stages 3 and 4. The fine-grained air quality map is generated after inferring the AQI in each grid with urban dynamics from its ROIs. Fig. 8a shows an air quality map at 2014-09-19 11:00:00 AM, which is an example of our visualization demo for city-wide air quality map in Shenzhen, assuming air quality inside each $2\ \mathrm{km} \times 2\ \mathrm{km}$ grid is the same (demo available at http://www.youtube.com/watch?v=CwfQvTeftbM&feature=youtu.be). Fig. 8b shows the air quality map of Hong Kong at 2015-04-09 7:00:00 PM, with grid size set to be around $1\ \mathrm{km} \times 1\ \mathrm{km}$. The selection of grid size in the two cities is elaborated in Section 6.1.



(a) Shenzhen (b) Hong Kong

Fig. 8. Air quality map.

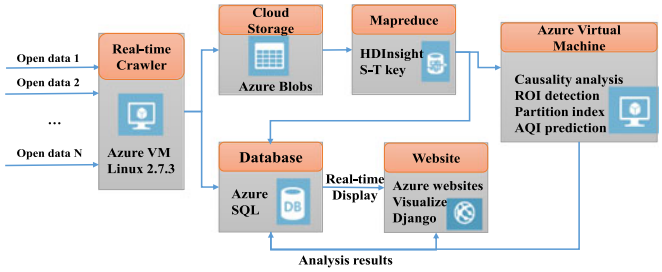


Fig. 9. Deployment on Windows Azure.

5 Performance Evaluation

5.1 System Deployment

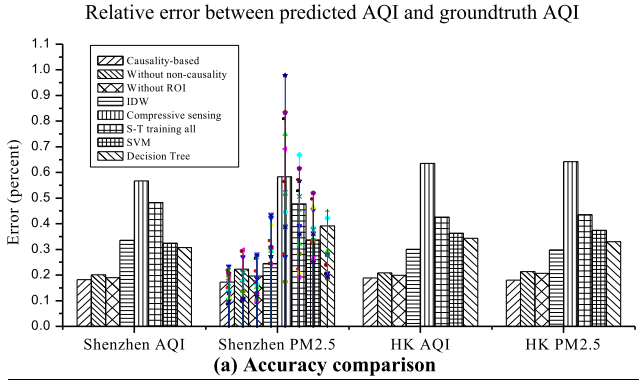
We deploy the whole system on Windows Azure Platform (http://azure.microsoft.com), with a flow diagram shown in Fig. 9. Data crawler is implemented on an 8-core Linux 12.04 virtual machine, with about 0.8T/month raw data collected. We synchronize the collected real-time data in Azure Blob storage for Mapreduce indexing, with 24 core HDInsight. Processed data are stored in two SQL servers each of 250 GB. Visualization is deployed at an Azure website service.

5.2 Accuracy and Time Efficiency

The accuracy and time efficiency of our approach are evaluated and compared with five reference methods, i.e., causality-based approach without non-causality analysis, causality approach without ROI detection, inverse distance weighting (IDW), compressive sensing, and spatially and temporally (ST) training all the urban dynamics. If there are N stations, we use N—1 stations to predict the remaining one. The error is calculated by comparing the predicted value with the ground truth value for each of the N stations and averaging them, with the error for each monitoring station represented as $Error = (AQI_{\_predicted} - AQI_{\_groundtruth}) / AQI_{\_groundtruth}$. We predict real values of AQI and PM2.5.

Fig. 10a shows causality-based AQI estimation performs better than the other seven methods. Performances drop a bit when we remove the non-causality detection phase or ROI-detection phase separately. We also note that when removing the ROI detection phase, the error increases less compared to removing the non-causality detection phase. This indicates the influence of data diversity could be more critical than the ST influence.

For IDW, we generate the air quality map based on 2D linear interpolation, with different weights reflecting different layouts [7]. Compressive sensing is a signal processing technique for efficiently acquiring and reconstructing a sparse signal [8]. Consider a city with N monitoring stations and being separated into M grids. We reconstruct a fine-grained air quality map (a signal vector with M values) from N monitored values based on $||l_1||$ minimization. "ST training all" is a method where we predict the AQI in one grid based on all the dynamics and their ST correlations by backpropagation neural network [9]. Taking AQI Station 5 in Shenzhen as an example, the estimation is based on training the dynamics category 1-18 within time $[t - L, t]$ and distance radius R (1-25 km). We set L to be 10, thus the input size will be $15 \times 10 \times 25 + 3 \times 1 \times 25 = 93{,}825$, since features are averaged for each radius and there is only one timestamp for static geographic data. We set $R_{\max} = 25$, $L = 10$ because this



(a) Accuracy comparison

Time efficiency for training (second)		
Approach	Shenzhen	HK
Causality-based (1VM, 2 cores)	311904	246732
Causality-based (4VMs, 8 cores)	121132	94813
S-T training all (1VM, 2 cores)	1158510	813543
S-T training all (1VM, 8 cores)	284666	205487

(b) Time efficiency comparison

Fig. 10. The performance of proposed causality-based approach with other methods.

parameter setting generates the best performance for the "ST training all" scenario. Our causality-based approach outperforms the "ST training all" approach, meaning it is desirable to decompose "big data" into "small data" and just process the most influential data.

Granger causality is basically a linear regression-based model and a feature-level-based data fusion method [27]. We added two more feature-level-based data fusion method, namely, support vector machine (SVM) [29] and decision tree (DT) [30], in addition to the existing neural network based "S-T training all" method, for a more comprehensive comparison.

To estimate the uncertainty of the model, we add a simulation study for the errors of PM2.5 at N = 14 monitoring stations in Shenzhen, and scatter the result on Fig. 10a. The variance of error for our method is 0.097, which is less than all of the other baseline methods (the order of variance of error is our model < our model without ROI < our model without non-causality < IDW < decision tree < SVM < ST training all < compressive sensing). The variance of error for IDW is 0.113, indicating higher uncertainty for the estimation model.

Fig. 10b compares the efficiency of our causality-based approach and the "ST training all" approach. The causality-based approach deployed on 4 independent virtual machines based on the ROI data improves the time efficiency by 2.5-9.56 times. We did not compare the error directly with the method in [1], because [1] only estimated AQI levels (1-6) not AQI values, and it is impossible to achieve exact training parameters for a reasonable comparison. However, based on comparing estimation results in [1] at 3 non-public AQI stations in Shenzhen from our database, we observe our approach gives about 5 percent higher precision when transforming AQI value to levels.

6 Dealing with Inaccuracies Due to Grid Size

6.1 Inaccuracies of Grid Size

Basically, we assume that each category of urban dynamics (time series at a timestamp or static data) at all the

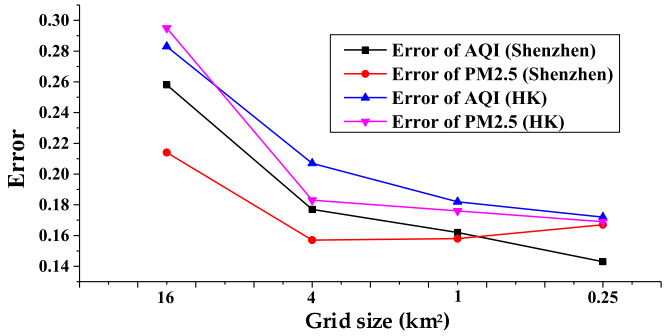


Fig. 11. Estimation error versus grid size.

locations inside one grid share one value. However, due to the different layouts and POIs throughout a city, the grid size usually causes significant inaccuracies in the causality model. For example, too large a grid size will average different geographical factors, while too small a grid size will cause duplicated computation since urban dynamics at neighborhood grids may have almost the same values. Fig. 11 illustrates how different grid sizes, i.e., 4×4, 2×2, 1×1, and 0.5×0.5 square kilometer, could affect the accuracy of the causality model. Four tasks, i.e., the city-wide AQI and PM2.5 estimation in Shenzhen and Hong Kong, are evaluated based on the estimation error. We observe the error tend to decrease as we decrease the grid size.

However, the computational complexity of the causality model is proportional to the number of grids. If we decrease grid size from 16 km^2 to 0.25 km^2, the mean error will be reduced from 0.2625 to 0.1628 by about 38 percent but the time consumption will increase by $16/0.25 = 64$ times. To address this, we need a proper grid size that trades off between performance and time efficiency. Based on the results shown in Fig. 11, we choose $2 \times 2 = 4\ \text{km}^2$ as the default grid size in Shenzhen and $1 \times 1 = 1\ \text{km}^2$ for Hong Kong. Since the two selected grid sizes achieve a satisfactory performance (error < 0.2) for the causality model while the computation time is affordable.

6.2 Flexible Grid-Based Estimation Algorithm

6.2.1 Motivation

The causality based air quality estimation model is grid-based, yet there are many non-grid based application scenarios. For example, there are social-economic data available for 412 constituency areas in Hong Kong since the last census in 2011, with each constituency area (CA) has a population of around 17,000 people on average that we assume do not change (see details at http://www.elections.gov.hk/dc2011/eng/summary.html). For the social study research, e.g., the correlation between air quality and income, health, education, etc., we need to estimate the air quality in each CA.

To minimize the inaccuracies caused by grids, and adapt the causality model to non-grid based application scenarios, a methodology is needed to convert the grid-based estimation to non-grid-based. Fig. 12 illustrates an example of the non-grid-based scenario for air quality estimation in the 412 CAs of Hong Kong (as shown in Fig. 12a). In this case, most CA covers more than one grid, but each grid is usually

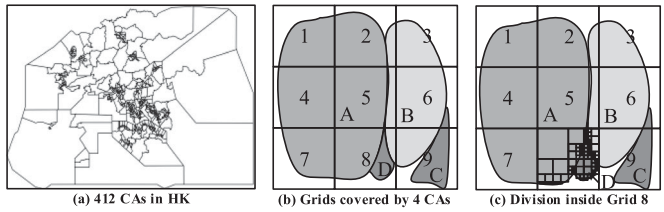


Fig. 12. An example of non-grid based application scenario.

shared by many CAs. For example, CA "A" covers grids 1, 2, 4, 5, 7, and 8, while grids 2, 5, and 8 are all shared by more than 2 CAs, as shown in Fig. 12b. This will make all the CAs to have similar estimated air quality values, especially when the grid size is large, since the shared grids will be included in the calculation of the corresponding CAs.

One way to reduce the effect of similar estimated values at non-grid-based areas is to decrease the grid size, thus reducing the number of shared grids for the averaging calculation. Yet this will greatly increase the computation time, since each time we divide a grid into 4 sub-grids, and iterate this type of division 4 times, the computation time will increase by $4^4 = 256$ times.

Since not all grids are needed to be divided, and the causality model has already generated the air quality map at $1 \times 1 = 1\ \text{km}^2$ grid size in Hong Kong which could best reduce the inaccuracies and guarantee time efficiency of the original causality model, we propose a flexible grid-based estimation algorithm (FGEA afterwards) that 1) selectively divides the grid which needs to be divided into four sub-grids (as shown in Fig. 12c, where Grid 8 covered by CA "A", "B" and "D" is divided on demand 4 times), and 2) updates the air quality in the sub-grids based on the parameters learnt previously by the causality model.

6.2.2 Algorithm Description

Fig. 13 shows the description of the proposed FGEA, by adding a module to the previous ROI detection and AQI estimation work flow Stage 5 (as shown in Fig. 13a).

Originally for Hong Kong, we estimate the grids with no AQI values by its $n_{mig} = 5$ most influencing grids with AQI values (shown as Fig. 13b), based on the causality measure-

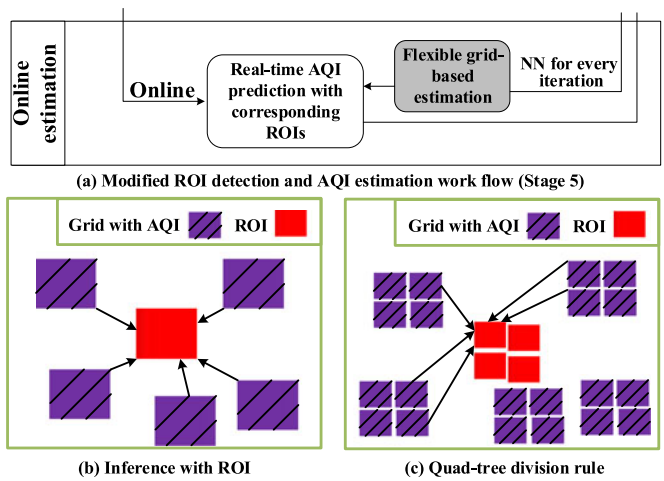


Fig. 13. Description of the flexible grid-based estimation algorithm (FGEA).

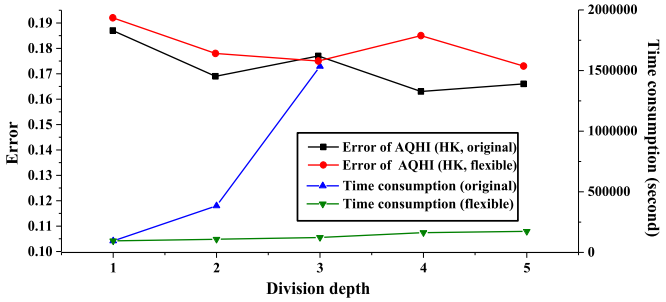


Fig. 14. Estimation accuracy and time consumption of FGEA, compared with the original causality model.

ment factor F_c. In the FGEA module, we first judge whether to divide a grid into four sub-grids. If division is required, we further divide the 5 most influencing grids by quad-tree (a tree data structure most frequently used to subdivide a region into four quadrants or regions [24]), and select the further top 5 sub-grids that have the highest F_c to the target sub-grid. The reason for choosing new most influencing grids from the sub-grids of the previous most influencing grids is due to the assumption that the sub regions of the most influential regions after division is also the most influential ones to a target grid.

The division decision includes the following four conditions, given a division depth H, ($H = 5$) in the experiment:

Case 1: A grid is covered by more than two (>2) CAs
Quad-tree based division;
Case 2: A grid is just covered by one CA
Stop division;
Case 3: A grid is covered by two (= 2) CAs with the smaller one covers > 10 percent.
Quad-tree based division;
Case 4: A grid is covered by two (= 2) CAs with the smaller one covers < 10 percent.
Stop division, and allocate the grid to the CA that has larger coverage;

All the urban dynamics in a new grid will be recalculated. Since the static factors are pixel based, when the grid changes, the number of pixels within each grid will change accordingly. The air quality values in the sub-grids are estimated by the same neural network trained previously. Since the proposed algorithm just chooses the subsets of the most influencing grids for each division, the computation complexity is greatly reduced. Assuming there are M grids at the beginning, for each division, there will be $4 \times n_{mig} \times 4 = 80$ new calculations for the causality factor F_c instead of $4 \times (M-1) \times 4$ calculations, where $n_{mig} \ll M$. Another advantage of the algorithm is it uses the previously trained parameters for new estimation in the sub-grids, instead of training these parameters again.

6.2.3 Evaluation

This section evaluates the effectiveness of FGEA in terms of its estimation accuracy and time consumption, as well as a metric we defined to represent the similarity between different non-grid-based areas.

Fig. 14 demonstrates the estimation accuracy and time consumption of the FGEA estimation algorithm. As mentioned in Section 6.2.1, we start from the grid size of $1\ \mathrm{km}^2$ and set the the division depth from 1 to 5, i.e., the smallest grid size after division is $\frac{1}{2^{5\times2}}\ \mathrm{km}^2$, which can best fit the CA-based division in Hong Kong. For comparison, we use the original causality model without the FGEA estimation algorithm, which divides all the grids to four sub-grids per division depth and trains the causality model again.

TABLE 4
Similarity Metric versus Division Depth for Different Methods

Methods\Divison depth	1	2	3	4	5
Causality model with FGEA	0.721	0.782	0.858	0.897	0.923
Causality model without FGEA	0.673	0.705	0.777	0.792	0.834
Grid-based IDW	0.63	0.691	0.661	0.714	0.726

The left Y-axis shows the errors of the non-grid-based air quality estimation with and without FGEA. The errors are similar and show a slight decreasing trend, indicating the accuracy of FGEA is comparable to the original proposed causality model. Note that the initial grid size for division is $1\ \mathrm{km}^2$, not a larger grid size. This is because $1\ \mathrm{km}^2$ grid size best reduces the inaccuracies, and is thus capable of estimating the air quality with the most influencing sub-grids in the same area covered by the previous most influencing grids.

The right Y-axis illustrates the time consumption for air quality estimation with and without FGEA. The computation time for the original causality model without FGEA consumes exponential amount of time as the division depth increases, since the computation complexity of the causality model is proportional to the number of grids. Since there will be too much computation when the division depth H is larger than 3, we skip the cases of $H = 4\ or\ 5$. For the air quality estimation with FGEA, the time consumption does not increase much as H increases. FGEA's significant improvement of the time efficiency is due to both the simplified most influencing grid selection and the utilization of previously trained paymasters.

We also define a similarity metric to observe the relative similarity of estimated air quality values in different areas. The similarity metric is defined as:

$$\text{Similarity } metric = \frac{variance\ of\ AQHI\ at\ all\ the\ estimated\ areas}{variance\ of\ AQHI\ at\ N\ monitoring\ stations}$$

$Metric = 1$ indicates the estimated air quality at all the areas have the same variance as the N air quality values monitored by stations. The smaller the metric the more similar the estimated values. Table 4 lists the similarity metrics for the averaged AQHI in Hong Kong during Jan 2015, versus the division depth, using causality model with FGEA, causality model without FGEA, and a grid-based IDW [7] estimation. The AQHI value in each area is the averaged value of all the grids covered by the area. The division depth is also set from 1 to 5, with each grid divided to 4 sub-grids when the depth is increased by 1.

The grid-based IDW method generated the smallest similarity metric (the highest similarity), since this method is based on interpolation. The averaging of linearly interpolated values at different grids will greatly

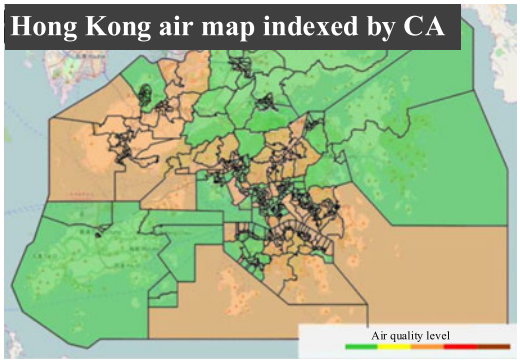


Fig. 15. Non-grid-based AQHI map visualization in Hong Kong at the same timestamp as Fig. 8.

decrease the relative variance. The causality model without FGEA have smaller similarity metrics compared to the causality model with FGEA. This is because for causality model without FGEA, there are many grids shared by different areas, thus reducing the overall variance of all areas. The causality model with FGEA demonstrates the minimum similarity between different non-grid-based areas as the division depth increases and the number of shared grids decreases.

Fig. 15 visualizes the Hong Kong air quality map indexed by 412 CAs, as a non-grid-based application scenario enabled by FGEA. The time stamp is 2015-04-09 7:00:00 PM, which is the same as in Fig. 8.

7 Conclusion

In this paper, we estimated city-wide air quality with limited AQI stations based on Granger causality analysis, and overcame the challenges due to ST heterogeneous urban big data. We proposed an ST extended Granger causality model which analyzes all the causalities between urban dynamics and air pollution in a consistent manner. Irrelevant categories of urban dynamics are ruled out by implementing non-causality test. To deal with time efficiency, we also proposed an approach to detect the ROI, thus decomposing "big data" into "small data". The city-wide air quality map is inferred and visualized by combining the training of the urban dynamics with the detection of ROI. Results show our causality model outperforms other interpolation or training approaches considering all the ST correlations, supporting the approach for urban big data computing which just processes the most influential data. We further propose an algorithm enabling flexible grid-based city-wide air quality inference, to reduce the inaccuracies caused by a fixed grid size, and make the Granger causality model feasible for non-grid-based application scenarios.

Appendix A

To know why and what will happen, "cause and effect" has long been discussed from three perspectives, i.e., the unit-level causality, graphical causality, and predictive causality. Note that identifying the absolute causality, such as "what caused a headache?" is usually impossible for mathematical expression. The causes can be a cold, a sleepless night, or an unexpected trouble. Thus in practice, the causality models mostly focus on the relative causality, with probability languages to represent uncertainties, e.g., using $Pr(Y \mid X)$ to represent the effect of Y given X.

This section introduces the basic concepts, models, and methodologies regarding the three main streams of causality, from "stringent" causality to "loose" causality.

A.1 Unit-Level Causality

The stringent causality models originate from Rubin's causality model [3], based on intervention and counterfactual analysis. The explanation comes in the form of "Had C not happened, E wouldn't have happened."

We start the mathematical representation of unit-level causality from the probabilistic representation of association (another form of correlation).

Suppose there exists a population U of "units", $Y(u)$ is the response variable for each unit u. Let A be a second variable defined on U, then the joint distribution of Y and A on U is specified by:

$$\Pr(Y = y, A = a) = proportion\ of\ u\ in\ U \quad where\ Y(u) = y\ and\ A(u) = a \tag{9}$$

The association parameter is determined by this joint distribution:

$$\Pr(Y = y|A = a) = \Pr(Y = y, A = a)/\Pr(A = a) \tag{10}$$

And the association can be represented by the conditional dependency:

$$\mathrm{E}(Y|A = a) \tag{11}$$

Causality is actually different from association. Rubin's causality assumes $Y_t(u)$ and $Y_c(u)$ are two potential responses when the unit u is exposed to two different causes, i.e., treatment t and control c. Then the effect of treatment t on each unit u is measured by $Y_t(u) - Y_c(u)$, with the causal-and-effect represented by the corresponding expectation:

$$\mathrm{E}(Y_t - Y_c) = \mathrm{E}(Y_t) - E(Y_c) \tag{12}$$

Applications extended from the unit-level causality frameworks have been developed to estimate the cause-and-effect of medicine on recovery [33], advertising on behavior change [34], genes on phenotype [35], etc.

A.2 Graphical Causality

Graphical causality represents the cause-and-effect with a directed acyclic graph (DAG) [4] or influence diagram (ID) [5]. The graphical model is very effective for modelling the causality among multiple variables. However, this type of modelling is limited by strict assumptions, for example, Markovian condition is assumed where one child node is locally dependent on its parents.

A.3 Predictive Causality

With the requirement of making the causality analysis more operable, predictive causality such as Granger causality [2] is widely deployed. Causality is loosely defined as one time series successfully predicting another, represented as the following format:

$$Y_t = \Sigma_{k=1}^{L} a_k Y_{t-k} + \Sigma_{k=1}^{L} b_k X_{t-k} + \xi_t \quad (13)$$

Here ξ_t are uncorrelated random variables with zero mean and variance σ^2, L is the number of timestamps, vectors $\boldsymbol{a} = \{a_k\}$ and $\boldsymbol{b} = \{b_k\}$ $(k = 1, 2, \ldots, L)$ are the correspondent weights for two processes X_t and Y_t. The null hypothesis that X_t does not cause Y_t is supported when $\boldsymbol{b} = 0$, reducing (12) to:

$$Y_t = \Sigma_{k=1}^{L} a_k Y_{t-k} + \widetilde{\xi}_t \quad (14)$$

This means process Y_t is caused by its own history, not by process X_t. For another extreme case where $\boldsymbol{a} = 0$ and $\boldsymbol{b} \neq 0$, process Y_t is only caused by process X_t, not by its own history.

This research belongs to the third perspective of causality modelling, and attempts to represent how urban dynamics cause air pollution in a predictive way, thus enabling more accurate and time-efficient air quality estimation.

Acknowledgments

This research was supported in part by a grant from Microsoft Research Asia. We also wish to thank the Microsoft Azure team for technical support.

References

[1] J. Y. Zheng, F. Liu, and H. P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 1436—1444.
[2] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: J. Econometric Soc.*, vol. 37, no. 3, pp. 424–438, 1969.
[3] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.
[4] J. Pearl, *Causality: Models, Reasoning and Inference*, vol. 29. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[5] A. P. Dawid, "Influence diagrams for causal modelling and inference," *Int. Statistical Rev.*, vol. 70, no. 2, pp. 161–189, 2002.
[6] W. W. S. Wei, *Time Series Analysis*. Reading, MA, USA: Addison-Wesley, 1994.
[7] Y. Han, J. K. C. Lam, J.Y. Zhu, V. O. K. Li, and J. Bacon-shone, "Are socially deprived exposed to more air pollution in Hong Kong? Air pollution and environmental justice in Hong Kong" submitted to Environmental Science & Technology.
[8] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, 2007.
[9] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neural Networks and Learning Machines*, vol. 3. Upper Saddle River, NJ, USA: Pearson Education, 2009.
[10] J. Schwartz, "Lung function and chronic exposure to air pollution: A cross-sectional analysis of NHANES II," *Environ. Res.*, vol. 50, no. 2, pp. 309–321, 1989.
[11] L. G. Chestnut, J. Schwartz, D. A. Savitz, and C. M. Burchfiel, "Pulmonary function and ambient particulate matter: Epidemiological evidence from NHANES I," *Archives Environ. Health: An Int. J.*, vol. 46, no. 3, pp. 135–144, 1991.
[12] D. W. Wong, L. Yuan, and S. A. Perlin, "Comparison of spatial interpolation methods for the estimation of air quality data," *J. Exposure Sci. Environ. Epidemiology*, vol. 14, no. 5, pp. 404–415, 2004.
[13] D. K. Jha, M. Sabesan, A. Das, N. Vinithkumar, and R. Kirubagaran, "Evaluation of interpolation technique for air quality parameters in Port Blair, India," *Universal J. Environ. Res. Technol.*, vol. 1, no. 3, pp. 301–310, 2011.
[14] P. Gupta, S. A. Christopher, J. Wang, R. Gehrig, Y. Lee, and N. Kumar, "Satellite remote sensing of particulate matter and air quality assessment over global cities," *Atmospheric Environ.*, vol. 40, no. 30, pp. 5880–5892, 2006.
[15] X. Yu, et al., "Efficient sampling and compressive sensing for urban monitoring vehicular sensor networks," *IET Wireless Sens. Syst.*, vol. 2, no. 3, pp. 214–221, 2012.
[16] L. Li, Y. Zheng, and L. Zhang, "Demonstration abstract: Pimi air box: a cost-effective sensor for participatory indoor quality monitoring," in *Proc. 13th IEEE Int. Symp. Inf. Process. Sens. Netw.*, 2014, pp. 327–328.
[17] EPA, "Emissions inventory guidance for ozone national ambient air quality standards (NAAQS) implementation and regional haze regulations - draft," Environmental Protection Agency, Washington, DC, USA, 2014.
[18] M. W. Martin Adams and Kristin Rypdal, "European environment agency (EEA) air pollutant emission inventory guidebook 2013," European Environment Agency, Kongens Nytorv 6, 1050 Copenhagen K, Denmark, Tech. Rep. 1725-2237, 2013.
[19] EPA, "Evaporative emissions from on-road vehicles in motor vehicle emission simulator (moves) 2014," Environmental Protection Agency, Washington, DC, USA, Tech. Rep. EPA-420-R-14-014, 2014.
[20] S. P. Arya, *Air Pollution Meteorology and Dispersion*. New York, NY, USA: Oxford Univ. Press, 1999.
[21] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, and E. Chang, "Diagnosing New York City's noises with ubiquitous data," in *Proc. 16th Int. Conf. Ubiquitous Comput.*, 2014, pp. 715–725.
[22] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1027–1036.
[23] J. Y. Zhu, C. Sun, and V. O. K. Li, "Granger-causality-based air quality estimation with spatio-temporal (S-T) heterogeneous big data" in *Proc. Int. Workshop Smart Cities and Urban Informat.*, 2015, pp. 612–617.
[24] R. A. Finkel and J. L. Bentley, "Quad trees a data structure for retrieval on composite keys," *Acta Informatica*, vol. 4, no. 1, pp. 1–9, 1974.
[25] P. H. Ryan and G. K. LeMasters, "A review of land-use regression models for characterizing intraurban air pollution exposure," *Inhalation Toxicology*, vol. 19, no. sup1, pp. 127–133, 2007.
[26] Y. Zheng, et al., "Forecasting fine-grained air quality based on big data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 2267–2276.
[27] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE Trans. Big Data*, vol. 1, no. 1, pp. 16–34, Jan.-Mar. 2015.
[28] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, 2014, Art. no. 38.
[29] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Inf. Process.-Lett. Rev.*, vol. 11, no. 10, pp. 203–224, 2007.
[30] G. K. F. Tso, and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, 2007.
[31] J. Nichol and M. S. Wong, "Mapping urban environmental quality using satellite data and multiple parameters," *Environ. Planning B: Planning Des.*, vol. 36, no. 1, pp. 170–185, 2009.
[32] J. Heaton, "Introduction to neural networks for C#, 2nd edition," Heaton Research, Incorporated, 2008.
[33] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.
[34] W. Sun, P. Wang, D.Yin, J. Yang, and Y. Chang, "Causal inference via sparse additive models with application to online advertising," in *Proc. 29th AAAI Conf. Artif. Intell.*, pp. 297–303, 2015.
[35] D. S. Wald, M. Law, and J. K. Morris, "Homocysteine and cardiovascular disease: Evidence on causality from a meta-analysis," *BMJ*, vol. 325, no. 7374, p. 1202, 2002.
[36] R. A. L. P. H. D'Agostino and E. S. Pearson, "Tests for departure from normality. Empirical results for the distributions of b2 and $\sqrt{}$ b1," *Biometrika*, vol. 60, no. 3, pp. 613–622, 1973.
[37] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," *Phys. Rep.*, vol. 441, no. 1, pp. 1–46, 2007.
[38] D. A. Dickey and W. A. Fuller, "Likelihood ratio statistics for autoregressive time series with a unit root," *Econometrica: J. Econometric Soc.*, vol. 49, no. 4, pp. 1057–1072, 1981.

Julie Yixuan Zhu received the BE degree in electronic engineering from Tsinghua University, Beijing, China, in 2012, and the PhD degree in the Department of Electrical & Electronic Engineering, The University of Hong Kong (HKU), in August 2016. Her research interests include urban computing, spatio-temporal data mining, and big data analytics.

Chenxi Sun received the BE degree in electrical and electronic engineering from The University of Hong Kong (HKU), Hong Kong, in 2015. She is working toward the PhD degree in the Department of Electrical & Electronic Engineering, The University of Hong Kong (HKU). Her research interests include spatiotemporal data analytics, network science and applications. She is a student member of the IEEE.

Victor O.K. Li (S'80–M'81–F'92) received the SB, SM, EE and ScD degrees in electrical engineering and computer science from MIT, in 1977, 1979, 1980, and 1981, respectively. He is chair professor of information engineering and head of the Department of Electrical and Electronic Engineering with the University of Hong Kong (HKU). He has also served as associate dean of engineering and managing director of Versitech Ltd., the technology transfer and commercial arm of HKU. He served on the board of China.com Ltd., and now serves on the board of Sunevision Holdings Ltd., listed on the Hong Kong Stock Exchange. Previously, he was a professor of electrical engineering with the University of Southern California (USC), Los Angeles, California, and the director of the USC Communication Sciences Institute. His research interests include technologies and applications of information technology, including clean energy and environment, social networks, wireless networks, and optimization techniques. Sought by government, industry, and academic organizations, he has lectured and consulted extensively around the world. He has received numerous awards, including the PRC Ministry of Education Changjiang Chair Professorship at Tsinghua University, the UK Royal Academy of Engineering senior visiting fellowship in Communications, the Croucher Foundation senior research fellowship, and the Order of the Bronze Bauhinia Star, Government of the Hong Kong Special Administrative Region, China. He is a registered professional engineer and a fellow of the Hong Kong Academy of Engineering Sciences, the IEEE, the IAE, and the HKIE. He is a fellow of the IEEE.