ISEIS

Journal of Environmental Informatics 38(2) 93-105 (2021)

Journal of Environmental Informatics

www.iseis.org/jei

Geospatial Information Diffusion Based on Self-Learning Discrete Regression

C. F. Huang[1, 2, 3*]

[1] *Key Laboratory of Environmental Change and Natural Disaster, Ministry of Education, Beijing Normal University, Beijing 100875, China*
[2] *State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China*
[3] *Faculty of Geographical Science, Academy of Disaster Reduction and Emergency Management, Beijing Normal University, Beijing 100875, China*

Received 17 July 2019; revised 10 January 2020; accepted 28 April 2020; published online 15 December 2020

ABSTRACT. When studying a phenomenon on the earth surface, such as natural disaster, water pollution and land use, the data in some geographic units may be insufficient. Most interpolation models cannot estimate missing data because they rely on continuous assumptions, however most geospatial data is not continuous. In this article, we develop an information diffusion technique, called self-learning discrete regression (SLDR), to infer the missing data of the gap units. To show how to use the suggested model, a virtual case based on flood experience in China is studied, where flood losses of the gap units are inferred with background data: population, per-capita GDP and relative exposure of the unit to flood. To the case, a comparison shows that SLDR is obviously superior to geographically weighted regression (GWR) and the back propagation neural network (BP network), reducing the error about 60% and 33%, respectively. To substantiate the special case arguments, ten simulation experiments are done with pure random seed numbers. The statistical average results show that the validity of GWR for filling gap units is doubtful, and SLDR is more accurate than BP network.

Keywords: information diffusion, geographic unit, regression, flood loss, simulation experiment

1. Introduction

The observations are very important for studying phenomena on the earth surface, such as natural disasters, water pollution and land use. However, in many cases, some units lack observations. For example, one day after the 2008 great Wenchuan earthquake (Zhang et al., 2009), the rescuers did not have disaster information for places where communication was lost. When it is impossible to perform an on-site investigation to collect observations, some mathematical models would be employed to estimate the missing data. The most practical method is the interpolation method, which estimates the values of a curve at any position between known points and is widely used in geographic information systems (GISs) (Eldrandaly and Abu-Zaid, 2011) and risk assessment (Stavrou and Ventikos, 2014). Any interpolation is based on the mathematical hypothesis that the corresponding interpolation space is continuous. In this case, the researcher considers the data gaps because data are only collected at discrete points, and blank data can be calculated according to a suitable continuous function. Polynomial interpolation, Hermite interpolation, spline approximation and series fitting are common interpolation models. Inverse distance weighting (De Mesnard, 2013) and ordinary kriging (Gutiérrez de Ravé et al., 2014), both of which are used to interpolate data from surround ing geographic units, fall into this category. However, except for the temperature physics field, most geospatial data are not continuous. In particular, data related to human society are usually non-continuous. Spatial data can be approximated as continuous only if the grid of the geographic units is notably small. In other words, when we study a phenomenon on the earth surface in geographic units at the township level, the continuity hypothesis for interpolation does not hold.

A variety of prediction methods are often used to extrapolate data. Among these, the grey system method is rather confusing. It appears that with the use of an accumulating generation operator on a limited amount of data, the grey method can accurately estimate the behavior of unknown systems, such as prediction of stock prices (Kayacan et al., 2010). Many researchers have noted that generations of grey sequences exhibit an exponential trend, however it is not well known that only time series data generalized from an energy system are subject to an exponential trend. For estimation of the behavior of an energy system, exponential regression (Dette et al., 2006) might be superior to the grey model. Obviously, the grey system model cannot be used to fill the data gaps in geographic units because the spatial data are not time series data.

Geographically weighted regression (GWR) is a potentially well-suited spatial predictive model (Lieske and Bender, 2011). An advantage of GWR is that it allows the actual parameters for each location in space to be estimated and mapped as opposed to the fitting of a trend surface to the parameters. Essentially, any GWR model is a statistical regression model. The least squares method is commonly used to estimate the coefficients

* Corresponding author. Tel.: +(86) 13693121969; Fax: +(86-10)58805817
E-mail address: hchongfu@bnu.edu.cn (C. F. Huang).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.202000439

in the model. GWR also offers extensions of generalized linear models, including logistic and Poisson regressions (Fotheringham et al., 2002). If we clearly know what type of function can express the relationships between *y* and *x*, any highly nonlinear or nonmonotonic relationship can be modeled using a GWR model with the known function, the parameters of which are determined using observations. However, it is notably difficult to determine which nonlinear function is appropriate for expressing a phenomenon on the earth surface. The main limitation of GWR is that, in many cases, we do not know which kind of function is appropriate to regress the given observations.

In our study of estimating the missing data, in order to avoid the mistakes caused by the wrong assumption of a statistical relationship, using the artificial neural networks (ANNs) may be an option. ANNs are well known as a tool for estimating relationships between input and output via learning on data samples. For example, an ANN model exhibited better results for the prediction of arsenic contamination in groundwater (Purkait et al., 2008). A highly nonlinear neural network model outperformed advanced statistical methods and more effectively reduced risk in managerial decision making (Marcek, 2013). The most popular ANN is the back propagation neural network (BP network) whose training algorithm is the well-known gradient descent method (Sen, 2006). In theory, ANN multilayer networks using arbitrary squashing functions can approximate any continuous function to any degree of accuracy given that sufficient hidden units are available (Hornik et al., 1989). In practice, because networks are implemented on computers, the property of universal approximation does not hold (Wray and Green, 1995). More importantly, an ANN does not converge if the training samples are contradictory due to random disturbances in the real world (Huang and Moraga, 2004).

The research goal of this article is to find a universal model that reasonably supplements incomplete spatial data when studying a phenomenon on the earth surface. The objectives of the suggested model are that, (i) It is able to process non-continuous geospatial data; (ii) It is a universal approximation, and does not rely on any assumption of a statistical relationship; (iii) It can converge even if there are contradictions in the training sample due to random interference; (iv) When the size of a training sample is small, the data estimated by the model also have higher accuracy. As an information diffusion technique, the self-learning discrete regression (SLDR) model suggested in this article has achieved these four objectives. One of the contributions of this paper is that, based on flood experience in China, it provides a full virtual case to show how to use the suggested model and it is superior to GWR and BP network in reducing the error about 60% and 33%, respectively. Another contribution is that, ten simulation experiments are done with pure random seed numbers, and results show that the validity of GWR is doubtful to supplement incomplete spatial data, and SLDR is more accurate than BP network.

2. Information Diffusion in Probabilistic Space

The concept of information diffusion was suggested in function learning from a small sample of data (Huang, 1997). The approximate reasoning of information diffusion was used to estimate probabilities and fuzzy relationships from scant and incomplete data for grassland wildfires (Liu et al., 2010). An information diffusion model was used to estimate the probability density function for average daily rainfall in an assessment of risk variability among several months in the three north-eastern provinces of China (Zhao and Zhang, 2012). An information diffusion method was also applied to calculate the risk values shown in risk radar for emergency management in a community (Huang et al., 2016).

The simplest models of the information diffusion technique are linear information distribution and normal diffusion. The latter is more convenient to use. Mathematically, normal diffusion can be illustrated in a fuzzy set, as shown in the following (Huang, 2002):

Let $X = \{x_i \mid i = 1, 2, ..., n\}$ be a given sample, and let $U = \{u\}$ be its universe of discourse (the range of all possible values). The function in Equation (1) is known as a normal diffusion function, which diffuses the information carried by observation *x* to the monitoring point *u* in the normal approach:

$$\mu(x, u) = \exp[-\frac{(x - u)^2}{2h^2}], \; x \in X, \; u \in U \tag{1}$$

The diffusion coefficient *h* can be calculated using Equation (2) (Huang, 2012):

$$h = \begin{cases} 0.8146\,(b-a), & n = 5; \\ 0.5690\,(b-a), & n = 6; \\ 0.4560\,(b-a), & n = 7; \\ 0.3860\,(b-a), & n = 8; \\ 0.3362\,(b-a), & n = 9; \\ 0.2986\,(b-a), & n = 10; \\ 2.6851(b-a)/(n-1), & n \geq 11. \end{cases} \tag{2}$$

where $b = \max_{1 \leq i \leq n} \{x_i\}$, $a = \min_{1 \leq i \leq n} \{x_i\}$.

Using a diffusion function $u(x, u)$, we change a given sample point *x* (observation) into a fuzzy set with membership function $\mu_x(u) = \mu(x,u)$ in universe *U*. The principle of information diffusion guarantees that reasonable diffusion functions exist to improve the non-diffusion estimates if the given samples are incomplete (Huang, 1997).

Current information diffusion models are based on an assumption that a given sample *X* is drawn from a population with probability density function $p(x)$. The basic advantage of using the information diffusion technique is that it can naturally fill the gaps in incomplete data using reasonable fuzzy sets to improve the estimation of $p(x)$.

If we change a random sample point *x* into a set-value sample point $\mu_x(u)$, we essentially diffuse the information carried by *x* in the probability space *U* (or subspace) using a diffusion function.

It has been proven with analytic geometry that the principle of information diffusion is applicable not only in probability space but also in any measurable space (Makó, 2005). This statement

implies that the information diffusion technique can be developed to fill the gaps caused by incomplete geospatial data.

3. Incomplete Geospatial Data

When we study a phenomenon on the earth surface in a study area, some information, such as the locations of the geographic units, is easy to obtain, and population and gross domestic product (GDP) data are also easy to collect. However, other information might not be as easy to obtain. For example, when we study the natural disaster risk in a county, it is difficult to collect historical disaster data for each township. In emergency rescue during a natural disaster, rescuers cannot obtain the disaster data of the townships where traffic and communication are interrupted.

Let G be a study area, where the phenomenon under study is denoted as F. Suppose that G is composed of n geographic units $g_1, g_2, ..., g_n$, i.e.:

$$G = \{g_1, g_2, ..., g_n\} \tag{3}$$

Furthermore, suppose that the phenomenon F could be recognized using n observations taken from the n geographic units. An observation, denoted as w, is a number or a vector that could be obtained directly or might be estimated. For example, when we study the phenomenon of "earthquake risk" in Yunnan Province, China, which is composed of 129 counties and municipal districts, the phenomenon can be written as E_{risk}, the study area G is Yunnan, and each county and municipal district is a geographic unit; thus, we write the area as the following:

$$G_{Yunnan} = \{g_1, g_2, ..., g_{129}\} \tag{4}$$

To recognize E_{risk} in G_{Yunnan}, we must know the seismic hazard and seismic vulnerability for each unit. In G_{Yunnan}, Tonghai, where a 7.8 magnitude earthquake occurred in 1970, is marked as g_{27}. In this case, the observation of g_{27} is a vector that includes "seismic hazard" and "seismic vulnerability".

All of the observations for recognizing phenomenon F in an area G form a set. When all of the observations are known, we say that the set is complete, and otherwise, it is incomplete. Formally, we give the following definition.

Definition 1: Let G be an area composed of geographic units $g_1, g_2, ..., g_n$, and let w_i be an observation of g_i. If a phenomenon F in G can be recognized using a set of observations taken from the geographic units, i.e.:

$$W = \{w_1, w_2, \ldots, w_n\} \tag{5}$$

when all w_i, $i = 1, 2, \ldots, n$, are assigned, we say that W is a complete data set with respect to F in G, and otherwise, it is incomplete.

If W is an incomplete data set, the data in W are referred to as incomplete geospatial data. An unassigned unit is known as a gap. Obviously, whether a data set is complete is related to what phenomena must be recognized and which area must be studied. The more complex or meticulous the phenomenon, the more difficult it is to obtain complete data. The greater the number of geographical units involved in the study area, the more difficult to collect complete data.

4. Information Diffusion in Geographical Space

If a sample X drawn from a population with a probability density function $p(x)$ is small, we can directly apply the information diffusion technique to improve the estimation of $p(x)$. If a data set W taken from a study area G is incomplete, to recognize a phenomenon F, we cannot directly apply the information diffusion technique to fill the data gaps.

According to Tobler's first law of geography, "Everything is related to everything else, however close things are more related than distant things" (Tobler, 1970). Inverse distance weighted (IDW) interpolation is most often used by GIS analysts to fill the data gaps (Eldrandaly and Abu-Zaid, 2011). In the IDW method, z_0, z_i, and d_i are used to denote the estimated value at point 0, the z value at known point i, and the distance between point i and point 0, respectively; z_0 is estimated using a series of known values, $z_1, z_2, \ldots, z_n$, with the aid of a series of distances $d_1, d_2, \ldots, d_n$ as media.

Let w_0 and w_i be the estimated value in geographic unit g_0 and the observed value in unit g_i, respectively. To mathematically express the information diffusion in geographical space, we first formally define the terms "observed unit", "gap unit", "media" and "background data".

Definition 2: Let g and o be two geographic units in a study area. If g is observed and assigned but o is not, g is known as an *observed unit* and o is known as a *gap unit* to recognize phenomenon F. For example, in a flood area, the units where the flood disasters have been investigated are observed units, and other units are gap units.

Definition 3: Let o be a gap unit. With data ω and a series of observed units, if a model exists to assign o to a value for the purpose of recognizing phenomenon F, ω are known as *media*. For example, in the IDW method, $\omega = \{d_1, d_2, \ldots, d_n\}$ are media. Selected attribute values describing the geographic features are media. The attribute values are known as background data.

Definition 4: Let o be a gap unit in area G:

$$G = \{g_1, g_2, ..., g_{n-1}, o\} \tag{6}$$

If a set Z_G of the attribute values describing selected geographic features are media, Z_G is known as a background data set:

$$Z_G = \{z_{g_1}, z_{g_2}, \cdots, z_{g_{n-1}}, z_o\} \tag{7}$$

For example, the attribute values of population and per-capita GDP features are background data for recognizing the losses of natural disasters. Thus, we can formally give a definition of information diffusion in geographical space:

Definition 5: Let W be an incomplete data set for recognizing phenomenon F in area G, and let Z_G be a background data

set. If a model γ can use Z_G to make W complete, it is said that γ uses Z_G to diffuse the information of W in G for recognition of F. Figure 1 shows the logical relationship among the objects in Definition 5.

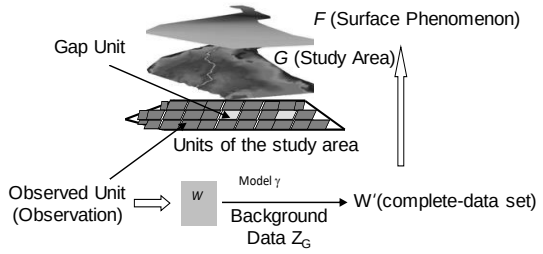

Figure 1. With support from background data Z_G, model γ diffuses information of W from observed units to fill gap units, and W becomes a complete data set W' that serves to recognize a phenomenon F in study area G.

5. An Information Diffusion Model with Background Data

Without loss of generality, suppose that a study area G is composed of $n - q$ observed units $g_1, g_2, ..., g_{n-q}$ and q gap units $g_{n-q+1}, ..., g_n$, i.e.:

$$G = \{g_1, g_2, ..., g_{n-q}, g_{n-q+1}, ..., g_n\} \tag{8}$$

Furthermore, suppose that the attribute values of t geographic features are background data. In unit g_i, the attribute value of the jth feature is z_{ij}. The information of G is shown in Table 1.

Table 1. Observations and Background Data in Area G

Geographic Unit	Background Data			Observation
g_1	z_{11}	z_{12}	$\dots z_{1t}$	w_1
g_2	z_{21}	z_{22}	$\dots z_{2t}$	w_2
…	…			…
g_{n-q}	z_{n-q1}	z_{n-q2}	$\dots z_{n-qt}$	w_{n-q}
g_{n-q+1}	$z_{n-q+1,1}$	$z_{n-q+1,2}$	$\dots z_{n-q+1,t}$	Unknown
…	…			…
g_n	z_{n1}	z_{n2}	$\dots z_{nt}$	Unknown

Recalling the above interpolation methods and prediction methods, we know that unless the size of the geographic unit in Table 1 is very small, or we know which type of function is appropriate to express a dependent variable w_i with independent variables $z_{i1}, z_{i2}, ..., z_{it}$, we cannot use those methods to fill the gap units $g_{n-q+1}, ..., g_n$. In this section, we suggest an information diffusion model to fill the gap units with background data.

This model consists of two parts. The first is a multiple normal diffusion to construct a relationship matrix. The second is an approximate reasoning to infer values in the gap units with background data.

5.1. Constructing a Relationship Matrix

Let $U_1, U_2, ..., U_t$ be t monitoring spaces that serve to diffuse background data of t features, respectively, and let U_{t+1} be a monitoring space that diffuses observations obtained from the observed units. Let $\lambda = t + 1$, we define a λ-dimensional monitoring space:

$$U_1 \times U_2 \cdots \times U_\lambda \tag{9}$$

where $U_j = \{u_{j1}, u_{j2}, \dots, u_{jm_j}\}$, $j = 1, 2, \dots, \lambda$. In this subsection, two kinds of subscripts are used: double or multiple subscripts and variable subscripts. They are all expressed in strict mathematics. For example, a double subscript is used in u_{j1}, i.e., j and 1, which can also be written as $u_{j,1}$. A variable subscript is used in u_{jm_j}, i.e., m_j. For different j, m may be different or the same.

Let $\tau = n - q$. From Table 1, we obtain a λ-dimensional sample X with size τ:

$$X = \{(x_{i1}, x_{i2}, \cdots, x_{i\lambda-1}, x_{i\lambda}) \mid i = 1, 2, \cdots, \tau\} \tag{10}$$

where $x_{i1} = z_{i1}, x_{i2} = z_{i2}, \dots, x_{i\lambda-1} = z_{it}, x_{i\lambda} = w_i$, $i = 1, 2, \dots, \tau$.

For a λ-dimensional sample point:

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{i\lambda}) \in X, \tag{11}$$

and a λ-dimensional monitoring point:

$$u = (u_{1k_1}, u_{2k_2}, \cdots, u_{\lambda k_\lambda}) \in U_1 \times U_2 \cdots \times U_\lambda, \tag{12}$$

where $k_j \in \{1, 2, \dots, m_j\}$, $j = 1, 2, \dots, \lambda$, we use the λ-dimensional normal diffusion formula in Equation (13) to diffuse the information of x to point u:

$$\mu(x_i, u) = \prod_{j=1}^{\lambda} \exp[-\frac{(x_{ij} - u_{jk_j})^2}{2h_j^2}] \tag{13}$$

where the diffusion coefficient h_j can be calculated using Equation (2) with the attribute values (background data) of the jth feature and the observations in Table 1. Let:

$$Q_{k_1 k_2 \cdots k_\lambda} = \sum_{i=1}^{\tau} \prod_{j=1}^{\lambda} \exp[-\frac{(x_{ij} - u_{jk_j})^2}{2h_j^2}] \tag{14}$$

We can obtain an information matrix of X in $U_1 \times U_2 \cdots \times U_\lambda$, as shown in Equation (15):

$$Q = \{Q_{k_1 k_2 \cdots k_{\lambda-1} k_\lambda}\}_{m_1 \times m_2 \times \cdots \times m_{\lambda-1} \times m_\lambda}, \ \forall k_\lambda \in \{1, 2, ..., m_\lambda\} \tag{15}$$

Let:

$$S_{k_\lambda} = \max_{\substack{1 \le k_j \le m_j \\ 1 \le j \le \lambda}} \{Q_{k_1 k_2 \cdots k_{\lambda-1} k_\lambda}\}, \tag{16}$$

and $$r_{k_1 k_2 \cdots k_{\lambda-1} k_\lambda} = \frac{Q_{k_1 k_2 \cdots k_{\lambda-1} k_\lambda}}{S_{k_\lambda}}. \tag{17}$$

Table 2. Losses and background Data in the Flood Disaster Area

Unit	Background Data			Loss	Unit	Background Data			Loss
g	z_1	z_2	z_3	*w*	*g*	z_1	z_2	z_3	*w*
1	1,911	2,571	0.55	656,548	55	1,391	2,748	0.55	2,759,376
2	1,078	4,075	0.68	2,476,669	56	1,720	3,570	0.04	1,136,662
3	2,530	4,868	0.81	6,755,590	57	670	2,654	0.58	1,869,182
4	1,419	1,947	0.44	Unknown	58	3,693	5,620	0.77	12,672,757
5	2,556	3,753	0.57	5,002,579	59	2,083	2,124	0.40	1,850,408
6	1,033	3,540	0.51	1,839,030	60	1,007	3,386	0.48	2,440,424
7	1,761	2,001	0.80	297,269	61	4,106	3,059	0.81	7,077,901
8	2,296	2,113	0.43	1,886,001	62	1,802	4,483	0.41	4,522,170
9	2,476	1,230	0.65	1,879,989	63	1,748	4,714	0.59	5,497,062
10	1,747	3,494	0.65	985,527	64	1,339	2,507	0.25	2,458,746
11	560	3,319	0.37	1,444,768	65	3,211	3,107	0.77	6,449,972
12	2,614	954	0.72	1,357,580	66	3,404	3,168	0.71	6,205,189
13	1,580	3,304	0.62	4,217,023	67	3,331	2,138	0.81	4,062,944
14	1,786	4,437	0.61	2,739,781	68	2,977	4,224	0.76	7,279,226
15	1,770	3,114	0.06	1,643,408	69	2,093	2,667	0.36	2,176,381
16	2,040	1,531	0.60	Unknown	70	2,081	3,881	0.78	2,664,408
17	3,190	2,318	0.73	3,910,232	71	1,364	2,395	0.65	3,696,478
18	2,155	3,932	0.69	2,707,680	72	2,006	3,613	0.54	3,004,322
19	1,174	1,813	0.29	1,192,620	73	1,036	2,338	0.63	740,752
20	2,159	3,230	0.70	5,432,413	74	2,307	3,447	0.41	4,649,048
21	613	6,407	0.29	1,165,556	75	2,675	2,523	0.78	3,349,675
22	1,587	1,152	0.40	1,687,707	76	1,767	744	0.80	245,698
23	1,735	3,554	0.35	3,657,313	77	1,543	5,134	0.61	Unknown
24	1,101	4,817	0.80	3,754,194	78	766	4,201	0.45	1,007,248
25	3,174	3,326	0.74	5,907,755	79	1,565	2,510	0.52	360,735
26	2,671	3,916	0.54	3,356,160	80	2,780	2,783	0.80	3,476,874
27	1,653	961	0.56	634,680	81	764	4,091	0.54	2,620,863
28	1,555	6,628	0.09	3,512,912	82	1,851	2,650	0.47	3,149,617
29	2,039	923	0.55	1,059,176	83	343	3,958	0.23	1,374,069
30	2,803	4,328	0.56	5,530,820	84	1,085	1,743	0.36	923,417
31	1,179	2,829	0.69	7,844	85	2,482	3,010	0.66	3,554,269
32	950	3,640	0.60	Unknown	86	1,611	3,127	0.51	Unknown
33	2,489	3,181	0.61	5,106,087	87	1,985	1,579	0.80	1,903,301
34	2,459	5,866	0.94	9,188,236	88	2,434	3,689	0.55	5,978,810
35	2,842	4,487	0.65	8,085,900	89	1,400	1,127	0.86	2,105,238
36	1,369	3,417	0.65	2,060,827	90	2,567	3,307	0.56	3,496,726
37	2,201	3,420	0.42	3,011,924	91	2,963	2,198	0.66	4,069,675
38	1,362	2,427	0.51	1,153,395	92	1,525	4,106	0.45	3,041,438
39	2,318	3,997	0.63	3,730,012	93	2,766	3,875	0.60	4,465,330
40	3,272	4,588	0.66	Unknown	94	2,014	5,150	0.73	5,640,014
41	1,990	4,269	0.48	4,572,379	95	2,313	4,468	0.88	6,518,108
42	862	2,568	0.53	1,989,547	96	2,651	2,710	0.79	2,500,870
43	1,067	3,815	0.63	3,473,850	97	1,968	4,835	0.54	5,492,120
44	3,027	4,721	0.63	7,171,829	98	2,439	2,067	0.78	2,863,983
45	2,743	3,392	0.37	5,859,156	99	1,584	1,209	0.64	1,109,807
46	2,477	4,659	0.46	6,514,906	100	2,009	2,942	0.91	3,625,000
47	3,412	1,884	0.77	3,006,031	101	860	1,948	0.32	Unknown
48	3,226	2,989	0.72	4,460,642	102	1,500	1,914	0.55	1,211,300
49	1,423	2,518	0.78	1,230,468	103	1,460	4,993	0.30	5,126,319
50	1,231	2,191	0.43	2,641,717	104	2,686	4,615	0.69	6,516,619
51	2,519	2,230	0.73	3,517,937	105	2,479	3,071	0.22	4,047,061
52	2,291	4,732	0.63	5,600,220	106	3,357	4,567	0.93	7,496,821
53	1,892	3,386	0.85	Unknown	107	497	5,641	0.33	62,251
54	877	5,306	0.46	572,849	108	2,227	2,969	0.67	3,441,624

Note: z_1-Population; z_2-Per-capita GDP (RMB Yuan); z_3-Relative exposure to flood; *w*-Flood loss (RMB Yuan).

We can obtain a relationship matrix for the background data and observations in Table 1, written as:

$$R = \{r_{k_1 k_2 \cdots k_{\lambda-1} k_\lambda}\}_{m_1 \times m_2 \times \cdots \times m_{\lambda-1} \times m_\lambda} \quad (18)$$

5.2. Inferring with Background Data

Let $z = (z_1, z_2, ..., z_t)$ be the background data of a gap unit in Table 1 and let:

$$u_{\lambda-1} = (u_{1k_1}, u_{2k_2}, \cdots, u_{\lambda-1 k_{\lambda-1}}) \in U_1 \times U_2 \cdots \times U_{\lambda-1} \quad (19)$$

where z can be changed into a fuzzy set in the universe of discourse $U_1 \times U_2 \cdots \times U_{\lambda\text{-}1}$ using the λ-1-dimensional normal diffusion formula in Equation (20) and normalizing by the maximum value, as shown in Equation (21):

$$\mu(z, u_{\lambda-1}) = \prod_{j=1}^{\lambda-1} \exp[-\frac{(z_j - u_{jk_j})^2}{2h_j^2}] \quad (20)$$

$$\begin{cases} a_{k_1 k_2 \cdots k_{\lambda-1}} = \dfrac{q_{k_1 k_2 \cdots k_{\lambda-1}}}{s}, \quad s = \max\limits_{\substack{1 \le k_j \le m_j \\ 1 \le j \le \lambda-1}} \{q_{k_1 k_2 \cdots k_{\lambda-1}}\}, \\ q_{k_1 k_2 \cdots k_{\lambda-1}} = \prod\limits_{j=1}^{\lambda-1} \exp[-\dfrac{(z_j - u_{jk_j})^2}{2h_j^2}] \end{cases} \quad (21)$$

The fuzzy set is written as $\tilde{A}$ with memberships $a_{k_1 k_2 \cdots k_{\lambda-1}}$, $k_j = 1, 2, \ldots, m_j$, $j = 1, 2, \ldots, \lambda - 1$. For the fuzzy input $\tilde{A}$, using the approximate reasoning operator represented in Equation (22), we can obtain a fuzzy output $\tilde{B}$ with a membership function $\mu_B(u_{\lambda k_\lambda})$:

$$\mu_B(u_{\lambda k_\lambda}) = \max_{\substack{1 \le k_j \le m_j \\ 1 \le j \le \lambda-1}} \min\{a_{k_1 k_2 \cdots k_{\lambda-1}}, r_{k_1 k_2 \cdots k_{\lambda-1} k_\lambda}\} \quad (22)$$

Finally, using the center-of-gravity method in Equation (23), we obtain a crisp value for w:

$$w = \frac{\sum_{k_\lambda=1}^{m_\lambda} \mu_B(u_{\lambda k_\lambda}) u_{\lambda k_\lambda}}{\sum_{k_\lambda=1}^{m_\lambda} \mu_B(u_{\lambda k_\lambda})} \quad (23)$$

The model consisting of Equations (13) ~ (23) is known as self-learning discrete regression (SLDR).

6. A Virtual Case for Filling Gaps with Background Data

Based on China's flood experience, we give the following virtual case: a disastrous flood occurred in an area, causing serious economic losses. We quickly obtained data on flood losses in some geographic units, however we are unable to obtain data in the units where traffic and communication have been disrupted. In this section, we apply the developed information diffusion technique to complete the flood loss data.

A total of 108 units are located in the flood disaster area, where 100 units are the observed units in loss w and 8 units are gap units. The information in the area is shown in Table 2. The background data for inferring the loss in a gap unit include population z_1, per-capita GDP z_2, and relative exposure z_3 of the unit to flood, which is determined by the flood path and the terrain of the unit.

The background data z_1, z_2, and z_3 are randomly generated by using a three-dimensional normal distribution, and the loss w is obtained by using a nonlinear function with the background data, and superimposed random interference. In total, 108 data sets are completely generated. Then, the first 100 data sets serve as the 100 observed units. After artificially deleting the loss values in the next 8 units, we consider the units as 8 gap units. Figure 2 shows a subarea in this case, where g_1, g_2, and g_3 are observed units and g_4 is a gap unit.

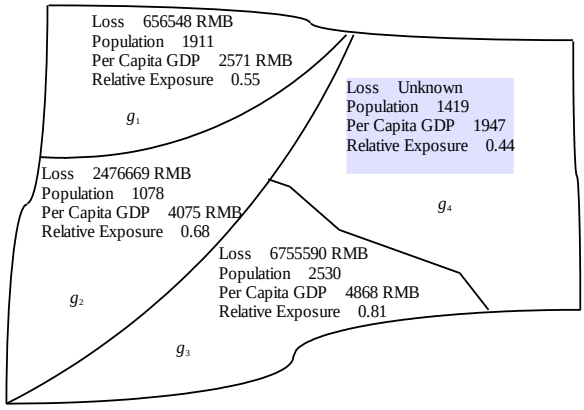

Figure 2. Subarea of a flood disaster area. The flood loss is unknown in unit g_4.

To make the virtual case more realistic, 100 complete data sets and 8 incomplete data sets are randomly assigned to the 108 units. From the 100 observed units, we have a 4-dimensional sample X shown in Equation (24):

$$\begin{aligned} X &= \{(x_{1,1}, x_{1,2}, x_{1,3}, x_{1,4}), (x_{2,1}, x_{2,2}, x_{2,3}, x_{2,4}), ..., \\ &\quad (x_{100,1}, x_{100,2}, x_{100,3}, x_{100,4})\} \\ &= \{(1911, 2571, 0.55, 656548), \\ &\quad (1078, 4075, 0.68, 2476669), ..., \\ &\quad (2227, 2969, 0.67, 3441624)\} \end{aligned} \quad (24)$$

Please note that there are 100 sample points in Equation (24) rather than 108 because 8 units in Table 2 are gap units. Let:

$$X_1 = \{x_{1,1}, x_{2,1}, ..., x_{100,1}\} = \{1911, 1078, ..., 2227\}$$

$$X_2 = \{x_{1,2}, x_{2,2}, ..., x_{100,2}\} = \{2571, 4075, ..., 2969\}$$

$$X_3 = \{x_{1,3}, x_{2,3}, ..., x_{100,3}\} = \{0.55,\ 0.68,\ ...,\ 0.67\}$$
$$X_4 = \{x_{1,4}, x_{2,4}, ..., x_{100,4}\}$$
$$= \{656548,\ 2476669,\ ...,\ 3441624\} \quad (25)$$

Using Equation (2) for X_1, X_2, X_3, and X_4, respectively, we obtain a set of diffusion coefficients in Equation (26):

$$H = \{h_1,\ h_2,\ h_3,\ h_4\}$$
$$= \{102.061,\ 159.587,\ 0.024,\ 343500.594\} \quad (26)$$

According to the accuracy we require, we apply the following U_1, U_2, U_3, and U_4 in Equation (27) with equal step lengths as monitoring spaces for diffusing the population, per-capita GDP, relative exposure and flood loss, respectively:

$$U_1 = \{u_{1,1},\ u_{1,2}, \cdots,\ u_{1,30}\} = \{343,\ 472.76, \cdots,\ 4106\}$$
$$U_2 = \{u_{2,1},\ u_{2,2}, \cdots,\ u_{2,30}\} = \{744,\ 946.9, \cdots,\ 6628\}$$
$$U_3 = \{u_{3,1},\ u_{3,2}, \cdots,\ u_{3,30}\} = \{0.04,\ 0.07, \cdots,\ 0.94\}$$
$$U_4 = \{u_{4,1},\ u_{4,2}, \cdots,\ u_{4,30}\}$$
$$= \{7844,\ 444565.12, \cdots,\ 12672754\} \quad (27)$$

In theory, the more points that monitoring space U contains, the better. However, too many points will only increase the amount of calculations but not improve the accuracy of the model. Usually, we use the amount about three times the value of the Otness-Encysin formula Equation (28) (Otness and Encysin, 1972) to determine the number of points:

$$m = 1.87(n-1)^{2/5} \quad (28)$$

where m is the number of intervals for constructing a histogram and n is the size of the given sample. In our case, n = 100, and we obtain m = 11. Then, 30 is approximately equal to three times 11. The first and last elements of U are the minimum and maximum values of the sample, respectively.

Using Equation (13), we diffuse the information of X in Equation (24) in the 4-dimensional monitoring space $U_1 \times U_2 \times U_3 \times U_4$. For example, if we diffuse the sample point:

$$x_{70} = (x_{70,1}, x_{70,2}, x_{70,3}, x_{70,4}) = (2675, 2523, 0.78, 3349675)$$

to the monitoring point

$$u = (u_{1,19}, u_{2,10}, u_{3,25}, u_{4,8}) = (2678.66, 2570.07, 0.78, 3064891.50),$$

we have the following:

$$\mu(x_{70}, u) = \exp[-\frac{(x_{70,1} - u_{1,19})^2}{2h_1^2} - \frac{(x_{70,2} - u_{2,10})^2}{2h_2^2}$$
$$- \frac{(x_{70,3} - u_{3,25})^2}{2h_2^2} - \frac{(x_{70,4} - u_{4,8})^2}{2h_4^2}] = 0.665.$$

Note that x_{70} is not the background data and loss in unit g_{70} in Table 2, and x_{70} corresponds to unit g_{75} because the gap units of Table 2 are not in sample X. By summing all diffused information with Equation (13), we obtain an information matrix $\boldsymbol{Q}$ of X in monitoring space $U_1 \times U_2 \times U_3 \times U_4$. For example, corresponding to the above monitoring point $\boldsymbol{u}$, the element of the matrix is written as follows:

$$Q_{19,10,25,8} = 0.934.$$

By normalizing $\boldsymbol{Q}$ with Equations (16) ~ (17), we obtain a relationship matrix R among the population, per-capita GDP, relative exposure and flood loss. For example, corresponding to the above monitoring point $\boldsymbol{u}$, the element of the relationship matrix is written as follows:

$$r_{19,10,25,8} = 1.$$

Our task is to infer the losses in the gap units in Table 2 using the relationship matrix R. First, using Equations (20) and (21), we change $\boldsymbol{z} = (z_1, z_2, z_3)$ into a fuzzy set in $U_1 \times U_2 \times U_3$ defined in Equation (26). For example, for gap unit g_4, the background data are $\boldsymbol{z}$ = (1419, 1947, 0.44), which derives a fuzzy input $\tilde{A}_4$ with memberships:

$$\cdots, a_{9,7,13} = 0.533,\ a_{9,7,14} = 1,\ a_{9,7,14} = 0.372, \cdots$$

Employing Equation (22) with relationship matrix R and input $\tilde{A}_4$, we obtain a fuzzy output $\tilde{B}_4$ with memberships:

$$\cdots, \mu_{B_4}(u_{46}) = 0.148, \mu_{B_4}(u_{47}) = 0.316, \mu_{B_4}(u_{48}) = 0.15, \cdots$$

Using the center-of-gravity method in Equation (23), from $\tilde{B}_4$, we obtain a crisp value w_{A4} = 2266178. Similarly, we infer the flood losses in the other gap units. Finally, we have Table 3 to make Table 2 complete.

Table 3. Estimated Flood Losses in the Gap Units by SLDR

Unit g_i	Loss w	Unit g_i	Loss w
4	2,266,178	53	3,355,138
16	1,226,910	77	5,101,474
32	3,250,729	86	2,230,053
40	7,437,407	101	1,044,825

Note: w-Flood loss (RMB Yuan)

7. Comparisons with Geographically Weighted Regression and BP Network

To demonstrate the advantages of the new information technique in filling the gaps caused by incomplete data, we compare it with GWR and BP network. In general, to fill the gaps caused by incomplete data, other common methods are less accurate than GWR and BP networks. For example, the accuracy of inverse distance weighted (IDW) interpolation is lower than that of GWR (Deng et al., 2018). Therefore, in this ar-

ticle, the suggested method is only compared with GWR and BP network.

7.1. Comparison with Geographically Weighted Regression

GWR modeling is based on the distance between geographical points. Equation (29) gives the basic GWR:

$$y(u, v) = b_0(u, v) + b_1(u, v)x_1 + \varepsilon(u, v) \quad (29)$$

where y is the dependent variable with a Gaussian distribution, x is the independent variable, u and v are the coordinates of the data, b_0 is the intercept term, b_1 is the coefficient to be estimated and ε is the random error term. Equation (30) describes multiple GWR:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{m} \beta_k(u_i, v_i)x_{ik} + \varepsilon_i \quad (30)$$

where y_i denotes the dependent variable, $\beta_0(u_i, v_i)$ is the intercept coefficient at location i, x_{ik} is the value of the k^{th} explanatory variable at location i, and $\beta_k(u_i, v_i)$ is the local regression coefficient for the kth explanatory variable. Furthermore, (u_i, v_i) denotes the Cartesian x and y point coordinates, and ε_i denotes the random location specific error term.

In GWR, u and v are the coordinates of the data, not the monitoring point in SLDR. The coordinates of a unit determine the relative exposure of the unit to the flood. This statement implies that the data in Table 2 include the geographic information of the units. We use GWR in Equation (30) to statistically regress the sample in Equation (24), where $m = 4$.

If we use GWR to process samples directly, data overflow occurs in the computer program used during the statistical calculation process. To avoid this problem, we divide z_1, z_2 and w in Table 2 by 10,000 and obtain z'_1, z'_2 and w', respectively. For example, in unit g_1, loss w = 656,548 divided by 10,000 results in w' = 65.6548.

Table 4. Estimated Flood Losses in the Gap Units by GWR

Unit g_i	Loss w	Unit g_i	Loss w
4	2,500,897	53	1,829,942
16	2,678,546	77	2,042,115
32	6,217,089	86	2,562,986
40	4,827,113	101	2,340,886

Note: w-Flood loss (RMB Yuan)

Using GWR to process z'_1, z'_2, z_3 and w', we obtain a regression function shown in Equation (31):

$$W' = -419.628 + 1774.305z'_1 + 975.237z'_2 + 145.3z_3 \quad (31)$$

Let $w = 10000w'$. Table 4 shows the estimates of the losses in the 8 gap units shown in Table 2.

Comparing Tables 3 and 4, we cannot judge which of SLDR and GWR is superior. A comparison between the root mean squared error (RMSE) values, calculated by Equation (32), of SLDR and GWR gives an answer. The RMSE statistic is available to evaluate the performance of an interpolation method:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(w_i - w_i)^2} \quad (32)$$

where w_i, w_i are the observed and the estimated value at the sampling point i (i = 1, 2, …, n); n is the number of the size of the sample used for the estimation. In our case, w and w are the observed and predicted losses, respectively, and n = 100. For the 100 observed units in Table 2, Table 5 shows the observed loss w, predicted losses w by using GWR and SLDR, respectively, and their errors.

The predicted values in Table 5 are quite different from the observed values. For example, $w_{1(\text{SLDR})}$ = 1,157,367 exceeds the observed value of 656548 by approximately 76%, and $w_{1(\text{GWR})}$ = 2500897 exceeds that value by approximately 106%. We might think that the error between the observed value and predicted value is still obvious. It is true. However, in reality, it is quite practical when the estimated value is not less than half of the true value nor more than double the true value. For example, with high-precision data, the assessment result of death in Ludian M_S6.5 earthquake, occurred in 2014, China, is has only reached the accuracy of 59.8% (An et al., 2015). Reducing the error from 106 to 76% (reducing the error by 28%) has clearly improved the estimation accuracy. Also, for some points, w (SLDR) are quite accurate, for example, w_{95} = 5126319, $w_{95(\text{SLDR})}$ = 5,130,228, the error is 3,909, only 0. 76%.

When we change the measurement of the losses from RMB Yuan to million Yuan, the values in Table 5 decrease. For example, 1,157,367 changes to approximately 1.16. In particular, we have:

$$\rho = \frac{RMSE_{\text{GWR}} - RMSE_{\text{SLDR}}}{RMSE_{\text{GWR}}} = \frac{1177703 - 456126.8}{1177703} = 0.613$$

The relative error ρ = 0.613 between GWR and SLDR means that, to fill the gaps in Table 2, the information diffusion technique suggested in this article reduces the error by approximately 60% compared with geographically weighted regression. This benefit indicates that SLDR is obviously superior to GWR.

7.2. Comparison with BP Network

To demonstrate that the suggested method is more general and robust than an ANN, we compare it with BP network, shown in Figure 3, with three input nodes, nine hidden nodes and one output node, trained by a normalized sample.

Since the output values of the sigmoid function for the BP network always fall in the interval [0, 1], we cannot directly train the BP network with the original sample X in Equation (24). Using Equation (33), we normalize ($x_{i,1}$, $x_{i,2}$, $x_{i,3}$, $x_{i,4}$) of X to be ($x'_{i,1}$, $x'_{i,2}$, $x'_{i,3}$, $x'_{i,4}$), i = 1, 2, ..., 100, which is used to train the BP network shown in Figure 3:

$$x'_{i,j} = \frac{c_{j,0} + c_{j,1}(x_{i,j} - a_j)}{b_j - a_j},\ i = 1, 2, \cdots, 100;\ j = 1, 2, 3, 4 \quad (33)$$

Table 5. Predicted Losses by Using GWR and SLDR and Their Errors

No.	Observed w	GWR Prediction $w_{(GWR)}$	GWR Error	SLDR Prediction $w_{(SLDR)}$	SLDR Error
1	656,548	2,500,897	-1,844,349	1,157,367	-500,819
2	2,476,669	2,678,546	-201,877	2,741,901	-265,232
3	6,755,590	6,217,089	538,501	6,771,059	-15,469
…	…	…	…	…	…
94	1,211,300	1,130,927	80,373	1,270,269	-58,969
95	5,126,319	3,699,460	1,426,859	5,130,228	-3,909
96	6,516,619	6,072,786	443,832	6,980,360	-463,741
97	4,047,061	3,516,832	530,229	4,060,035	-12,974
98	7,496,821	7,565,254	-68,432	7,544,761	-47,940
99	62,251	2,666,348	-2,604,097	219,693	-157,442
100	3,441,624	3,624,081	-182,457	4,099,178	-657,554
		$RMSE_{GWR}$ = 1177703		$RMSE_{SLDR}$ = 456126.8	

Note: w_{100} is not the loss in unit g_{100}, but is the loss in unit g_{108} in Table 2 because there are 8 gap units.

where $b_j = \max_{1 \le i \le 100} \{x_{i,j}\}$, $a_j = \min_{1 \le i \le 100} \{x_{i,j}\}$, $c_{j,0} = 0.1$ and $c_{j,1} = 0.8$ are employed to compress $x'_{i,j}$ into a smaller interval than [0, 1] so as not to include the points 0 and 1.

Let the momentum rate be $\eta = 0.9$ and the learning rate be $\alpha = 0.7$. After 23273 iterations, the normalized system error is 0.0009. More iterations show that it is impossible to significantly reduce the error. With the sample, we also trained a BP network with three hidden layers, however the error was also not significantly reduced.

Using the parameters in Equation (33), we normalize the background data of the gap units in Table 2 to be the inputs of the trained BP network for estimating the flood losses in the units. We inverse the results from the trained network into the primary universe by:

$$x_{i,4} = \frac{x'_{i,4}(b_4 - a_4) - c_{4,0}}{c_{4,1}} + a_4 \quad (34)$$

Finally, using the trained BP network, we obtain Table 6 to make Table 2 complete. The RMSE and relative error of the BP network are:

$RMSE_{BP}$ = 678225.1

$$\rho = \frac{RMSE_{BP} - RMSE_{SLDR}}{RMSE_{BP}} = \frac{678\,225.1 - 45\,6126.8}{678\,225.1} = 0.327$$

This means that, to fill the gaps in Table 2, SLDR reduces the error by approximately 33% compared with BP network.

Table 6. Estimated Flood Losses in the Gap Units by the BP Network

Unit g_i	Loss w	Unit g_i	Loss w
4	1,468,647	53	2,505,539
16	1,599,238	77	8,650,776
32	2,582,922	86	3,650,178
40	6,552,073	101	1,401,586

Note: w-Flood loss (RMB Yuan)

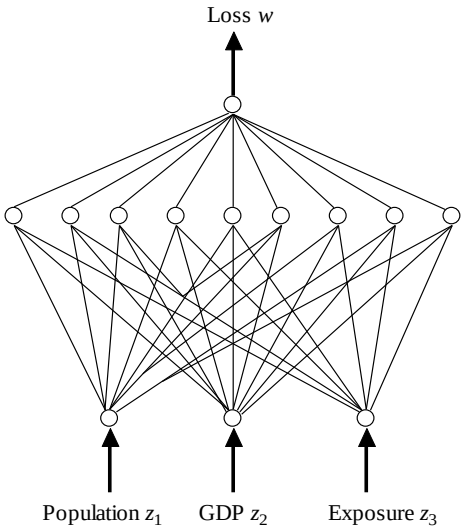

Figure 3. The architecture of a BP network trained by a sample consisting of observations obtained from observed units. It is a topology 3-9-1 BP network.

7.3. Q-Q Plots

In addition to using RMSE to judge the accuracy of a statistical prediction method, a quantile-quantile (Q-Q) plot shows a match between the estimated values and the expected (theoretical) values in detail through the reference line $y = x$.

As a probability plotting technique, a Q-Q plot is primarily suggested for testing whether a dataset follows a normal distribution (Stine, 2016), where n ordered sample values play the role of the sample quantiles and form the points of a Q-Q plot with theoretical quantiles of a normal distribution. In general, a Q-Q plot is a graphical technique for determining if two data sets come from populations with a common distribution. More generally, any pair of data can be in a Q-Q plot to show the difference between the two components through the reference line.

Let the loss w of a gap unit, obtained by using a nonlinear function with the background data and superimposed random

interference, be a target value. Let w be an estimated value obtained by using statistical prediction SLDR, GWR and the BP network. Table 7 lists the sorted target values and corresponding estimated values of the 8 gap units in Table 2, the sorting is in ascending order according to the target values. Figure 4 shows the Q-Q points of the 8 gap units.

Table 7. Sorted Target Values and Corresponding Estimated Vaues by SLDR, GWR and BP Network

Target Value w	$w_{(SLDR)}$	$w_{(GWR)}$	$w_{(BP)}$	Unit g_i
303,708	1,044,825	2,340,886	1,401,586	101
1,015,227	3,250,729	6,217,089	2,582,922	32
1,825,431	2,266,178	2,500,897	1,468,647	4
2,776,973	1,226,910	2,678,546	1,599,238	16
3,212,392	2,230,053	2,562,986	3,650,178	86
3,600,762	3,355,138	1,829,942	2,505,539	53
3,665,050	5,101,474	2,042,115	8,650,776	77
9,092,449	7,437,407	4,827,113	6,552,073	40

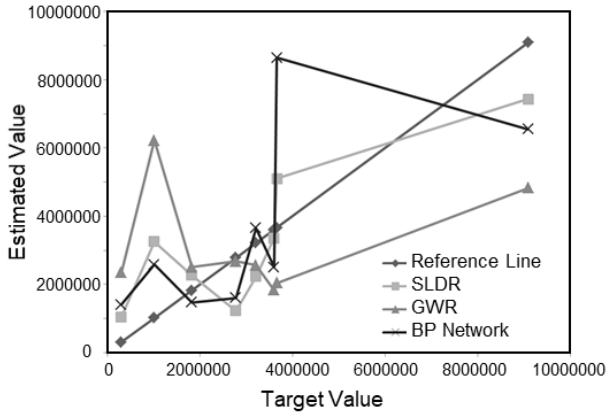

Figure 4. Q-Q plot to show the differences between the target values and estimated values of flood losses in the 8 gap units through the reference line $y = x$. The SLDR plot is near the line.

8. More Simulation Results

To substantiate the special case arguments based on Table 2, we need more numerical results to compare SLDR, GWR and the BP network with respect to the accuracy and validity of a statistical prediction by using a sample with noise. In this section, we give more simulation results. The simulation experiments are done by the following five steps.

Step 1: Generating random numbers

Running program MVN (Uebersax, 2006), we can easily generate random multivariate normal numbers. To a set of parameters serving for a probability distribution, different seed number will generate different set of random numbers. A simulation conclusion is believable if and only if it is based on many simulations with different pure random seed numbers.

Step 2: Modelling theoretical value of flood loss and superimposing random interference

Employing Equation (35) to model a theoretical flood loss, and Equation (36) to superimpose random interference to make the loss more realistic, we have a sample X in Equation (37).

$$w' = 0.59 z_1 z_2 [1 - \exp(\frac{-30 z_3}{7})] \tag{35}$$

$$w = w' + \delta \tag{36}$$

$$X = \{(z_{1i}, z_{2i}, z_{3i}, w_i) \mid i = 1, 2, \ldots, n\} \tag{37}$$

where z_1, z_2 and z_3 are to simulate population, per-capita and relative exposure, respectively, randomly generated by using program MVN. w' and δ are theoretical loss and random interference, respectively. δ is also randomly generated by program MVN. The subscript i in observation (z_{1i}, z_{2i}, z_{3i}, w_i) is to index the sequence number of a sample point. n is sample size determined by the researcher to simulate. The meaning of X in Equation (37) is same as the meaning of X in Equation (10), however the new expression is more targeted for this section.

Step 3: Taking training sample and validation data

Let the main part of X be a training sample and the other part of X be validation data.

Step 4: Calculating RMSE and RMSFE

Training SLDR, GWR and BP network by using above training sample, we have the RMSE of each model. Then, employing above validation data, we have the root mean squared forecasting error (RMSFE) of each model.

Step 5: Calculating average of simulation results and comparing them

Repeating above four steps with different seed numbers, we calculate the average of simulation results. Comparing average RMSE and RMSFE, we can compare the accuracy and validity of the three models, with respect to the sample X. Let A and B be two model trained by above training sample. When $RMSE_A > RMSFE_A$, model A is invalid to learning the given sample. When A and B are valid, if $RMSE_A < RMSE_B$, A is more accurate than B.

To simple expression, in this section, we use million RMB Yuan as the unit of flood loss w in the sample point to replace RMB Yuan in Table 2. According to our experience of flood disasters occurred in China, we use parameters in Table 8 to simulate a three-dimensional normal distribution with respect to population z_1, per-capita z_2 and relative exposure z_3.

Randomly selecting a seed number, such as 175, to run program MVN with parameters in Table 8 we generate a pseudo-random sample Z with 140 three-dimension points. Randomly selecting another seed number, such as 76453, to run program

MVN with mean $\mu = 0$ and standard deviation $\sigma = 100$, we generate a pseudo-random sample V with 140 one-dimension points to be random interference.

Table 8. Parameters of A Three-Dimensional Normal Distribution for Simulation Experiments

Random Variable	Mean	Standard Deviation	Covariance Matrix		
			z_1	z_2	z_3
z_1	2,000	600	1	-0.042	0.598
z_2	3,100	1,000	-0.042	1	-0.174
z_3	0.800	0.200	0.598	-0.174	1

Employing Equation (35) to the data of Z and employing Equation (36) with random interference V by point and point, we have a sample X' with 140 four-dimension points. From X', we select the first 108 points whose all components are non-negative and $z_3 \leq 1$ to form a set X shown in Equation (36) with background data z_1, z_2, z_3 and loss w on 108 geographic units in a flood disaster area. A part of the X generated with seed number 175 and 76,453 is shown in Table 9. Any interested reader can check if the data are really generated by MVN with the two seed numbers.

Because the data of X are randomly generated, we use the first 100 points to form a training sample, denoted as X (a, b), and other 8 points to form a validation set T (a, b), where a, b are seed numbers for randomly generating background data Z and interference V, respectively, i.e.:

$$X\ (175, 76453) = \{(z_{1i}, z_{2i}, z_{3i}, w_i) \mid i = 1, 2, \cdots, 100\}$$
$$= \{(3259, 3465, 0.53, 4.0523), (3533, 2533, 0.73, 6.5950),$$
$$\cdots, (2403, 1943, 0.86, 1.9627)\} \subset X$$

$$T\ (175, 76453) = \{(z_{1j}, z_{2j}, z_{3j}, w_j) \mid j = 1, 2, \cdots, 8\}$$
$$= \{(1669, 3154, 0.53, 3.5032), (674, 3551, 0.65, 0.2018),$$
$$\cdots, (3354, 1436, 0.73, 3.3601)\} \subset X$$

Training SLDR model composed of Equations. (13) ~ (23) by X (175, 76453), we have $RMSE_{\text{SLDR}} = 0.3861$. Using the trained SLDR to validation set T (175, 76453), we have testing error $RMSFE_{\text{SLDR}} = 1.0973$. Similarly, dealing with X (175, 76453) and T (175, 76453) by GWR and BP network, we have their RMSE and RMSFE shown in the third row of Table 10.

Randomly selecting another 9 groups of seed numbers and dealing with the samples and validation sets generated with the seed numbers, we have their *RMSEs* and testing errors shown in 4 ~ 12 rows of Table 10.

The averages in Table 10 could verify the performance of the studied models. It is interesting to note that:

$$RMSE_{\text{GWR}} = 1.1973 > 1.0480 = RMSFE_{\text{GWR}}.$$

It means that the error of a GWR after training with a sample is greater than the error without training. This is equivalent to saying that a tourist has travelled around Europe but not Africa, however the tourist's description of Africa is more precise than Europe. This is obviously very ridiculous. In other words, validity of GWR for filling gaps caused by incomplete is doubtful. The reason is the linear assumption in GWR model. Because

$$RMSE_{\text{SLDR}} = 0.3972 < 0.6904 = RMSE_{\text{BP}},$$

$$RMSFE_{\text{SLDR}} = 1.2562 < 1.5095 = RMSFE_{\text{BP}},$$

The results of 10 simulations show that SLDR is more accurate than BP network for filling the gaps. The reason is that the BP network does not converge when there is a random interference in the training sample.

The reason why the normal diffusion function used in SLDR is reasonable is that an information diffusion, as a simple process (without the aid of an intermediary) and without birth-death (in a seal system where the sum of information is kept 1), would obey the normal law, similarly as a molecule diffusion through a small unit (Huang and Shi, 2002).

There are many kinds of populations from that samples are drawn. They have performance between two shapes: the normal distribution and exponential distribution. The former is a symmetry curve; the latter is a monotone decreasing curve. In other words, if a mode is advantage for both of the normal distribution and exponential distribution, we can say that the model is absolutely advantage. Therefore, it is enough to simulate the two distributions. For our case, normal distribution is more reasonable to randomly model population, per-capita GDP, relative exposure and random interference.

The parameters of the normal distributions used in simulation experiment affect the values of *RMSE* and *RMSFE*. However, for the three statistical models used in our research, even we change the parameters, the conclusions based on simulation results in Table 10 will not change. In other words, the conclusions have general significance. Of course, for the information diffusion model SLDR, there is a lot of room for improvement in both the form of the diffusion function in Equation (13) and the diffusion coefficient in Equation (2).

9. Conclusions

Whether it is the study of the static phenomena on the earth surface, such as land use, or the study of dynamic phenomena, such as changes of natural disasters, people often encounter the problem of lack data on some geographic units. From the inverse distance weighted interpolation to geographically weighted regression (GWR), many models have been suggested to predict the lack data, however these models are not universal due to subject to continuous assumption or statistical forms. In theory, artificial neural networks, such as the back propagation neural network (BP network) which can be used to statistically predict lack data, are universal. However, the accuracy of BP network is not high when training sample is randomly interfered, due to convergence problem.

In this article, we develop an information diffusion technique, called self-learning discrete regression (SLDR), to infer

Table 9. Background Data *z* Randomly Generated with Seed Number 175 and Loss *w* with Random Interference

No.	z_1	z_2	z_3	Theoretical Loss	Random Interference	*w*
1	3,259	3,465	0.5300	5.9632	-1.9110	4.0523
2	3,533	2,533	0.7300	5.0512	1.5438	6.5950
…	…	…	…	…	…	…
100	2,403	1,943	0.8600	2.6868	-0.7241	1.9627
101	1,669	3,154	0.5300	2.7812	0.7220	3.5032
102	674	3,551	0.6500	1.3266	-1.1248	0.2018
…	…	…	…	…	…	…
108	3,354	1,436	0.7300	2.7142	0.6458	3.3601

Note: z_1 - Population; z_2 - Per-capita GDP (RMB Yuan); z_3 - Relative exposure; *w* - Flood loss (million RMB Yuan); Random interference data are generated with seed number 76453

Table 10. Results of 10 Simulations by SLDR, GWR and BP Network

Seed No.		SLDR		GWR		BP	
Sample	Interference	RMSE	RMSFE	RMSE	RMSFE	RMSE	RMSFE
175	76,453	0.3861	1.0973	1.1887	1.1953	0.6859	1.1367
374	61,473	0.4050	1.1164	1.1401	0.9235	0.6652	0.9801
6,423	4,863	0.4581	1.1094	1.2469	0.9315	0.9493	1.9815
6,503	624,352	0.4459	1.7739	1.3506	1.5853	0.7144	3.2710
9,126	6,453	0.4005	1.5569	1.0489	0.7874	0.6939	1.0386
35,267	48,329	0.4824	0.8509	1.3514	0.9792	0.6399	1.8307
37,091	543	0.2818	1.1211	1.1438	1.0803	0.5807	1.3851
86,754	573	0.3154	1.6627	1.2014	1.0974	0.6989	1.4153
291,347	9,861	0.4717	0.9290	1.1474	0.8141	0.6429	0.7206
679,341	397,454	0.3252	1.3441	1.1535	1.0855	0.6330	1.3353
	Average	0.3972	1.2562	1.1973	1.0480	0.6904	1.5095

Note: SLDR - Self-Learning Discrete Regression; GWR - Geographically Weighted Regression; BP - Back Propagation Neural Network; RMSE - Root Mean Squared Error; RMSFE - Root Mean Squared Forecasting Error

lack data. Since the author suggested the information dissemination technology 30 years ago, it can be used only in probability space. This article first develops it for geographic space. In this study, the geographic units that lack necessary data are called gap units. The units with necessary data are called observed units. The principle of SLDR is that, with the aid of background data on all units as the media, the model diffuses the information of observed units to gap units, and then infers lack data.

A virtual case based on China's flood experience is stud- ied, where flood losses of the gap units are inferred by using a relationship matrix with background data: population, per-capita GDP and relative exposure of the unit to flood. The result shows that the suggested model is a universal approximation. To this case, a comparison shows that SLDR is obviously superior to GWR. The new technique reduces the error by approximately 60%. It is also superior to BP network, reducing the error by approximately 33%.

To substantiate the special case arguments, ten simulation experiments are done with pure random seed numbers. The statistical average results show that (1) the validity of GWR for filling gaps caused by incomplete is doubtful; (2) SLDR is more accurate than BP network for filling the gaps. The validity of GWR is questionable due to its linear assumptions. The low accuracy of BP is due to that the model does not converge when there is a random interference in the training sample.

Geospatial information diffusion for filling gaps with the information of observed units is a new approach to supplementing incomplete spatial data to make the data complete. Information diffusion in probability space is an unconstrained diffusion. However, geospatial information diffusion in geographyical space is restricted by background data. SLDR using multiple normal diffusion is simply one option. In particular, if we can replace the relative exposure with coordinates, rivers, and terrain data, the suggested method is expected to be more effective.

In addition to the ability of the information diffusion techniques recognizing nonlinear systems with random interference, another reason for higher accuracy is that the techniques have the ability to optimally process small samples. Usually, if the type of a population from which a 1-dimensional sample is drawn is unknown, we need least 30 sample points for estimating the probability distribution of the population more accurately. Therefore, the support of a binary regression model requires a sample with a size of 900 (i.e., 30×30) if we do not know the type of input-output function that corresponds to the sample. When the sample size is small, the accuracy of SLDR is naturally higher than that of GWR and BP network.

Acknowledgements. This project was supported by the National Natural Science Foundation of China (41671502) and the National Key Research and Development Plan (2017YFC1502902).

References

An, J.W., Xu J.H., Nie, G.Z., and Bai, X.F. (2015). Earthquake disaster rapid assessment for emergency response supported by high-precision data of hazard bearing body. *Seismol. Geol.*, 37(4) 1225-1241. (in Chinese). http://dx.doi.org/10.3969/j.issn.0253-4967.2015.04.022

De Mesnard, L. (2013). Pollution models and inverse distance weighting: Some critical remarks. *Comput. Geosci.*, 52, 459-469. https://doi.org/10.1016/j.cageo.2012.11.002

Deng, Y., Liu, J.P., Liu, Y., and Xu, S.H. (2018). Spatial distribution esti- mation of PM2.5 concentration in Beijing by applying Bayesian geographic weighted regression model. *Sci. Surv. Map.*, 43(10), 39-45. http://dx.doi.org/10.16251/j.cnki.1009-2307.2018.10.006

Dette, H., Lopez, I.M., Rodriguez, I.M.O., and Pepelyshev, A. (2006). Maximin efficient design of experiment for exponential regression models. *J. Stat. Plan. Infer.*, 136(12), 4397-4418. https://doi.org/10.1016/j.jspi.2005.06.006

Eldrandaly, K.A. and Abu-Zaid, M.S. (2011). Comparison of six GIS-based spatial interpolation methods for estimating air temperature in western Saudi Arabia. *J. Environ. Inf.*, 18(1), 38-45. https://doi.org/10.3808/jei

Fotheringham, A.S., Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, John Wiley & Sons, Chichester.

Gutiérrez de Ravé, E., Jiménez-Hornero, F.J., Ariza-Villaverde, A.B., and Gómez-López, J.M. (2014). Using general-purpose computing on graphics processing units (GPGPU) to accelerate the ordinary kriging algorithm. *Comput. Geosci.*, 64, 1-6. https://doi.org/10.1016/j.cageo.2013.11.004

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5), 359-366. https://doi.org/10.1016/0893-6080(89)90 020-8

Huang, C.F. (1997). Principle of information diffusion. *Fuzzy Set Syst.*, 91(1), 69-90. https://doi.org/10.1016/S0165-0114(96)00257-6

Huang, C.F. (2002). Information diffusion techniques and small sample problem. *Int. J. Info. Tech. Dec. Mak.*, 1(2), 229-249. https://doi.org/10.1142/S0219622002000142

Huang, C.F. (2012). *Risk Analysis and Management of Natural Disaster*, Science Press, Beijing. (in Chinese).

Huang, C.F. and Moraga, C. (2004). A diffusion-neural-network for learning from small samples. *Int. J. Approx. Reason.*, 35, 37-161. https://doi.org/10.1016/j.ijar.2003.06.001

Huang, C.F. and Shi, Y. (2002). *Towards Efficient Fuzzy Information Processing - Using the Principle of Information Diffusion*, Physica-Verlag (Springer), Heidelberg, Germany. https://doi.org/10.1007/978-3-7908-1785-0_5

Huang, C.F., Wu, T., and Renn, O. (2016). A risk radar driven by internet of intelligences serving for emergency management in community. *Environ. Res.*, 148, 550-559. https://doi.org/10.1016/j.envres.2016.03.016

Kayacan, E., Ulutas, B., and Kaynak, O. (2010). Grey system theory-based models in time series prediction. *Expert Syst. Appl.*, 37(2), 1784-1789. http://dx.doi.org/10. 1016/j.eswa.2009.07.064

Lieske, D.J. and Bender, D.J. (2011). A robust test of spatial predictive models: geographic cross-validation. *J. Environ. Inf.*, 17(2), 91-101. https://doi.org/10.3808/jei.201100191

Liu, X.P., Zhang, J., Cai, W.Y., and Tong, Z.J. (2010). Information diffusion-based spatio-temporal risk analysis of grassland fire disaster in northern China. *Knowledge-Based Syst.*, 23(1), 53-60. https://doi.org/10.1016/j.knosys.2009.07.002

Makó, Z. (2005). Approximation with diffusion-neural-network. *Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence*, November 18-19, Budapest, 589-600.

Marcek, D. (2013). Risk scenes of managerial decision-making with incomplete information: an assessment in forecasting models based on statistical and neural networks approach. *J. Risk Anal. Crisis Response*, 3(1), 13-21. https://doi.org/10.2991/jrarc.2013.3.1.2

Otness, R.K. and Encysin, L. (1972). *Digital Time Series Analysis*. John Wiley, New York.

Purkait, B., Kadam, S.S., and Das, S.K. (2008). Application of artificial neural network model to study arsenic contamination in groundwater of Malda District, Eastern India. *J. Environ. Inf.*, 12(2), 140-149. https://doi.org/10.3808/jei.200800132

Sen, M. (2006). *Lecture Notes on Intelligent Systems. Department of Aerospace and Mechanical Engineering*, University of Notre Dame, IN 46556, USA.

Stavrou, I.D. and Ventikos, N.P. (2014). Ship to ship transfer of cargo operations: Risk assessment applying a fuzzy inference system. *J. Risk Anal. Crisis Response*, 4(4), 214-227. https://doi.org/10.2991/jrarc.2014.4.4.3

Stine, R.A. (2016). Explaining normal quantile-quantile plots through animation: the water-filling analogy. *Am. Statistician*, 71(2), 145-147. https://doi.org/10.1080/00031305.2016.1200488

Tobler, W.R. (1970). A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.*, 46, 234-240. https://doi.org/10.2307/143141

Uebersax, J.S. (2006). MVN program-Generate Random Multivariate Normal Numbers Easily, http://john-uebersax.com/stat/ mvn.htm.

Wray, J. and Green, G.G.R. (1995). Neural networks, approximation theory, and finite precision computation. *Neural Netw.*, 8(1), 31-37. https://doi.org/10.1016/0893-6080(94)00056-R

Zhang, Y., Xu, L.S., and Chen, Y.T. (2009). Spatio-temporal variation of the source mechanism of the 2008 great Wenchuan earthquake. *Chinese J. Geophys*, 52(2), 379-389. (in Chinese)

Zhao, S.J., Zhang, Q. (2012). Risk assessment of crops induced by flood in the three Northeastern provinces of China on small space-and-time scales. *J. Risk Anal. Crisis Response*, 2(3), 201- 208. https://doi.org/10.2991/jrarc.2012.2.3.7