Published in final edited form as:

Ann GIS. 2019 ; 25(1): 1–8. doi:10.1080/19475683.2018.1555188.

Estimation of Environmental Exposure: Interpolation, Kernel Density Estimation, or Snapshotting

Xun Shi[a], **Meifang Li**[b,a], **Olivia Hunter**[c], **Bart Guetti**[a], **Angeline Andrew**[d], **Elijah Stommel**[e], **Walter Bradley**[f], and **Margaret Karagas**[d]

[a]Department of Geography, Dartmouth College, Hanover, NH 03755, USA

[b]School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China

[c]Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA

[d]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, NH 03755, USA

[e]Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756, USA

[f]Department of Neurology, Miller School of Medicine, University of Miami, Miami, FL 33136, USA

Abstract

In environmental health researches and practices, spatial analysis became an important approach to estimation of environmental exposure of human subjects under concern. A typical situation in this kind of application is that the data of pollution are available only at certain locations, and thus inference is needed to convert a limited number of values at discrete locations into a continuous surface. This paper intends to clarify the distinction among three methods that can be used to achieve this conversion, namely interpolation, kernel density estimation (KDE), and *snapshotting*. Due to the apparent similarity of the three, they may cause confusions that lead to misuses. We compare and contrast the three methods, in terms of nature of the input data, mathematical process of the inference, and essential meaning of the output. For each method we suggest appropriate applications within the context of estimation of environmental exposure.

Keywords

Environmental exposure; interpolation; kernel density estimation; snapshotting; health

Introduction

In environmental health researches and practices, spatial analysis became an important approach to estimation of environmental exposure of human subjects under concern. The underlying assumption of this approach is that the concentration of a certain pollutant at a human subject's location (e.g., residential location) can be taken as a substitute of the subject's actual exposure to that pollutant in an epidemiological analysis. A typical situation that the researcher deals with in this kind of application is that the data of pollution are available only at particular locations, and thus a certain inference is needed to estimate concentrations at human subjects' locations. Such an inference usually appears to be a process of converting a limited number of values at discrete locations into a continuous surface. Commonly employed methods to achieve such a conversion include interpolation and kernel density estimation (KDE).

Interpolation has been widely used in environmental health studies to map contaminations of soil (e.g., Carlon et al. 2001; Hooker and Nathanail 2006; Aelion et al. 2013) and sediments (e.g., Katz et al. 2013), model chemical concentrations in groundwater (e.g., Goovaerts et al. 2005; Ayotte et al. 2006), and quantify the degree of local air pollution (e.g., Yuan et al. 2015; Borge et al. 2016; Martens et al. 2017). On the other hand, studies that employ KDE are often under topics of material flow, environmental fate, and pollution emission (e.g., Jerrett et al. 2005; Tonne et al. 2006; Hu, Liebens and Rao 2008; Gottschalk, Scholz and Nowack 2010), and access to food outlets and recreational facilities (Kestens et al. 2010; Thornton, Pearce and Kavanagh 2011). In our own studies, we realized that when the available data are values measured right at the sources for selected time points, neither interpolation nor KDE is directly applicable. We are not aware of works in the literature that explicitly address this latter situation. We developed a process and associated software tool to perform inference based on such data, and named the process *snapshotting*.

The similarity that all three can generate continuous surfaces in the raster format from discrete representations, typically points, may cause confusion and lead to misuses. This paper intends to clarify the distinction among the three methods, without overusing statistical concepts and terminology, and suggest appropriate applications for each.

For the convenience of description, we only discuss the situations that the input data are in the format of points. In real-world exposure estimations based on either of the three analyses, lines and polygons are usually first to be discretized into points, and thus the discussion here can be directly applied to the situations with lines and polygons. Also for the convenience of description, in many places of this paper we imply negative physical environmental factors by using words such as contamination and pollution, but the same idea and process are certainly apply to other types of environmental factors, including physical, socioeconomic, and behavioural attributes, and many of them can be positive to human health, e.g., infrastructure that improves walkability, healthy food outlets, recreational facilities, and greenspaces.

Interpolation

In this paper, to simplify the description we use *interpolation* to refer to *spatial interpolation*. Interpolation is based on *samples*. Samples means that they are selected (sampled) from a much larger set of values (i.e., the value at each and every location in the study area). The reason for sampling is that we are not able to obtain and/or handle values of all locations in the study area, and thus we intend to infer information about those unselected values (i.e., values at those un-sampled locations) based on the samples. This intention requires samples to be *representative*. Interpolation is one way to infer values at un-sampled locations based on values at sampled locations. Basically, interpolation is a weighted average calculation:

$$v_o = \sum_{i=1}^{n} w_i v_i \quad (1)$$

where v_o is the value at a given un-sampled location; v_i is the value of sample *i*; w_i is the weight of sample *i*; and *n* is the total number of samples. The weight w_i is greatly related to the distance between locations *i* and *o*. For example, in a simple inverse distance weighting (IDW) method, w_i is entirely determined by that distance; in a sophisticated kriging process, w_i is determined by a sample-driven quantification of spatial autocorrelation, which takes in both spatial and attribute information. Since it is a weighted average calculation, it is constrained by:

$$\sum_{i=1}^{n} w_i = 1 \quad (2)$$

This constraint indicates that the weight of a sample point is not determined by itself, but by its relationships with other samples. The result of a simple IDW interpolation that employs a linear distance decay function can be illustrative about its weighted-averaging nature (Figure 1): the resulting continuous surface passes through the samples, i.e., the sample values are on the surface, and an interpolated value is intermediate between the sample values it is inferred from.

The result of a kriging interpolation may not have the original sample values right on the interpolated surface. This is because it uses a derived global model to determine the weight of each sample. Nevertheless, the goal of the derivation is to make the model most optimally fit all the samples (or certain derived characteristics of them, in this case, the semivariogram). This is like that a derived regression line may not pass through all sample values, but the goal is still to best fit the line to the samples.

When estimating the environmental exposure, interpolation should be used when the available data are indeed from a limited number of sample locations. The term *sample locations* means that these locations are selected from a much larger number of locations – theoretically including each and every location in the study area – where measurements can be obtained. Typical examples include PM2.5 measurements from air quality monitoring stations, heavy metal concentrations in soils from sample locations in a farm, and cyanobacteria concentrations in water from sample locations in a lake, and arsenic concentrations measured with sampled water from wells.

Kernel Density Estimation

In non-parametric statistics, the observed points in the kernel density estimation (KDE) are considered realizations of a *probability surface* (Silverman 1986), and KDE is a process of inferring the probability density at each and every location, i.e., constructing the probability density surface, based on the locations of realizations (the points). This notion, in a two-dimensional space, can be formally represented as (adapted from Shi 2010):

$$\hat{p}(o) = \frac{1}{n\pi h^2}\sum_{i=1}^{n} K\left(\frac{d_{i,o}}{h}\right) \quad (3)$$

where $\hat{p}(o)$ is the estimated probability density at a given location o; n is the total number of points under concern; h is the kernel bandwidth; $d_{i,o}$ is the distance between point i and location o; and K is a kernel function characterizing how the relevance of point i varies as a function of $d_{i,o}$. To ensure that K is a probability density, it must have the normalization feature:

$$\int_{-\infty}^{+\infty} K(u)du = 1 \quad (4)$$

Equation 4 specifies that all probability values calculated by the kernel around a point should sum to 1. In a two-dimensional space, under the polar coordinate system this normalization feature is represented as:

$$\int_{0}^{2\pi} d\theta \int_{0}^{h} K\left(\frac{r}{h}\right)dr = 1 \quad (5)$$

where Θ is the azimuth and r is the radius.

When the purpose is to estimate probability density at a location, $\hat{p}(o)$ goes through two normalizations. One is global, represented by n in the denominator of Equation 3, which gives the likelihood that a point would occur within the kernel of point i; the other normalization is local, represented by Equation 4, or specifically in a two-dimensional space, by Equation 5, which evaluates to what extent $\hat{p}(o)$ is affected by point i.

Within a physical context, when the points represent locations of *sources* of the matter that is under concern (e.g., pollutants), KDE is employed to *spread* the matter from the points to their vicinities. This process should not go through the global normalization, because usually the concentrations at those locations are not realizations of an underlying probability density distribution, rather they can be considered independent from each other. Thus, in this situation (e.g., when estimating the environmental exposure sourced from certain locations), Equation 3 should become:

$$v_o = \frac{1}{\pi h^2}\sum_{i=1}^{n} K\left(\frac{d_{i,o}}{h}\right)v_i \quad (6)$$

where v_o is the estimated concentration (exposure) value at a given location o; and v_i is the value of source point i. Equation 6 indicates that in the result of KDE, the value at a given location is the sum of all quantities it receives from all sources in the study area.

To ensure that the KDE represented by Equation 6 is a *spreading* process, the local normalization implemented by Equation 5 is still necessary. From a physical perspective, it specifies that the quantities spread to the vicinity of a source should sum up to the original total quality at the source.

Equations 6 (KDE) and 1 (interpolation) are very similar, both being combinations of values from relevant points. The difference between the two is on how the coefficients of points in this combination are determined. The coefficient of each point in Equation 6 is not a relative weight like that in Equation 2; rather, in Equation 6 the coefficient of a point has nothing to do with any other points but is determined independently by the K function applied to that point. As a counterpart of Figure 1, the result of KDE using Equation 6 can be illustrated by Figure 2.

Another important difference between the interpolation and the KDE-based spreading process is that the points in the latter are not samples, but a complete set of the points that should be taken into account, e.g., all pollution sites in an area that should be included in the analysis.

Here we present a case study of estimating spatial distribution of pesticide pollution in New Hampshire (NH), US, which calculates the kernel density based on areal rather than point data. From the NH Department of Environmental Service, we acquired data that comprehensively describe the application history of various pesticides in all NH farms during 1965–1994. Each farm is represented by one or multiple polygons in the provided Shapefile. A separate Excel sheet lists details of pesticide applications, including the quantity of a particular pesticide in an application for a farm, the date of the application, and the acreage of the application. Here we use Maneb (a pesticide) in a single year as an example to describe the estimation process. The same process was applied to each pesticide in each year. The reason for aggregating the estimates about individual applications into years is that we need to correspond the estimates (environmental exposure) to the data of patients' migration histories that are organized by years in an environmental health study of a certain disease.

KDE requires the input data to be points. Thus the general idea in this analysis is to first convert each polygon into an agglomeration of points, then attach the application quantity of Maneb to those points, and finally use the value points generated in this way to calculate the kernel density, which is an estimate of quantity of Maneb at each and every location in NH. The analysis encountered a number of challenges sourced from the complexity in the data. Among others, one challenge is that the record of an application is about a farm, but a farm may have multiple fields (polygons), and there is no information about which field(s) the pesticide was applied to; also, the recorded acreage of an application are often much smaller than the area of a polygon, and there is no information about to which portion(s) of the polygon the pesticide was applied. To compile the data into a set of reasonable points, each attached with a reasonable quantity of applied Maneb, so that the KDE can be performed, we went through a four-step process described as follows, and illustrated by Figure 3.

First, if the farm has multiple fields, without further information, we assumed that the quantity of Maneb, as well as the acreage, of a single application had been proportionally allocated to those farmlands based on their areas (implying that an application area is homogeneous in terms of the concentration of Meneb). Based on this assumption, we calculated the quantity and acreage of an application for each field polygon.

Second, considering the spatial concentration of an application (i.e., it is not likely that a small acreages of application in a relatively large farmland would be evenly scattered across the entire field), and also balancing the precision and computing burden, we chose to use one point to represent an area of 4 hectare. With this setting, we calculated the number of points that should be used to represent a polygon, based on the acreage of the application that was allocated to the polygon.

Third, according to the number of points for a polygon, we generated random points for the polygon. The total quantity of Maneb applied to the polygon were equality divided among the points and the value was attached to each point.

Fourth, using the generated value points, we created raster layers of the Maneb distribution through KDE. We tested two bandwidths, including 500 m and 1,000 m, representing how far the Maneb from a source could reach.

Snapshotting

The KDE represented by Equation 6 is modeling a spreading process, with the initial state being each source having the total quantity of the pollutant, and the final state being the result of KDE, representing the end of the spread it models (not necessarily the end of spread in the real world). However, in a real-world environmental exposure research or practice, often the initial total quantity of pollutant at a source is unknown, or the source constantly generates new pollutants, making the measurement of total quantity less practically meaningful. In this situation, usually the available data are values sampled at the source location for certain time points, e.g., the $PM_{2.5}$ concentration measured at the location of a factory chimney at a certain time of a day. Noteworthy, while such data can also be called *samples*, they are *temporal samples*, i.e., values taken at certain time points selected from a much larger number of possible time points, different from the *spatial samples* used in the spatial interpolation.

These temporal samples at the sources give a snapshot of the sources during the pollutant spreading process. To estimate the environmental exposure at every location, we want to use the snapshot of the sources to infer a snapshot of the entire area. In this paper, we propose to use the term *snapshotting* to refer to such an inference.

In contrast to the KDE represented by Equation 6, during which the initial state (a source has the total quantity of the pollutants) is not part of the ending state (the pollutant has been spread), snapshotting is modelling a single state, and the temporal sample values are part (eventually, anchor points) of this modelled state. Technically, snapshotting is simply the KDE (Equation 6) without being adjusted by the kernel area:

$$v_o = \sum_{i=1}^{n} K\left(\frac{d_{i,o}}{h}\right) v_i \quad (7)$$

A fundamental feature of snapshotting that is not fully implied by Equation 7 is that in the result of snapshotting, the values right at the locations of input points maintain unchanged, whereas in the result of KDE the values at the locations of input points would be smaller than the original values, as illustrated by Figure 2. This source-invariable feature of *snapshotting* can be represented by Equation 8 and illustrated by Figure 4.

$$K_i(0) = 1 \quad (8)$$

In Equation 8, K_i denotes the kernel function applied to source i; and 0 indicates that the location is right at source i. Note that most kernel functions commonly used in KDE, which are designed to meet the constraint set by Equation 5, do not meet the constraint set by Equation 8.

Figure 4 only illustrates a simple situation, where all sources are far away enough from each other so that no source is within the kernel of another source. In a more general situation, when a source is possibly within kernels of other sources, the sample value measured at a source is the sum of the quantities it receives from all other sources, on top of its own, as illustrated by Figure 5.

Figure 5 demonstrates that the sampled values at the sources cannot be directly sent into Equation 7. Instead, one has to first use the sampled values to derive the sources' own values. If we denote the sampled values at sources $1, 2, \ldots, n$ as $x'_1, x'_2, \ldots, x'_n$; denote the sources' own values as $x_1, x_2, \ldots, x_n$; and denote $K\left(\frac{d_{i,j}}{h}\right)$ as $k_{i,j}$, where $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, n$, then:

$$\begin{aligned} x'_1 &= x_1 + k_{1,2}x_2 + \cdots + k_{1,n}x_n \\ x'_2 &= k_{2,1}x_1 + x_2 + \cdots + k_{2,n}x_n \\ &\cdots\cdots\cdots \\ x'_n &= k_{n,1}x_1 + k_{n,2}x_2 + \cdots + x_n \end{aligned} \quad (9)$$

Using Cramer's Rule in linear algebra, $x_1, x_2, \ldots, x_n$ can be solved as:

$$x_1 = \frac{D_1}{D}, x_2 = \frac{D_2}{D}, \cdots, x_n = \frac{D_n}{D} \quad (10)$$

where

$$D = \begin{vmatrix} 1 & k_{1,2} & \cdots & k_{1,n} \\ k_{2,1} & 1 & \cdots & k_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ k_{n,1} & k_{n,2} & \cdots & 1 \end{vmatrix} \quad (11)$$

$$D_j = \begin{vmatrix} 1 & \cdots & k_{1,j-1} & x'_1 & k_{1,j+1} & \cdots & k_{1,n} \\ k_{2,1} & \cdots & k_{2,j-1} & x'_2 & k_{2,j+1} & \cdots & k_{2,n} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ k_{n,1} & \cdots & k_{n,j-1} & x'_n & k_{n,j+1} & \cdots & 1 \end{vmatrix} \quad (12)$$

The sources' own values (x_1, x_2,…, x_n) calculated with Equations 10–12 will then be sent into Equation 7 (constrained by Equation 8) to calculate the *snapshotting* scenario. We have included a tool of *snapshotting* in *ArcHealth*, a software package especially for health-related geospatial analyses and runs as an *Extension* of ArcMap. The reader can contact us to request ArcHealth. As far as we know, this function has not been included in any other GIS packages.

Discussion and Summary

A rule of thumb we propose for selecting among the three methods: If we know that the pollution data are spatially about the source(s), either point sources or non-point sources, KDE or *snapshotting* should be the choice; on the other hand, if the data are not limited to the sources, especially if the spatial distribution of the data points follows a certain spatial sampling strategy, e.g., random or stratified, interpolation should be used.

Stemming from that rule, a requirement to the data for KDE and *snapshotting* is that they should be a complete set of all sources, not a sample set. Exceptions to this requirement might exist in some rare situations, where one can assume that the contribution of a source can be inferred from the data of other sources, and the locations included in the analysis are selected through a certain valid sampling strategy.

Some studies develop regression models to infer pollution (exposure) values at un-sampled locations based on the measured values at sample locations (serving as the dependent in the model) and data of related environmental factors (serving as the independents in the model) (e.g., Ayotte et al. 2006; Pająk, Halecki and Gąsiorek 2017) Noteworthy to point out, this kind of regression modelling, within the context of this paper, can be considered as a type of interpolation.

In both interpolation and *snapshotting*, the input values are part of the surface these two methods are inferring (even if in an interpolation, the input sample value may not be exactly the same as the interpolated value at the sample location, where the residual represents the difference between the derived global model and the local variability), whereas in a spreading KDE, the input values represent the initial state, and are not part of the ending state (the resulting density surface generated by the KDE).

The essential difference between interpolation and *snapshotting* is that in interpolation, the input values are *spatial samples* that are randomly (or using other sampling strategies) selected from a physical surface, and the inference is based on the spatial autocorrelation that is derived from the relationship among those samples, e.g., through calculating semivariance; whereas in *snapshotting*, the input values are *temporal samples* measured at a complete set of source locations, and their influences to their vicinities are mutually independent. While in *snapshotting* the influence also follows the rule of spatial autocorrelation, i.e., the influence has a distance decay, this spatial autocorrelation is represented by the kernel function independently applied to each source and is not related to other sources.

The choice of different kernel function (K) in KDE and in the snapshotting method reflects the researcher's understanding of the physical spreading process. However, Silverman (1986) states that the choice of mathematical function for K would not dramatically change the overall pattern of the estimation result, and thus has a less significant impact on the result than the choice of bandwidth. Nevertheless, in the environmental exposure estimation, the kernel function can/should incorporate or entirely revised by certain physical processes like wind and water flows, which makes the construction of kernel function itself a research topic, and it is a topic that concerns more physical process rather than mathematics.

The meanings and intentions of the three methods are fundamentally distinctive, which is summarized in Table 1, and when being used for estimating environmental exposure (and likely in other applications), the results from the three methods can be considerably different from one another, in terms of both absolute value and relative variability. To make a correct selection, the researcher needs to have an in-depth understanding of the nature of the available data, the mathematical process needed for the inference, and the essential meaning of the output.

Biographies

Xun Shi is a Professor of Geography at Dartmouth College, USA. His primary research interest is in spatial analysis and its application in human health studies.

Meifang Li is a Ph.D. candidate at School of Geography and Planning, Sun Yat-sen University, China. She is currently a visiting scholar at the Geography Department, Dartmouth College. Her primary research interest is in spatiotemporal modelling of communicable diseases.

Olivia Hunter is a senior undergraduate student at Dartmouth College, USA. She is majored in Biology, with a minor in Geography.

Bart Guetti is a freelance researcher specialized in GIS and spatial analysis.

Angeline Andrew is a Professor of Epidemiology at Geisel School of Medicine, Dartmouth College, USA. Her primary research area is in environmental health.

Elijah Stommel is a Professor of Neurology at Geisel School of Medicine, Dartmouth College, USA. He specialized in environmental impacts on neural diseases.

Walter Bradley is a Professor of Neurology at University of Miami, USA. He specialized in environmental impacts on neural diseases.

Margaret Karagas is a Professor of Epidemiology at Geisel School of Medicine, Dartmouth College, USA. Her primary research area is in environmental health.

References

Aelion CM, Davis HT, Lawson AB, Cai B, and McDermott S. 2013 “Associations between soil lead concentrations and populations by race/ethnicity and income-to-poverty ratio in urban and rural areas.” Environmental geochemistry and health 35(1):1–12. doi:10.1007/s10653-012-9472-0. [PubMed: 22752852]

Ayotte JD, Nolan BT, Nuckols JR, Cantor KP, Robinson GR, Baris D, Hayes L, Karagas M, Bress W, and Silverman DT. 2006 “Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment.” Environmental science & technology 40(11):3578–3585. doi: 10.1021/es051972f. [PubMed: 16786697]

Ayotte JD, Baris D, Cantor KP, Colt J, Robinson GR, Lubin JH, Karagas M, Hoover RN, Fraumeni JF, and Silverman DT. 2006 “Bladder cancer mortality and private well use in New England: an ecological study.” Journal of Epidemiology & Community Health 60(2):168–172. doi:10.1136/jech.2005.038620. [PubMed: 16415269]

Borge R, Narros A, Artíñano B, Yagüe C, Gómez-Moreno FJ, de la Paz D, Román-Cascón C, Díaz E, Maqueda G, and Sastre M. 2016 “Assessment of microscale spatio-temporal variation of air pollution at an urban hotspot in Madrid (Spain) through an extensive field campaign.” Atmospheric Environment 140:432–445. doi: 10.1016/j.atmosenv.2016.06.020.

Carlon C, Critto A, Marcomini A, and Nathanail P. 2001 “Risk based characterisation of contaminated industrial site using multivariate and geostatistical tools.” Environmental pollution 111(3):417–427. doi:10.1016/S0269-7491(00)00089-0. [PubMed: 11202746]

Goovaerts P, AvRuskin G, Meliker J, Slotnick M, Jacquez G, and Nriagu J. 2005 “Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan.” Water Resources Research 41(7):W07013. doi:10.1029/2004WR003705.

Gottschalk F, Scholz RW, and Nowack B. 2010 “Probabilistic material flow modeling for assessing the environmental exposure to compounds: Methodology and an application to engineered nano-TiO2 particles.” Environmental Modelling & Software 25(3):320–332. doi:10.1016/j.envsoft.2009.08.011.

Hooker PJ, and Nathanail CP. 2006 “Risk-based characterisation of lead in urban soils.” Chemical Geology 226(3–4):340–351. doi:10.1016/j.chemgeo.2005.09.028.

Hu Z, Liebens J, and Rao KR. 2008 “Linking stroke mortality with air pollution, income, and greenness in northwest Florida: an ecological geographical study.” International journal of health geographics 7(1):20. doi:10.1186/1476-072X-7-20. [PubMed: 18452609]

Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, Morrison J, and Giovis C. 2005 “A review and evaluation of intraurban air pollution exposure models.” Journal of Exposure Science and Environmental Epidemiology 15(2):185. doi:10.1038/sj.jea.7500388.

Katz DR, Cantwell MG, Sullivan JC, Perron MM, Burgess RM, Ho KT, and Charpentier MA. 2013 “Factors regulating the accumulation and spatial distribution of the emerging contaminant triclosan in the sediments of an urbanized estuary: Greenwich Bay, Rhode Island, USA.” Science of the total environment 443:123–133. doi:10.1016/j.scitotenv.2012.10.052. [PubMed: 23183224]

Kestens Y, Lebel A, Daniel M, Theriault M, and Pampalon R. 2010 “Using experienced activity spaces to measure foodscape exposure.” Health & Place 16(6):1094–1103. doi:10.1016/j.healthplace.2010.06.016. [PubMed: 20667762]

Martens DS, Cox B, Janssen BG, Clemente DB, Gasparrini A, Vanpoucke C, Lefebvre W, Roels HA, Plusquin M, and Nawrot TS. 2017 “Prenatal air pollution and newborns’ predisposition to accelerated biological aging.” JAMA pediatrics 171(12):1160–1167. doi:10.1001/jamapediatrics.2017.3024. [PubMed: 29049509]

Pająk M, Halecki W, and Gąsiorek M. 2017 “Accumulative response of Scots pine (Pinus sylvestris L.) and silver birch (Betula pendula Roth) to heavy metals enhanced by Pb-Zn ore mining and processing plants: Explicitly spatial considerations of ordinary kriging based on a GIS approach.” Chemosphere 168:851–859. doi:10.1016/j.chemosphere.2016.10.125. [PubMed: 27836278]

Shi X 2010 “Selection of bandwidth type and adjustment side in kernel density estimation over inhomogeneous backgrounds.” International Journal of Geographical Information Science 24(5): 643–660. doi:10.1080/13658810902950625.

Silverman BW 1986 Density Estimation for Statistics and Data Analysis Boca Raton, FL: Chapman & Hall/CRC Press.

Thornton LE, Pearce JR, and Kavanagh AM. 2011 “Using Geographic Information Systems (GIS) to assess the role of the built environment in influencing obesity: a glossary.” International Journal of Behavioral Nutrition and Physical Activity 8(1):71. doi:10.1186/1479-5868-8-71. [PubMed: 21722367]

Tonne C, Melly S, Mittleman M, Coull B, Goldberg R, and Schwartz J. 2006 “A case-control analysis of exposure to traffic and acute myocardial infarction.” Environmental health perspectives 115(1): 53–57. doi:10.1289/ehp.9587.

Yuan T, Shie R, Chin Y, and Chan C. 2015 “Assessment of the levels of urinary 1-hydroxypyrene and air polycyclic aromatic hydrocarbon in PM2. 5 for adult exposure to the petrochemical complex emissions.” Environmental research 136:219–226. doi: 10.1016/j.envres.2014.10.007. [PubMed: 25460640]

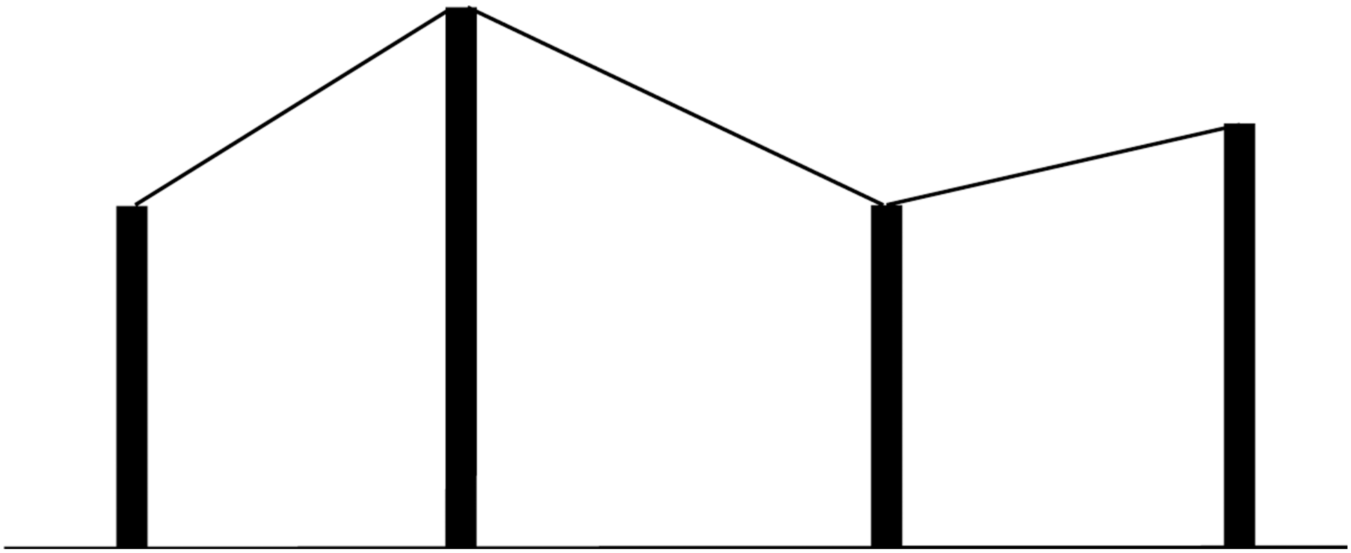

Figure 1.

An illustration of the result from an inverse-distance weighting (IDW) interpolation: each vertical bar represents a sample, with the height indicating the sample value; the thin lines passing through the tops of the bars represent the interpolated continuous surface. The result is generated using the decay factor $r = 1$.

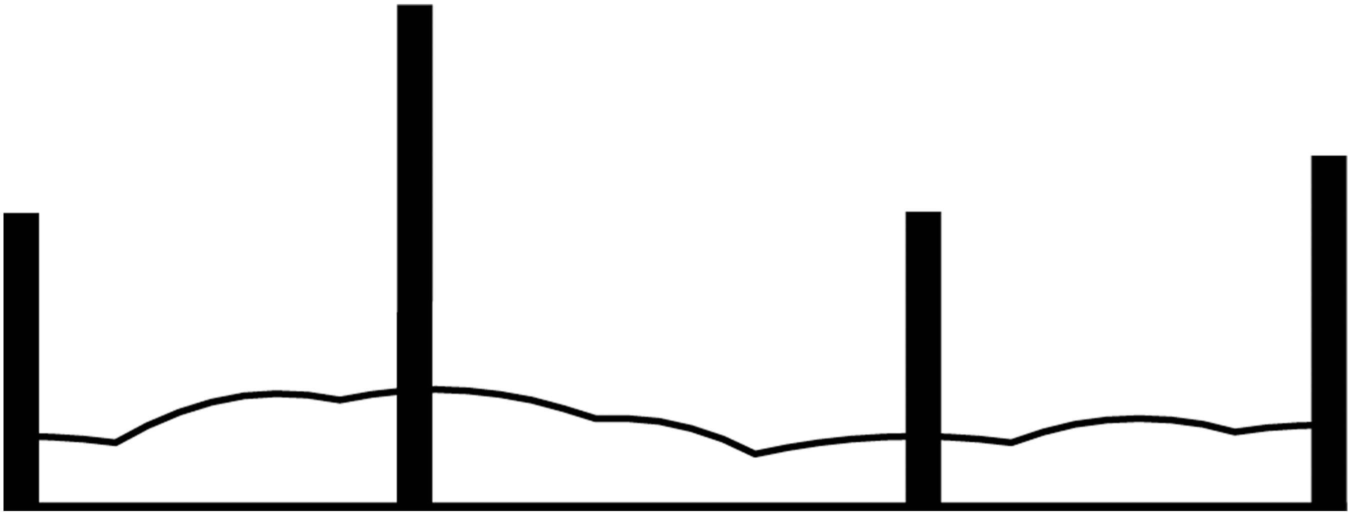

Figure 2.

An illustration of the result from kernel density estimation (KDE): each vertical bar represents a value point, with the height indicating the attribute value; the curvy thin line represents the estimation result, using the Epanechnikov function $(K(u) = \frac{3}{4}(1 - u^2))$.

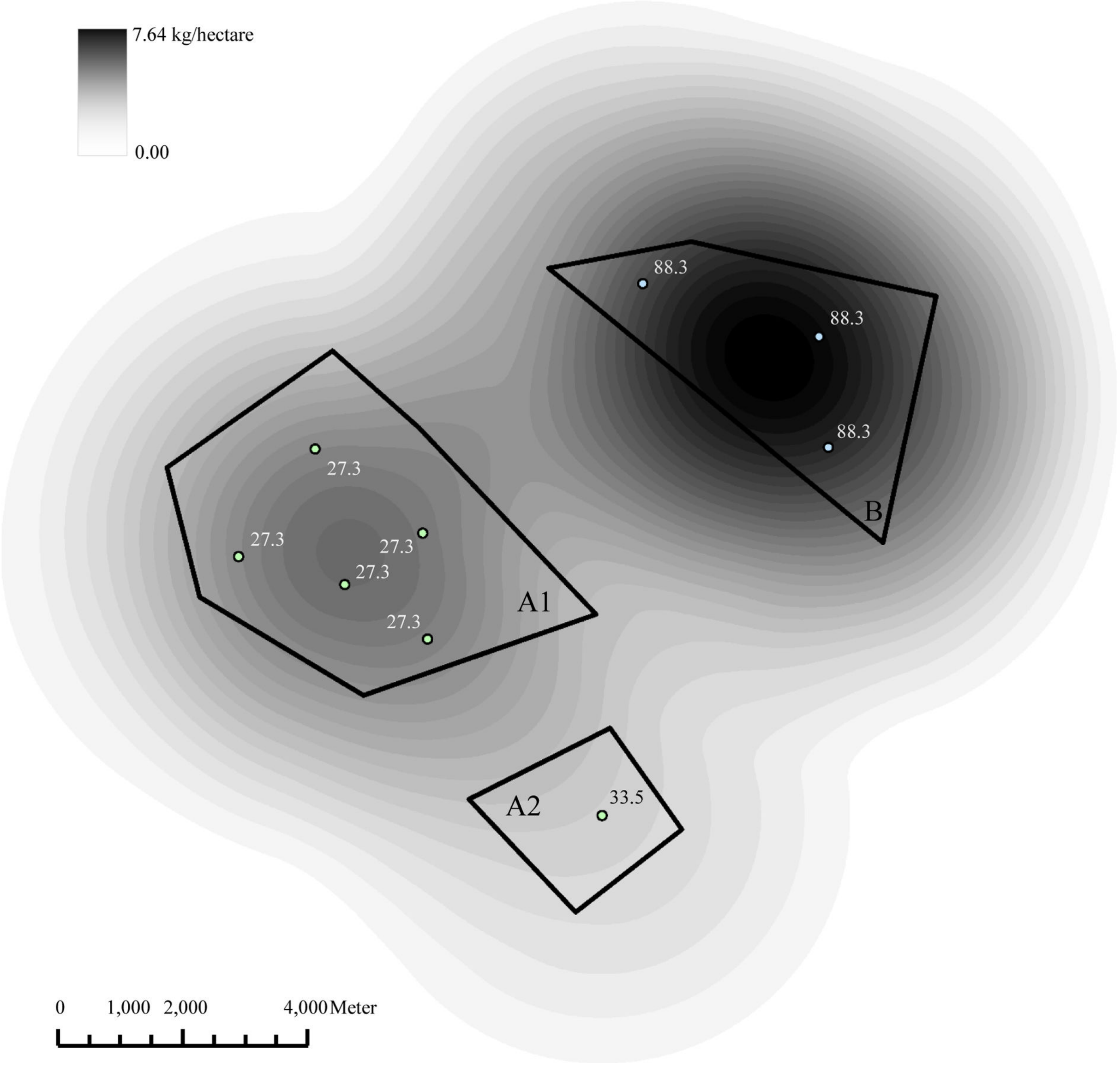

Figure 3.

A process of using kernel density estimation (KDE) to model spread of pesticides from farms. The polygons are farmlands, with A1 and A2 belonging to farm A, and B belonging to farm B. To run KDE, we first convert polygons into points by generating random points within each polygon, with the number of points to generate determined by the area of the polygon. The total amount of pesticide in an application of a farm is first proportionally allocated to different polygons of the farm, according to the areas of the polygons (not according to the number of points), and then equally allocated to each point in the polygon. These value points are then used to generate the density surface of pesticide. Note: that the points in A1 and A2 have different pesticide values is due to the precision loss in discretizing polygon area into number of points.

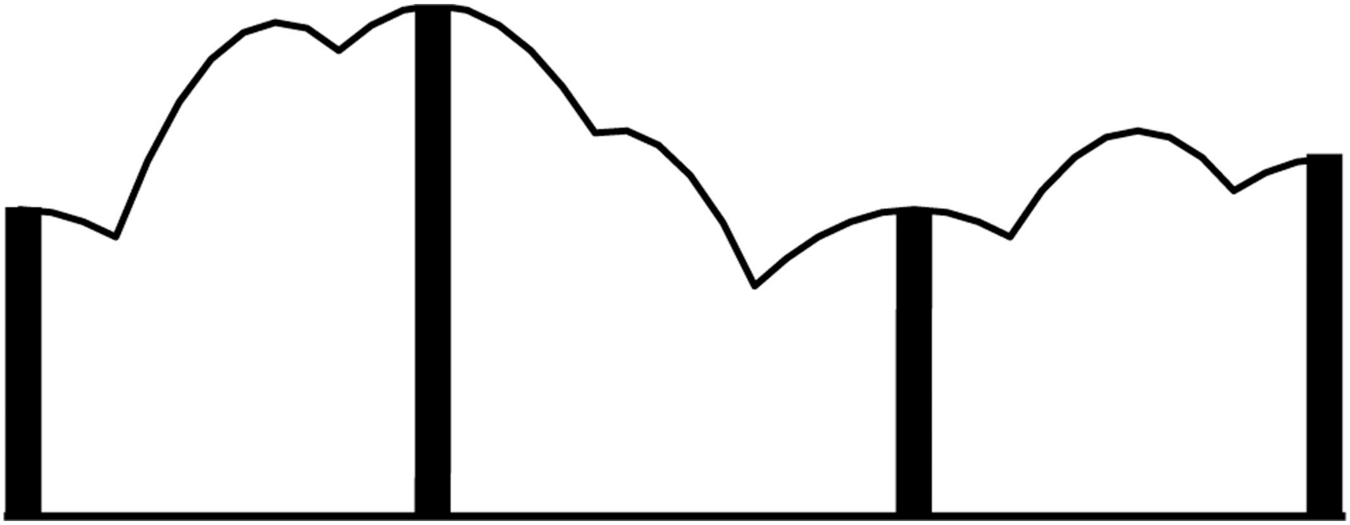

Figure 4.

An illustration of the result from *snapshotting*: each vertical bar represents a value point, with the height indicating the attribute value; the curvy thin line represents the estimation result, using a quadratic function $(K(u) = 1 - u^2)$. These points are far away enough from each other, i.e., none of them is within kernels of other points.

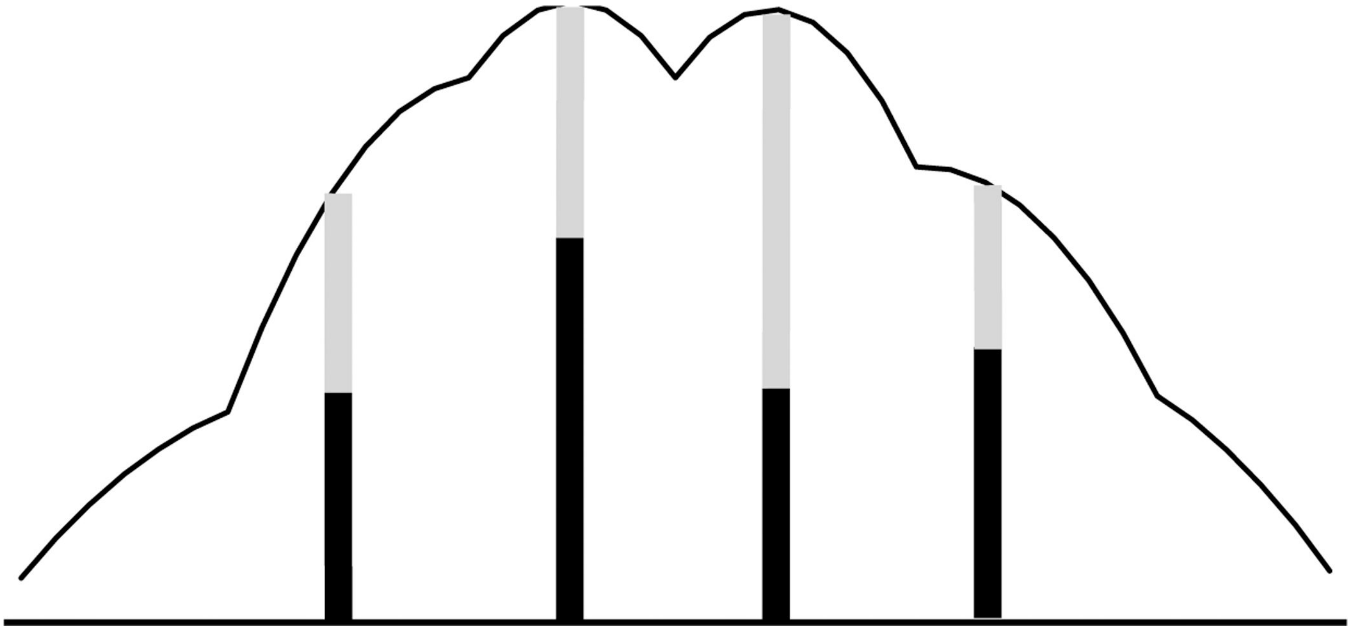

Figure 5.

An illustration of the result from *snapshotting*: each vertical bar represents a value point. The height of the dark part of a bar indicates the attribute value of the point itself; the grey part indicates the quantity it receives from other points; and the total height (dark + grey) indicates the measured value at the point. The curvy thin line represents the estimation result based on the value of each point itself (the dark part), using a quadratic function $(K(u) = 1 - u^2)$.

Table 1

summarizes the contrast among the spatial interpolation, spatial KDE, and spatial *snapshotting*.

	Equations	Input	Output	Application Examples
Interpolation	$v_o = \sum_{i=1}^{n} w_i v_i$ $\sum_{i=1}^{n} w_i = 1$	Spatial samples selected from a physical surface.	An inferred physical surface.	• Use measurements from randomly allocated monitoring stations to estimate the exposure to $PM_{2.5}$. • Use measurements from wells to estimate the exposure to arsenic in the groundwater.
KDE	$v_o = \frac{1}{\pi h^2} \sum_{i=1}^{n} K\left(\frac{d_{i,o}}{h}\right) v_i$ $\int_0^{2\pi} d\theta \int_0^h K\left(\frac{r}{h}\right) dr = 1$	Total quantity at a complete set of source locations.	The result of a spreading process.	• Use governmental records of farm applications to estimate the exposure to pesticides. • Use the factory inventory of chemical yearly release to estimate the exposure to a chemical.
Snapshotting	$v_o = \sum_{i=1}^{n} K\left(\frac{d_{i,o}}{h}\right) v_i$ $K_i(0) = v_1$	Temporal samples measured at a complete set of source locations.	A snapshot for a state during a spreading process.	• Use measurements from street intersections to estimate the exposure to $PM_{2.5}$. • Use measurements from brownfields to estimate the exposure to Trichloroethylene (TCE).