



## OPEN

## SUBJECT AREAS:

ENVIRONMENTAL  
SCIENCESMATHEMATICS AND  
COMPUTINGReceived  
3 October 2014Accepted  
30 January 2015Published  
3 March 2015Correspondence and  
requests for materials  
should be addressed to  
B.Z. (210010@csu.  
edu.cn)

# Performance comparison of LUR and OK in $PM_{2.5}$ concentration mapping: a multidimensional perspective

Bin Zou<sup>1,2</sup>, Yanqing Luo<sup>1</sup>, Neng Wan<sup>3</sup>, Zhong Zheng<sup>1</sup>, Troy Sternberg<sup>4</sup> & Yilan Liao<sup>5</sup>

<sup>1</sup>School of Geosciences and Info-Physics, Central South University, Changsha, 410083, China, <sup>2</sup>Shanghai Key Laboratory of Atmospheric Particle Pollution and Prevention (LAP<sup>3</sup>), Shanghai, 200433, China, <sup>3</sup>Department of Geography, University of Utah, Salt Lake City, UT 84112, USA, <sup>4</sup>School of Geography and Environment, Oxford University, Oxford, UK, <sup>5</sup>State Key Laboratory of Resources and Environmental Information System, Chinese Academy of Sciences, Beijing, 100001, China.

Methods of Land Use Regression (LUR) modeling and Ordinary Kriging (OK) interpolation have been widely used to offset the shortcomings of  $PM_{2.5}$  data observed at sparse monitoring sites. However, traditional point-based performance evaluation strategy for these methods remains stagnant, which could cause unreasonable mapping results. To address this challenge, this study employs ‘information entropy’, an area-based statistic, along with traditional point-based statistics (e.g. error rate, *RMSE*) to evaluate the performance of LUR model and OK interpolation in mapping  $PM_{2.5}$  concentrations in Houston from a multidimensional perspective. The point-based validation reveals significant differences between LUR and OK at different test sites despite the similar end-result accuracy (e.g. error rate 6.13% vs. 7.01%). Meanwhile, the area-based validation demonstrates that the  $PM_{2.5}$  concentrations simulated by the LUR model exhibits more detailed variations than those interpolated by the OK method (i.e. information entropy, 7.79 vs. 3.63). Results suggest that LUR modeling could better refine the spatial distribution scenario of  $PM_{2.5}$  concentrations compared to OK interpolation. The significance of this study primarily lies in promoting the integration of point- and area-based statistics for model performance evaluation in air pollution mapping.

Numerous studies have identified the negative impact of fine particulates ( $PM_{2.5}$ ) on respiratory health and human mortality<sup>1,2</sup>. However, understanding and monitoring harmful particulates  $PM_{2.5}$  has encountered several challenges so far, among which the most serious one is the insufficient  $PM_{2.5}$  data due to the expensive equipment and sparsely distributed field monitoring sites. This leads to difficulties in detecting the spatial characteristics and spatial-temporal dynamics of  $PM_{2.5}$  pollution and designing effective control strategy.

Several methods have been developed over the last decade to strengthen  $PM_{2.5}$  field monitoring which is critical in understanding global  $PM_{2.5}$  exposure. Efforts mainly include remote sensing image retrieval, air dispersion modeling, spatial interpolation, and land use regression (LUR) modeling<sup>3</sup>. However, investigation must explore advantages and shortcomings to determine the most effective approach for a specific situation.

While remote sensing techniques are able to retrieve particulate distribution over an image area-based on the relationship between aerosol optical depth (AOD) and  $PM_{2.5}$  concentration, the effectiveness is reduced when image acquisition phase-in fails. The limited spatial resolution (i.e. hundreds to thousands of meters) also makes it difficult to derive detailed  $PM_{2.5}$  spatial distribution characteristics in urban environments<sup>4</sup>. Similarly, air dispersion models can be used to simulate the  $PM_{2.5}$  concentration at preset receptors (i.e. grid points in this study) with various resolutions and coverage by using boundary layer turbulent diffusion theories and aerochemical theories. However, it requires copious data (e.g. emission, meteorological and terrain data) for hypothesis of the diffusion mode which makes it difficult to implement<sup>3,5–7</sup>.

Relatively, LUR modeling and Ordinary Kriging (OK) interpolation are two popular methods for mapping  $PM_{2.5}$  concentration based on the sparsely distributed observation data in diverse applications<sup>8–11</sup>. LUR modeling can produce  $PM_{2.5}$  concentration surfaces at fine resolutions by linking geographic elements with  $PM_{2.5}$  observation data using the least square method<sup>12</sup>. OK interpolation is suitable for  $PM_{2.5}$  concentration mapping based on the observation data with normal distribution and is the preferred unbiased geo-statistical technique in air pollution interpolation<sup>13</sup>. Unfortunately, implications of both LUR and OK methods are also limited by their inherent defects<sup>14</sup>. Shortcomings in unclear driving factors, non-standard predictor variable selection and poor time-space migration generally limit LUR model’s effectiveness<sup>9</sup>. OK interpolation usually fails to produce  $PM_{2.5}$



surface in the regions with sparse or missing data and is prone to over-amplify extreme variations due to its reliance on a single factor<sup>8</sup>. Consequently, accurate performance evaluation of LUR modeling and OK interpolation is particularly important for reliable air pollution mapping.

While studies have attempted to promote this work through comparing the performance of LUR, Kriging and air dispersion modeling in estimating  $PM_{10}$  concentration<sup>15</sup>, further improvements are still needed. Because model performance in these comparative studies was largely determined by similarities of causal mechanisms on air pollution concentrations between locations of test sample sites and training sample sites<sup>15–17</sup>. Model reliability is therefore dependent on test sites selected that are subject to evaluation errors<sup>18</sup>.

Information entropy, an area-based statistic indicator that was originally designed to describe the even spatial distribution of energy, has been increasingly used to evaluate the richness of image information. Since air quality concentration varies over space, information entropy has the potential of reflecting this variation based on the raster map of air quality concentrations<sup>19,20</sup>. Compared to traditional point-based statistics, information entropy is an effective index that can uniquely and objectively measure the information amount of a map and evaluate the capacity of this map in disclosing variation details of an element<sup>21,22</sup>.

This study therefore employed area-based information entropy along with traditional point-based statistics to evaluate the performance of LUR modeling and OK interpolation in mapping  $PM_{2.5}$  concentration in Houston from a multidimensional perspective. In order to better understand the meaning of information entropy values, an external profile analysis is also implemented.

As a large industrialized region in southeast Texas, the Houston metropolitan area covers 10 counties and 26,060 km<sup>2</sup> (Figure 1). In this study the city serves as a representative urban environment with documented high  $PM_{2.5}$  pollution rates. Prior works estimated a mean annual particulate concentrations in Houston that range from

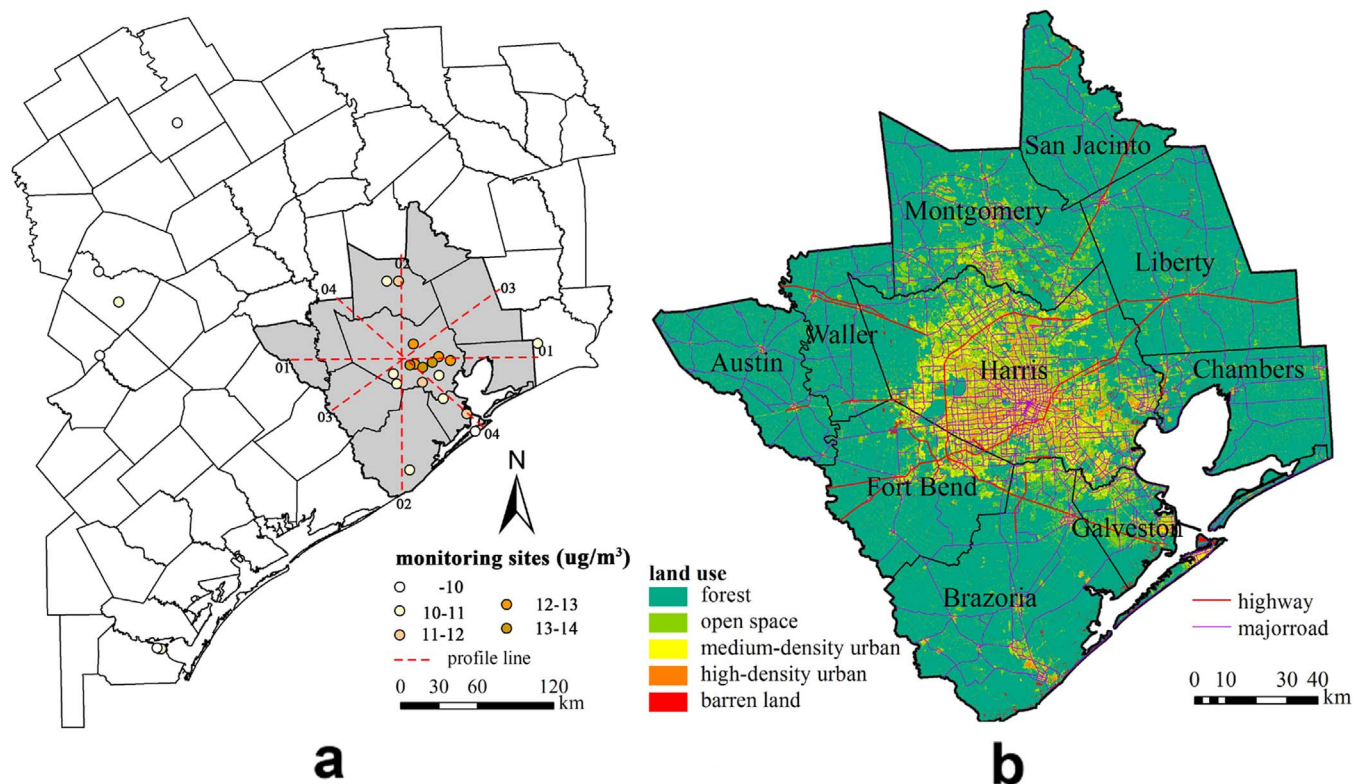
9.87  $\mu\text{g}/\text{m}^3$  (minimum) to 14.24  $\mu\text{g}/\text{m}^3$  (maximum) in the metropolitan area.

In the flat landscape, industrial and traffic emissions are the main pollutant sources in the multi-county area of 6 million residents in Houston metropolitan area according to U.S Environmental Protection Agency (EPA)<sup>23</sup>. Therefore, factors that contribute to Houston's  $PM_{2.5}$  pollution could be land-use type, road traffic, population distribution and geographic elements that represent location and climatic characteristics.

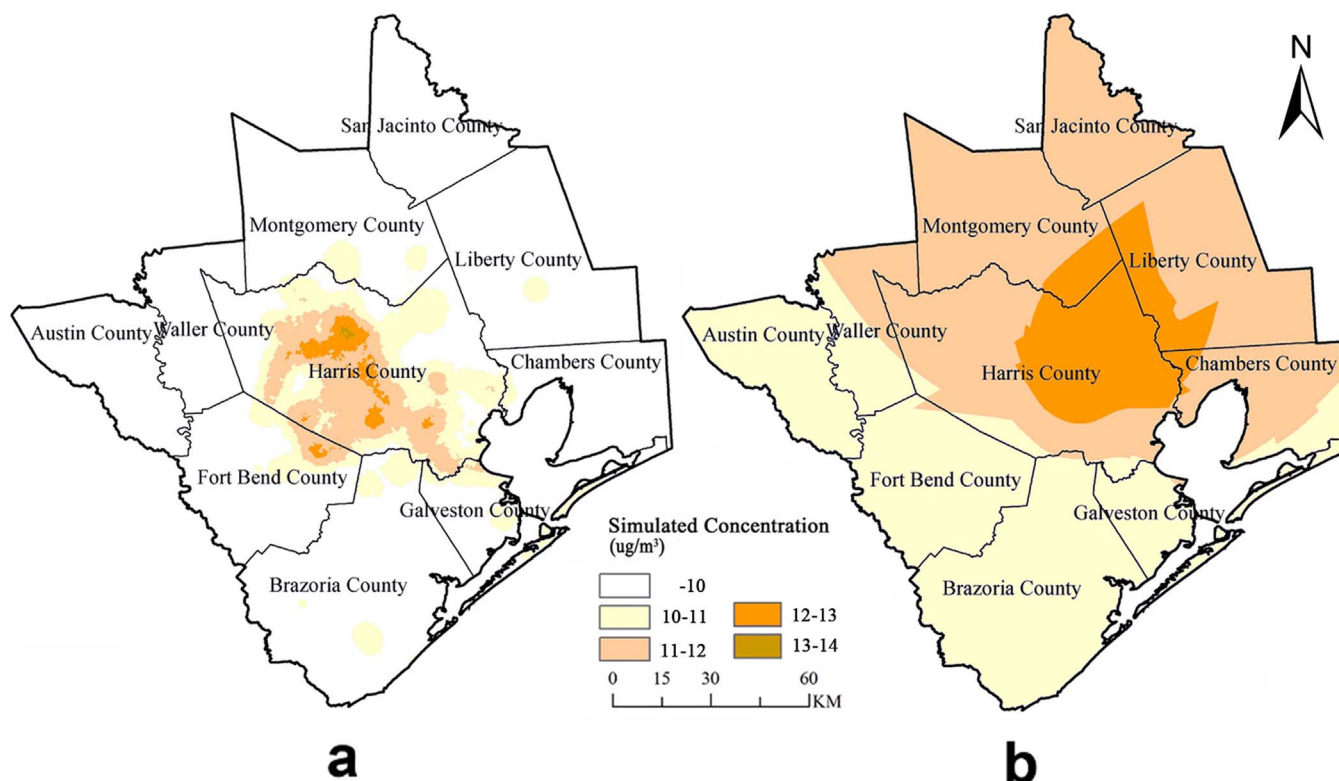
As a result, data used for LUR modeling in this study include the annual  $PM_{2.5}$  concentration at 17 monitoring sites (10 of them locate in Harris County) from the U.S. EPA's Air Quality System Technology Transfer network<sup>24</sup>. These  $PM_{2.5}$  concentrations are nearly distributed as normal fashion. Air quality monitoring on these sites complies with EPA's federal reference standard or federal equivalency standard, thus providing valid data for taking official air pollution measurements and quality assurance plans<sup>25</sup>. Land cover map with a spatial resolution of 30 m, road networks and demographic census data are respectively from the U.S. National Land Cover Database<sup>26</sup>, the Environmental Systems Research Institute (ESRI) nationwide street and geocoding databases<sup>27</sup>, and the U.S. Census database<sup>28</sup>.

## Results

**$PM_{2.5}$  concentration map.** Figure 2 shows the spatial distribution of annual  $PM_{2.5}$  concentrations in Houston metropolitan area produced by methods of LUR model and OK interpolation. Significant differences in  $PM_{2.5}$  concentrations can be observed across the covered counties from Figure 2 and confirmed by Figure 3. For the LUR model based map, high concentrations of simulated  $PM_{2.5}$  ( $>10 \mu\text{g}/\text{m}^3$ ) were found in urban Harris County whereas the surrounding suburban counties have lower concentrations ( $<10 \mu\text{g}/\text{m}^3$ ). In conjunction with Figure 3, this presents an obviously gradient of  $PM_{2.5}$  concentration in Houston from Harris to the surrounding areas. For the OK interpolation result map, the interpolated  $PM_{2.5}$  concentrations



**Figure 1** | A schematic map of study area created with the basic mapping function of ArcGIS (version 10.0): (a) monitoring sites distribution, (b) distribution of major geographical elements across Houston metropolitan.

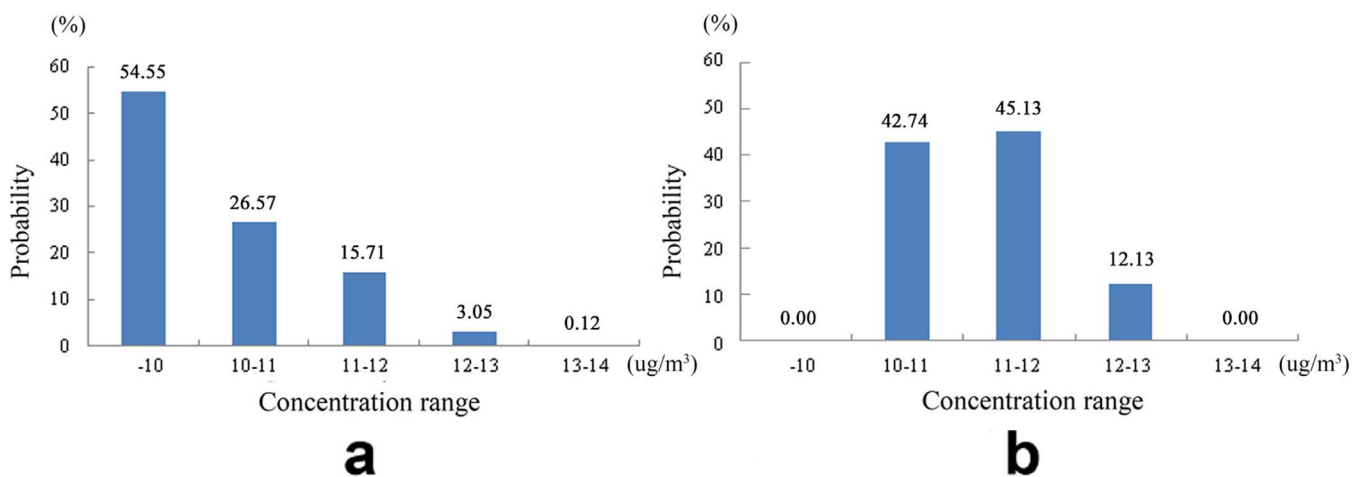


**Figure 2** | Spatial distribution map of PM<sub>2.5</sub> concentration in Houston metropolitan area produced by LUR model (a) and OK interpolation (b) with the spatial analysis and geostatistical function of ArcGIS (version 10.0).

reflected a clear zoning distribution with high levels of pollution (>10 µg/m<sup>3</sup>) in central and east Harris county followed by northeast and southwest Houston. Furthermore, comparison of Figure 2a/3a and Figure 2b/3b obviously demonstrates that the LUR simulated PM<sub>2.5</sub> concentrations showed a higher level of details and smoother variations than OK interpolated results.

**Performance comparison based on point-based statistics.** The performance of the LUR model and OK interpolation in PM<sub>2.5</sub> concentration estimation is evaluated by using the point-based statistics including absolute error, error rate, RMSE, and paired T-test. These statistics are calculated by using the typical N-1 cross validation strategy. Results listed in Table 1 show that the error rates of both LUR simulated- and OK interpolated PM<sub>2.5</sub> concen-

trations varied among monitoring sites. Whilst the absolute minimum and maximum errors of LUR simulated PM<sub>2.5</sub> concentrations were 0.02 µg/m<sup>3</sup> at site 9 and 2.04 µg/m<sup>3</sup> at site 15/16 with an average absolute error of 0.70 µg/m<sup>3</sup>, that of the OK interpolated PM<sub>2.5</sub> concentrations were 0.01 µg/m<sup>3</sup> at site 15 and 2.07 µg/m<sup>3</sup> at site 13 with an average absolute error of 0.80 µg/m<sup>3</sup>. Moreover, LUR model had an overall higher accuracy of simulated PM<sub>2.5</sub> concentrations compared to the OK interpolated ones although the paired T-test confirmed insignificant difference in site-based error rates between these two methods at P = 0.65. The LUR simulated- and OK interpolated PM<sub>2.5</sub> concentrations had respectively 5 and 7 sites with absolute error >1.00 µg/m<sup>3</sup>. The maximum error rates of these two methods were 15.95% and 20.34% with an average error rate of 6.13% and 7.01%, respectively. In addition, the RMSE evaluation



**Figure 3** | Statistic histograms of PM<sub>2.5</sub> concentrations illustrated in spatial distribution map from LUR model (a) and OK interpolation (b).



**Table 1 |** Point-based statistics of observed and simulated annual PM<sub>2.5</sub> concentrations. Paired *T* test is designed to test the significance of difference in error rates between LUR and OK in this table

Site ID	LUR				OK		
	Obser. (O) (µg/m <sup>3</sup> )	Simu. (S) (µg/m <sup>3</sup> )	Error (E) (µg/m <sup>3</sup> )	Error rate (E*) (%)	Simu. (S) (µg/m <sup>3</sup> )	Error (E) (µg/m <sup>3</sup> )	Error rate (E*) (%)
1	10.18	11.71	1.52	14.97	9.93	0.25	2.46
2	9.87	9.57	0.30	3.04	11.10	1.22	12.41
3	11.44	11.30	0.14	1.23	10.35	1.09	9.52
4	10.58	11.58	1.00	9.47	10.98	0.40	3.74
5	12.44	11.12	1.32	10.60	12.59	0.15	1.19
6	10.36	9.76	0.61	5.85	11.11	0.74	7.17
7	10.41	10.96	0.55	5.27	11.57	1.16	11.17
8	12.70	11.98	0.72	5.66	12.19	0.51	4.03
9	11.11	11.13	0.02	0.22	11.54	0.43	3.86
10	13.67	13.42	0.25	1.86	12.64	1.03	7.56
11	14.24	14.00	0.24	1.68	12.23	2.02	14.15
12	13.04	12.68	0.37	2.81	12.26	0.78	6.01
13	10.16	11.12	0.95	9.38	12.23	2.07	20.34
14	11.85	11.74	0.11	0.92	11.22	0.62	5.24
15	12.78	10.74	2.04	15.95	12.77	0.01	0.11
16	12.42	12.00	0.41	3.34	12.43	0.01	0.12
17	10.88	9.57	1.31	12.05	11.98	1.09	10.04
<b>Average</b>	<b>11.66</b>	<b>11.43</b>	<b>0.70</b>	<b>6.13</b>	<b>11.71</b>	<b>0.80</b>	<b>7.01</b>
<b>RMSE</b>			<b>0.89</b>			<b>1.00</b>	
P value of paired <i>T</i> test*					0.65		

results of LUR simulated and OK interpolated PM<sub>2.5</sub> concentrations (i.e. 0.89 and 1.00, respectively) were also consistent with those based on the absolute error and error rate.

#### Performance evaluation based on area-based information entropy.

Table 2 displays the values of information entropy and the related statistics calculated from the spatial distribution maps simulated by LUR model and interpolated by OK. The information entropy values (i.e. LUR: 7.79 vs. OK: 3.63) in Table 2 indicate that LUR model outperformed OK interpolation in illustrating detailed spatial variations of PM<sub>2.5</sub> concentrations across the Houston metropolitan area. The reliability of information entropy was echoed by the maximum, minimum and average PM<sub>2.5</sub> concentrations simultaneously shown in Table 2. Specifically, the LUR model generated PM<sub>2.5</sub> concentrations (9.57–13.52 µg/m<sup>3</sup>) were closer to the actual ground observation values (9.87–14.24 µg/m<sup>3</sup>) than did by OK interpolation (which ranges from 10.08–12.93 µg/m<sup>3</sup>). Additionally, profile analysis results in Figure 4 further confirmed above findings of information entropy evaluation. It can be observed that, along all four directions, the PM<sub>2.5</sub> concentrations interpolated by the OK method almost first demonstrated an increasing trend and then gradually decreased, while those simulated by LUR model at the same local sites were relatively lower and stable at two ends but higher and fluctuated in the middle. This difference suggests that the spatial distribution scenario of PM<sub>2.5</sub> concentrations could be better refined by LUR modeling rather than by OK interpolation.

#### Discussion

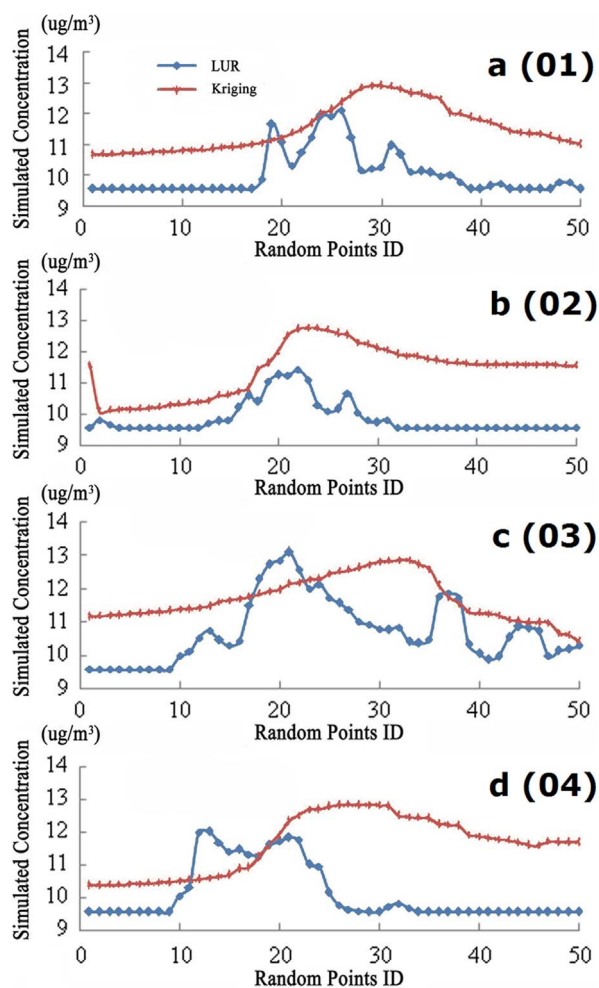
This study explored the differences in spatial distributions of PM<sub>2.5</sub> concentrations between LUR model and OK interpolation by com-

prehensively using point-based statistics and area-based information entropy for the first time. We found that, based on point-based statistics, the two methods produce similar results. However, highlighted significant differences were observed between the two methods based on area level information entropy and confirmed the better performance of LUR relative to OK. Our findings provide new insights for future air pollution research.

The optimal adjusted LUR model in this study has a fitting  $R^2$  of 0.69, which is much higher than that of the OK method ( $R^2 = 0.38$ ) as well as the results of previous studies (e.g. London, 0.45 to 0.60<sup>29</sup>; 0.56, 0.73 and 0.50, northern Europe<sup>30</sup>; Germany, 0.17<sup>31</sup>). This study applied backward Multiple Linear Regression (MLR) method<sup>30,32,33</sup> to achieve the best LUR model fitting. Due to the limited number of PM<sub>2.5</sub> monitoring sites in the Houston metropolitan area, this study utilized empirical LUR variable values and sampling-site numbers to screen individual modeling variables<sup>34–36</sup> and the strategy widely used in previous studies<sup>37,38</sup>. Variables of land use type and road traffic with strong prediction capacity are screened first. Population distribution and variables about distance to sea are then incorporated for model adjustments. Because Houston's PM<sub>2.5</sub> pollution is primarily from diesel emission, oil vehicles, road dust, barbecue, and wood burning<sup>23</sup>, Harris County which is highly urbanized and industrialized experiences relatively higher PM<sub>2.5</sub> concentration, while surrounding areas which are characterized by agricultural land use and fewer road networks have relatively lower PM<sub>2.5</sub> concentrations. This is reflected by the LUR model simulated result, which shows a decreasing trend from Harris County to surrounding areas. It also confirms that the simulation result of the LUR model is closer to the real PM<sub>2.5</sub> spatial distribution compared to that of the OK interpolation as shown by the statistics in Table 1, while the PM<sub>2.5</sub> annual concentration of OK

**Table 2 |** Area-based statistics of the spatial distribution maps of PM<sub>2.5</sub> concentration

	Max value (µg/m <sup>3</sup> )	Min value (µg/m <sup>3</sup> )	Ave. value (µg/m <sup>3</sup> )	Information Entropy
OK interpolated	12.93	10.08	11.21	3.63
LUR model Simulated	13.52	9.57	9.86	7.79
Observed	14.24	9.87	11.65	-



**Figure 4** | Variations of  $PM_{2.5}$  concentrations produced by LUR model and OK interpolation at four direction profiles in Houston metropolitan area: east-west a (01), south-north b (02), southeast-northwest c (03) and southwest-northeast d (04).

interpolation was zonally distributed in Houston. And, high concentration areas include the central eastern region of Harris County and the northeast and southwest regions of Houston.

The point-based statistics validation demonstrated no significant differences between the results from LUR model and OK interpolation. However, the LUR model achieved slightly better simulation accuracy than the OK interpolation (e.g. RMSE: 0.89 vs. 1.00). Given the fact that the quality of OK interpolation is dependent on the distribution of monitoring sites, the validation precision of OK interpolated  $PM_{2.5}$  concentrations at different monitoring sites would certainly vary, with poorest results being at the boundary area due to the insufficient observation data. Inversely, considering more factors such as land use, traffic, population, we believe LUR model is more reliable than OK interpolation, especially for the area without abundant observed  $PM_{2.5}$  concentrations but sufficient relevant auxiliary factors. Additionally, the point-based statistics validation process of LUR model and OK interpolation in this study is based on the typical N-1 cross validation strategy and should be the ‘best’ one we can use with discrete monitoring sites.

Furthermore, area-based information entropy evaluation revealed significantly different results between the LUR model and OK interpolation. The annual  $PM_{2.5}$  concentrations simulated by the LUR model have more spatial variations (greater information entropy values) than the OK interpolation. This is because LUR model integrates additional influencing factors that are closely related with the

emission and diffusion of  $PM_{2.5}$ , such as land use, road traffic, and climatic indicators. These factors strengthened the ability of LUR model in revealing the concentration variations through distinguishing the surrounding geographic differences of divergent positions, especially for areas with limited monitoring sites. These two advantages increased the information richness of the LUR simulated map and reflected the real-world scenario of  $PM_{2.5}$  concentration. These factors re-confirmed the superiority of information entropy in evaluating an air quality map’s capacity in disclosing variations, which could not be achieved by previous air quality mapping studies based on traditional point-based statistics<sup>7,18,39</sup>. According to the urban development pattern (e.g. Harris County is with high volume of traffic and is also the industrial and economic center) and  $PM_{2.5}$  sources in Houston, we believe the evaluation result based on LUR model is more reliable than OK interpolation.

Like previously reported studies with data from few monitoring sites (i.e. minimum site number is 13)<sup>40,41</sup>, while satisfactory results have been achieved with the data collected from 17 monitoring sites in this study, issues on monitoring sites and the predictor selection still need to be addressed in the future.  $PM_{2.5}$  concentration estimation with higher accuracy could be achieved with more monitoring sites. Moreover, while this study established a significant LUR model at an acceptable accuracy level using MLR without overestimates, the model’s performance definitely could be further enhanced by involving more predictors under sufficient monitoring sites that are evenly distributed in space.

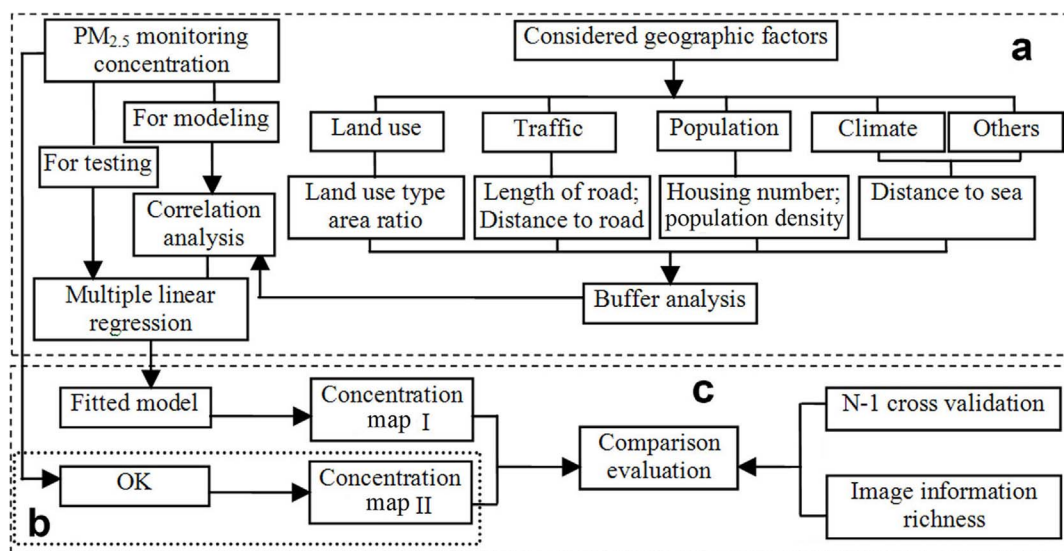
In summary, findings in this study imply that although the point-based statistics evaluation could accurately reflect a model’s performance in mapping air pollution concentration, its evaluation result is often limited by test site locations and their spatial distribution. In regions with densely centralized test sites and training sites, point-based statistics evaluation methods may overestimate the model accuracy (i.e. better or worse accuracy), and vice versa. Therefore, except for point-based statistics evaluation, the area-based information entropy evaluation proposed in this study is important and necessary for more comprehensive and accurate assessment of the air pollution concentration maps. In other words, the information entropy evaluation clearly confirms that LUR model is more accurate in representing the spatial distribution of annual  $PM_{2.5}$  concentrations of Houston metropolitan area than the OK interpolation in this study. Additionally, this study implies that the utilization of information entropy is a new measure to effectively evaluate the performance of other exposure models such as dispersion modeling, LUR modeling, and remote sensing based models, for which the spatial resolution is better than OK interpolation. And this could greatly enhance the reliability of findings for future environmental health studies.

## Methods

The methodology of this study is composed of three parts: LUR modeling, OK interpolation, and performance comparison between LUR and OK (Figure 5).

**LUR modeling.** LUR modeling links the air pollution concentration at a monitoring site with other geographic characteristics of that monitoring site. The modeling is composed of variable extraction and screening, regression model building, and model validation. The variable extraction and screening include selecting geographic elements and extracting characteristic variables of geographic elements.

Considering experiences from previous LUR studies<sup>10,11,30,32,33,42–44</sup> and  $PM_{2.5}$  pollution sources in Houston, this study utilizes annual  $PM_{2.5}$  concentrations as the outcome variable and develops predictors of various geographic elements including land use type ( $X_1$ ), road length ( $X_2$ ), distance to road ( $X_2$ ), population density ( $X_{31}$ ), house density ( $X_{32}$ ), and distance to sea ( $X_{41}$ ). Among them, the “measured values” of predictors with spatial scaling effect are extracted at 100 m, 300 m, 500 m, 800 m, 1000 m, 1500 m, 2000 m, 2500 m, 3000 m, 3500 m, 4000 m, 4500 m and 5000 m buffering radius due to the unclear ‘spatial scale dependency’<sup>33,42,45</sup>. Land use types are reclassified as forest ( $X_{11}$ ), open space ( $X_{12}$ ), medium-density urban ( $X_{13}$ ), high-density urban ( $X_{14}$ ) and barren land ( $X_{15}$ ) with the 11 initial land use types provided by United States Geological Survey. Road traffic data in this study includes highway ( $X_{21}$ ), major road ( $X_{22}$ ), local road ( $X_{23}$ ), minor road ( $X_{24}$ ) and other road ( $X_{25}$ ). The entire process is implemented with ArcGIS 10.0.



**Figure 5** | Framework of study procedure including LUR modeling (a), OK interpolation (b), and performance comparison between LUR and OK (c).

To screen out effective predictors appropriate for LUR modeling in Houston, Pearson coefficient values between all predictors and annual PM<sub>2.5</sub> concentration are calculated with SPSS 19.0. For predictors with spatial scaling effect, the optimal spatial scale of each predictor is defined as the one with calculated maximum Pearson coefficient in a scale range of 100 m to 5000 m. Consequently, the final predictors screened out for LUR modeling in this study are area fraction of land use type including  $X_{11}$ -5000,  $X_{12}$ -100,  $X_{13}$ -100,  $X_{14}$ -800,  $X_{15}$ -3000; road traffic including  $X_{22}$ -100 m,  $X_{23}$ -300 m,  $X_{24}$ -3000 m,  $X_{25}$ -1500 m and  $X_{26}$ ; and others including  $X_{31}$ -3000 m,  $X_{32}$ -1000 m,  $X_{41}$ .

A predictor-based regression model is established by using a multiple linear regression (MLR) equation (i.e. Equation 1) in this study. The equation is shown as

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \dots + a_nX_n + u \quad (1)$$

where  $Y$  is the annual PM<sub>2.5</sub> concentration,  $X$  denotes independent predictors,  $a_0$  is a constant,  $a_1$  to  $a_n$  are the regression coefficients for each predictor  $X_i$ , respectively, and  $u$  is the random error. An equation group is composed of  $n$  groups of observed values  $Y_i, X_{1i}, X_{2i}, X_{3i}, \dots, i = 1, 2, 3, \dots, n$ .

Area fraction of land use type and road traffic are the two major influencing factors of annual PM<sub>2.5</sub> concentration in the Houston area. Thus, this study starts with backward MLR by using the respective type of predictors under an optimal spatial scale as inputs to establish the preliminary optimal models (i.e. model with highest fitting  $R^2$ ) with SPSS 19.0. Thereafter, another round of backward MLRs is conducted for these preliminary optimal models by adding predictors such as population density, house density and distance to sea. As a result, the finalized LUR model is built as  $Y_{\text{Conc}} = X_{13-100} + X_{31-3000} + 8.357$  with significant coefficients at  $p < 0.05$ . The adjusted  $R^2$  of this finalized model is 0.69 with VIF values less than 10 to ensure non-multicollinearity.

Using the finalized LUR model, a continuous surface of annual PM<sub>2.5</sub> concentrations at the resolution of 3 km  $\times$  3 km (Figure 2) within the study area is generated with ArcGIS 10.0 taking into account the point-based high computational cost and spatial similarity of predictors within certain spatial scales (i.e. buffering area size). Specifically, grid points with the 3 km interval across the entire study area are pre-set firstly; then the 'measured values' of predictors in the finalized LUR model at these pre-set points are extracted and used to calculate the annual PM<sub>2.5</sub> concentrations at each pre-set point; these high density estimated PM<sub>2.5</sub> concentrations are used to produce the distribution map of annual PM<sub>2.5</sub> concentrations in the end.

**OK interpolation.** OK interpolation refers to the linear unbiased optimal estimation of unknown points according to the structural features of known sample points<sup>46</sup>. When the regional variable  $Z(x)$  is a constant (m) with unknown mathematical expectations, the OK method can be used for spatial interpolation. The interpolation formula is stated as

$$Z^*(x_0) = \sum_{i=1}^n \omega_i Z(x_i) \quad (2)$$

where  $Z(x_0)$  is the value of an unknown sample point,  $Z(x_i)$  is the value of a known sample point surrounding the unknown sample point, and  $Z^*(x_i)$  is the unbiased estimation of  $Z(x_i)$  (i.e.  $E[Z^*(x_0) - Z(x_0)] = 0$ ).  $\omega_i$  is the weight of the  $i^{\text{th}}$  known sample point to the unknown sample point and  $\sum_{i=1}^n \omega_i = 1$  where  $n$  is the amount of known sample points.

For the process of OK interpolation, an exploratory data analysis is firstly conducted on the training sample data of 17 monitoring sites within the Houston metropolitan area and external 4 expanding sites outside Houston metropolitan area to determine whether the data follow a normal distribution or are spatially correlated or not. Then, a continuous prediction map of annual PM<sub>2.5</sub> concentration is produced using the 'Spatial Interpolation' wizard of ArcGIS10.0. We did not use trend removal because of the relatively smooth variation of PM<sub>2.5</sub> concentration across these monitoring sites (i.e. 17 + 4). Considering the sparse distribution of monitoring sites in the study area, the searching number of neighborhood points is set as 4.

**Point-based statistics calculation.** Point-based statistics including absolute error, relative error and root-mean-square error ( $RMSE$ ) are employed to validate methods of LUR model and OK interpolation in this study by using the commonly N-1 cross validation strategy, which is suitable for limited data samples<sup>12,15</sup>. Following this, this study divides the 17 monitoring sites across the study area into 16 training sites and 1 validation site. The absolute error represents the deviation direction and size of the simulated/estimated concentration from the observed concentration. Relative error and  $RMSE$  represent the deviation degree of the simulated/estimated concentration from the observed concentration, which reflect the reliability of the estimation result of the model. The three error indices are calculated according to equations (3)–(5) with larger values indicating lower model accuracy.

$$E = |O - S| \quad (3)$$

$$E^* = E / \times 100\% \quad (4)$$

$$RMSE = \sqrt{\{(O_1 - S_1)^2 + (O_2 - S_2)^2 + \dots + (O_n - S_n)^2\} / n} \quad (5)$$

where  $E$ ,  $E^*$ , and  $RMSE$  respectively represent the absolute difference, relative error (i.e. error rate) and the root-mean-square error between observed concentration and estimated concentration.  $O$  is observed concentration,  $S$  is simulated/estimated concentration and  $n$  is sample size.

**Area-based information entropy.** "Entropy" is an indicator that was originally designed to describe the even spatial distribution of energy<sup>19</sup>. It has been expanded to indicate the richness of information in information theories. Given  $v = \{X_1, X_2, \dots, X_n\}$ , suppose the probability of  $X_i \in v$  is  $\rho_i = P(X_i)$ , the information entropy of  $v$  can be defined as:

$$E(v) = - \sum_{i=1}^n P(X_i) \log_2(P(X_i)) \quad (6)$$

where  $X_i$  represents the pixel of an image and  $P(X_i)$  is the probability of occurrence of  $X_i$ . The more heterogeneous  $X_i$  is, the larger the information entropy of the image will be, indicating more details of the spatial pattern.

In this study, information entropy is developed to depict the ability of LUR model and OK interpolation methods in mapping the variation of the annual PM<sub>2.5</sub> concentrations over the entire study area. Specifically, the distribution maps of annual PM<sub>2.5</sub> concentration are firstly produced and reclassified with natural break points. Then, the number of raster grids at each class are summed and divided by the total grid number of the raster map to calculate the probabilities  $P(X_i)$ . These probabilities are finally used to compute the value of information entropy according to equation (6). The calculations of information entropy for raster maps from LUR model and OK



interpolation are similar and both are implemented with the modules of spatial interpolation and algebraic computation in ArcGIS 10.0.

Additionally, a four-direction criterion (Figure 1), namely, east-west (01), south-north (02), southeast-northwest (03) and southwest-northeast (04) are employed to further confirm the necessity of area-based information evaluation considering factors that possibly caused the heterogeneity of PM<sub>2.5</sub> concentrations. In each direction, PM<sub>2.5</sub> concentrations at 50 randomly distributed sites are separately simulated and interpolated by LUR and OK methods.

- Kaiser, J. How dirty air hurts the heart. *Science* **307**, 1858–1859 (2005).
- Silva, R. A. *et al.* Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change. *Environ. Res. Lett.* **8**, 034005 (2013).
- Zou, B., Wilson, J. G., Zhan, F. B. & Zeng, Y. Air pollution exposure assessment methods utilized in epidemiological studies. *J. Environ. Monit.* **11**, 475–490 (2009).
- Levy, R. C., Remer, L. A., Mattoo, S., Vermote, E. F. & Kaufman, Y. J. Second-generation operational algorithm: Retrieval of aerosol properties over land from inversion of Moderate Resolution Imaging Spectroradiometer spectral reflectance. *J. Geophys. Res.* **112** (2007).
- Bellander, T. *et al.* Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in Stockholm. *Environ. Health Perspect.* **109**, 633 (2001).
- Wilson, J. G., Kingham, S., Pearce, J. & Sturman, A. P. A review of intraurban variations in particulate air pollution: Implications for epidemiological research. *Atmos. Environ.* **39**, 6444–6462 (2005).
- Zou, B., Zhan, F. B., Zeng, Y., Yorke, C. & Liu, X. Performance of Kriging and EWPM for Relative Air Pollution Exposure Risk Assessment. *Int. J. Environ. Res.* **5**, 769–778 (2011).
- Wilson, J. G. & Zawar-Reza, P. Intraurban-scale dispersion modelling of particulate matter concentrations: applications for exposure estimates in cohort studies. *Atmos. Environ.* **40**, 1053–1063 (2006).
- Hoek, G. *et al.* A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **42**, 7561–7578 (2008).
- Beelen, R. *et al.* Development of NO<sub>2</sub> and NO<sub>x</sub> land use regression models for estimating air pollution exposure in 36 study areas in Europe—The ESCAPE project. *Atmos. Environ.* **72**, 10–23 (2013).
- Wang, M. *et al.* Evaluation of land use regression models for NO<sub>2</sub> and particulate matter in 20 European study areas: the ESCAPE project. *Environ. Sci. Technol.* **47**, 4357–4364 (2013).
- Gilliland, F. *et al.* Air pollution exposure assessment for epidemiologic studies of pregnant women and children: lessons learned from the Centers for Children's Environmental Health and Disease Prevention Research. *Environ. Health Perspect.* **113**, 1447–1454 (2005).
- Jerrett, M. *et al.* A review and evaluation of intraurban air pollution exposure models. *J. Exposure Sci. Environ. Epidemiol.* **15**, 185–204 (2005).
- Sellier, Y. *et al.* Health effects of ambient air pollution: Do different methods for estimating exposure lead to different results? *Environ. Int.* **66**, 165–173 (2014).
- Gulliver, J., de Hoogh, K., Fecht, D., Vienneau, D. & Briggs, D. Comparative assessment of GIS-based methods and metrics for estimating long-term exposures to air pollution. *Atmos. Environ.* **45**, 7072–7080 (2011).
- Marshall, J. D., Nethery, E. & Brauer, M. Within-urban variability in ambient air pollution: comparison of estimation methods. *Atmos. Environ.* **42**, 1359–1369 (2008).
- Mercer, L. D. *et al.* Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NO<sub>x</sub>) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Atmos. Environ.* **45**, 4412–4420 (2011).
- Zou, B., Zhan, F. B., Wilson, J. G. & Zeng, Y. Performance of AERMOD at different time scales. *Simul. Modell. Pract. Theory* **18**, 612–623 (2010).
- Clausius, R. Über die bewegende Kraft der Wärme und die Gesetze, welche sich daraus für die Wärmelehre selbst ableiten lassen. *Ann. Phys.* **155**, 368–397 (1850).
- Tsai, D.-Y., Lee, Y. & Matsuyama, E. Information entropy measure for evaluation of image quality. *J. Digit. Imaging* **21**, 338–347 (2008).
- Ferraro, M., Boccignone, G. & Caelli, T. Entropy-based representation of image information. *Pattern Recognit. Lett.* **23**, 1391–1398 (2002).
- Wang, Y., Chen, Q. & Zhang, B. Image enhancement based on equal area dualistic sub-image histogram equalization method. *IEEE Trans. Consum. Electron.* **45**, 68–75 (1999).
- Fraser, M. P., Yue, Z. W. & Buzcu, B. Source apportionment of fine particulate matter in Houston, TX, using organic molecular markers. *Atmos. Environ.* **37**, 2117–2123 (2003).
- EPA, U. S. *Air Quality System*. Available at: <http://www.epa.gov/ttn/airs/airsaqs>. (Accessed: 17th November 2013).
- EPA, U. S. *Manuals and Guides*. Available at: <http://www.epa.gov/ttn/airs/airsaqs/manuals/>. (Accessed: 4th December 2013).
- EPA, U. S. *2001 National Land Cover Data (NLCD 2001)*. Available at: <http://www.epa.gov/mlrc/nlcd-2001.html>. (Accessed: 25th December 2010).
- Esri. *Esri Business Analyst Desktop*. Available at: <http://www.esri.com/software/arcgis/extensions/businessanalyst>. (Accessed: 29th November 2013).
- Census, U. S. *Census 2000 Gateway*. Available at: <http://www.census.gov/main/www/cen2000.html>. (Accessed: 18th November 2010).
- Briggs, D. J. *et al.* Mapping urban air pollution using GIS: a regression-based approach. *Int. J. Geogr. Inf. Sci.* **11**, 699–718 (1997).
- Brauer, M. *et al.* Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. *Epidemiology* **14**, 228–239 (2003).
- Hochadel, M. *et al.* Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information. *Atmos. Environ.* **40**, 542–553 (2006).
- Ross, Z., Jerrett, M., Ito, K., Tempalski, B. & Thurston, G. D. A land use regression for predicting fine particulate matter concentrations in the New York City region. *Atmos. Environ.* **41**, 2255–2269 (2007).
- Mao, L., Qiu, Y., Kusano, C. & Xu, X. Predicting regional space–time variation of PM<sub>2.5</sub> with land-use regression model and MODIS data. *Environ. Sci. Pollut. Res. Int.* **19**, 128–138 (2012).
- Green, S. B. How many subjects does it take to do a regression analysis. *Multivar. Behav. Res.* **26**, 499–510 (1991).
- Mulaik, S. A. The curve-fitting problem: An objectivist view. *Philos. Sci.* **68**, 218–241 (2001).
- Babyak, M. A. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom. Med.* **66**, 411–421 (2004).
- Liu, Y., Franklin, M., Kahn, R. & Koutrakis, P. Using aerosol optical thickness to predict ground-level PM<sub>2.5</sub> concentrations in the St. Louis area: A comparison between MISR and MODIS. *Remote Sens. Environ.* **107**, 33–44 (2007).
- Hu, X. *et al.* Estimating ground-level PM<sub>2.5</sub> concentrations in the southeastern US using geographically weighted regression. *Environ. Res.* **121**, 1–10 (2013).
- Zou, B., Wilson, J. G., Zhan, F. B., Zeng, Y. & Wu, K. Spatial-temporal variations in regional ambient sulfur dioxide concentration and source-contribution analysis: A dispersion modeling approach. *Atmos. Environ.* **45**, 4977–4985 (2011).
- Henderson, S. B., Beckerman, B., Jerrett, M. & Brauer, M. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environ. Sci. Technol.* **41**, 2422–2428 (2007).
- Chen, C., Wu, C., Yu, H., Chan, C. & Cheng, T. Spatiotemporal modeling with temporal-invariant variogram subgroups to estimate fine particulate matter PM<sub>2.5</sub> concentrations. *Atmos. Environ.* **54**, 1–8 (2012).
- Eeftens, M. *et al.* Development of land use regression models for PM<sub>2.5</sub>, PM<sub>2.5-10</sub> absorbance, PM<sub>10</sub> and PM<sub>coarse</sub> in 20 European study areas; results of the ESCAPE project. *Environ. Sci. Technol.* **46**, 11195–11205 (2012).
- De Hoogh, K. *et al.* Development of land use regression models for particle composition in twenty study areas in Europe. *Environ. Sci. Technol.* **47**, 5778–5786 (2013).
- Wang, M. *et al.* Performance of Multi-City Land Use Regression Models for Nitrogen Dioxide and Fine Particles. *Environ. Health Perspect.* **122**, 843–849 (2014).
- Zou, B., Peng, F., Wan, N., Wilson, J. G. & Xiong, Y. Sulfur dioxide exposure and environmental justice: a multi-scale and source-specific perspective. *Atmos. Pollut. Res.* **5**, 491–499 (2014).
- Journel, A. G. & Huijbregts, C. J. *Mining geostatistics*. **600** (Academic Press, London, 1978).

## Acknowledgments

This work is funded by the National Natural Science Foundation of China (No. 41201384, 41471377, 41101431), the Opening Project of Shanghai Key Laboratory of Atmospheric Particle Pollution and Prevention (LAP<sup>3</sup>), and the Hunan Provincial Natural Science Foundation of China (No. 12JJ3034). Bin Zou would also like to thank grants from the Key Laboratory of Geo-informatics of State Bureau of Surveying and Mapping (No. 201228, 2014J07) and State Key Laboratory of Resources and Environmental Information System, as well as the NieYing Talent Program of Central South University.

## Author contributions

B.Z. designed and performed the majority of experiments and data analysis, as well as coordinated and wrote the manuscript. Y.Q.L. participated in experimental designs and data analysis, and drafted the manuscript text. N.W., T.S., Z.Z. and Y.L.L. contributed to writing the manuscript. All authors reviewed and approved the manuscript.

## Additional information

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Zou, B. *et al.* Performance comparison of LUR and OK in PM<sub>2.5</sub> concentration mapping: a multidimensional perspective. *Sci. Rep.* **5**, 8698; DOI:10.1038/srep08698 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>