Investigation of the Ability to Estimate Values of Road Section Condition Indicators Based on Their Spatial Correlation

Nam Lethanh, Ph.D.[1]; Craig Richmond, Ph.D.[2]; and Bryan T. Adey, Ph.D.[3]

Abstract: In the management of road networks, it is often desired to know the condition of individual road sections, which is approximated using the values of condition indicators. The values of these indicators can be used, for example, to determine whether an intervention should be executed on the road sections in the upcoming year, or to predict the future condition of the road sections. Unfortunately, a common problem when working with these data is that there are numerous road sections where no information is available. This can happen either due to errors made during the inspection campaigns themselves or due to using multiple independent sets of geographical information system (GIS) indexed data, when the sets are recorded as noncoincidental GIS shapes. It is of interest to the road manager to estimate the values of the missing condition indicators as best as possible. In this paper, an investigation of the ability to estimate values of road section indicators based on their spatial correlation is presented. The investigation is done by estimating the values of condition indicators for surface defects, and longitudinal and transversal unevenness exploiting the spatial correlation between them, on the Swiss national highway network. It is shown that the values of road section indicators can be estimated based on their spatial correlation with reasonably high levels of accuracy. The variation of the predictive ability per condition indicator is shown. **DOI: 10.1061/(ASCE)IS.1943-555X.0000290.**

Author keywords: Kriging method; Pavement management; Spatial correlation; Missing data.

Introduction

In the management of road networks, it is often desired to know the condition of individual road sections, which is approximated using the values of condition indicators. The values of these indicators can be used, for example, to determine whether an intervention should be executed on the road section in the upcoming year, or to predict the future condition of the road sections. The information collected during each inspection campaign is often stored in the form of geographical information system (GIS) indexed data.

Two common problems when working with these data are that there are some road sections where no information has been collected or stored, or where the roads have been divided into different sections in successive years. The former can happen due to errors made during the inspection campaigns themselves, for example, if the values of a condition indicator for one section are not measured, or are entered incorrectly into the database. The latter can happen due to poor planning or coordination. These two cases are illustrated in Fig. 1, where the recording of information from two inspection campaigns in which information was stored at two different sizes of road sections is shown in the upper portion of the figure, and the desired recording of information is shown in the bottom portion of the figure.

In Fig. 1, the first inspection campaign is considered to have taken place at time t, and the second inspection campaign is considered to have taken place at time $t + 1$. The road link extends from Point A to Point B. In the inspection campaign at time t, the collected information is recorded and attributed to each GIS shape representing a 200-m-long road section. In the inspection campaign at time $t + 1$, the collected information is recorded and attributed to each GIS shape representing a 100-m-long road section, which is not the same as in the first campaign. Additionally, no information is recorded for the third 100-m road section from Point A, which may be due to a malfunction in equipment. When a road manager would like to determine whether an intervention should be executed on the road section in the upcoming year or to predict the future condition of the road sections, it is useful to have harmonized and complete data sets, i.e., the GIS shape of the road sections should be the same irrespective of inspection campaign, and all road sections should have a value for the condition indicator. In either case, it is possible to speak of missing data. In the first case, the data are only missing because of the division of the road sections, and the second is truly missing data. In Fig. 1, they are referred to as potential missing data and missing data, respectively. In both cases, it would be beneficial to be able to create the complete data sets. This figure is only for illustrative purposes; in many practical cases, in either first or second inspection campaigns, the GIS shapes are often not equally the same (e.g., at time t and $t + 1$, the shapes can differ by more than 200 m or 100 m).

When there are missing, or potentially missing, condition indicator values, there are generally two situations that might exists. The first is when previous values exists from earlier or later inspection campaigns, e.g., the value in 2014 is missing, but the values in 2013 and/or 2015 are available. In this situation, many pavement

[1]Research Associate, Institute of Construction and Infrastructure Management, Swiss Federal Institute of Technology (ETH), 8093 Zurich, Switzerland (corresponding author). E-mail: lethanh@ibi.baug.ethz.ch

[2]Research Associate, Institute of Construction and Infrastructure Management, Swiss Federal Institute of Technology (ETH), 8093 Zurich, Switzerland. E-mail: richmond@ibi.baug.ethz.ch

[3]Professor, Institute of Construction and Infrastructure Management, Swiss Federal Institute of Technology (ETH), 8093 Zurich, Switzerland. E-mail: adey@ibi.baug.ethz.ch

Note. This manuscript was submitted on August 27, 2014; approved on October 29, 2015; published online on February 5, 2016. Discussion period open until July 5, 2016; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Infrastructure Systems***, © ASCE, ISSN 1076-0342.

Fig. 1. Graphical illustration of problems exist across inspection campaigns for a same road link

behavioral models (e.g., Chu and Durango-Cohen 2007, 2008; Nakat and Madanat 2008; Reger et al. 2013; Anastasopoulos and Mannering 2015) can be used to extrapolate past data. The second is when values have been recorded nearby, e.g., the value on Road Section 3 in Fig. 1 (inspection campaign at time $t + 1$) is missing but the values for Road Sections 1, 2, 4, and 5 are available. In this situation, other techniques can be used to interpolate the data (Al-Zoubi et al. 2015). For example, a missing value can be interpolated using the mean of the values for the two nearest road sections. In such as case, there may be a bias introduced because the two sections may not be the same distance away. As roads are longitudinally connected, the values of the indicators along the road will be spatially correlated if they are sufficiently close to one another (Kestler et al. 1994). One way to estimate missing data is, therefore, to postulate that the values of the condition indicators on the nearby sections and exploit the correlation between the values. One method to do this is a Kriging model, which can be used to estimate the missing value at a target point from optimally derived weights applied to observations at nearby points (Wackernagel 1998). The predicted target value is then calculated as a linear combination of the observed values. Importantly, the weights depend both on the spatial correlation and on the spatial distribution of the points.

In this paper, an investigation of the ability to estimate values of road section indicators based on their spatial correlation is presented. The investigation is done by estimating the values of condition indicators for surface defects, as well as longitudinal and transversal unevenness by exploiting the spatial correlation between them, on the Swiss national highway network. A univariate Kriging model is used. It is shown that the values of road section indicators can be estimated based on their spatial correlation with reasonably high levels of accuracy. The variation of the predictive ability per condition indicator is shown.

The rest of the paper is structured as follows. Section "Background" provides a description of the problems and accompanying research that has been done with respect to missing data in pavement management and the use of spatial correlation to estimate the missing values. Section "Methodology" provides the proposed methodology used in the investigation. Section "Example" provides a description of the investigation. The last section includes the conclusions and a discussion of possible future research directions.

Background

Problem and Relevant Research

Missing data are a common problem in many disciplines of science. Missing data can have a significant effect on the inferences made in, and the conclusions of, a study. In the pavement management context, missing data can occur due to numerous reasons, such as malfunctions in inspection equipment, interpretation mistakes by the inspector, inadequate amounts of time or resources to collect data everywhere desired, and aggregating data to objects differently in different inspection campaigns.

The focus of research in the past,when dealing with the problem of missing data in the field of pavement management, has been predominately focused on the estimation of parameters to be used in predictive models. For example, Ben-Akiva et al. (1993) and Ben-Akiva and Ramaswamy (1995) proposed an statistic approach to predict latent deterioration of infrastructure objects in situations when panel data sets were either complete or incomplete. Later, Chu and Durango-Cohen (2008) proposed the use of an AutoRegressive Integrated Moving Average model and a Kalman filter to deal with panel data sets containing missing data. Others, such as Hong and Prozzi (2006), Kobayashi et al. (2012), and Lethanh et al. (2015) have used a Bayesian approach for estimating the parameters of predictive models when there are incomplete or heterogeneous data sets.

Research work on reestimating the missing data is rare. Some of the first research in this area has been done by Farhan and Fwa (2013, 2015) who used a stochastic multiple imputation model, which is basically a linear regression model with a stochastic error term, to estimate missing data using other temporally correlated data. Other work has been done by Al-Zoubi et al. (2015), who proposed a so-called systematic statistical analysis to estimate

the missing data using a set of techniques such as using average or mean values of nearby data, or linear regression model similar to the work of Farhan and Fwa (2013, 2015). Some of the techniques proposed are to be used for temporally correlated data, and others for spatially correlated data, and some for both.

As summarized by Al-Zoubi et al. (2015) the techniques used to estimate the values of missing pavement condition indicators can be grouped in two categories: (1) model-free replacement techniques, and (2) model-based replacement techniques. The model-free replacement techniques do not require any mathematical model to estimate the missing data. Instead they rely on simple extrapolation and interpolation techniques. For example, if the values of a condition indicator of a road section is missing, the values may be estimated as the mean value of nearby road sections, by extrapolating the previous condition indicators values using a moving average (Al-Zoubi et al. 2015). Such estimation techniques, however, can have biases and therefore may not be the best way to predict the missing values.

The model-based replacement techniques require a mathematical model to estimate the missing data. Their use requires the definition of a time-dependent or parameter-dependent function of the values of the condition indicators. An example of such function is the *sigmoidal* function used by the Texas Department of Transportation (Stampley et al. 1995) and the exponential function used to capture the evolution of international roughness index over time (Paterson 1986). Values of parameters in the predefined functions are often estimated from available data by means of various regression techniques (e.g., linear regression, cubic regression, or cubic spline based on serviceability indicators) (Al-Zoubi et al. 2015).

Although the research made to date is beneficial with respect to estimating missing data, some improvements are possible with respect to the exploitation of the spatial correlation of the values of condition indicators on road sections. An interesting investigation into the spatial correlation of the values of condition indicators on road sections is that by Kestler et al. (1994), who used geostatistical analysis to estimate the spatial dependency of the values of the Falling Weight Deflectometer tests at various distances on a road.

Keeping this in mind, the methodology investigated in this paper for use to estimate values of road section condition indicators is one that exploits the spatial correlation of the values of the condition indicators on road sections; a univariate Kriging model has been chosen to estimate the values of characteristic variables at one location. The rest of this section includes a brief overview of the Kriging model, a brief discussion of the longitudinal nature of road data, and a brief refresher of spatial correlation.

Kriging Models

Kriging models are normally used to estimate the values of characteristics variables at one location, the estimation point, given similar values at surrounding locations. For example, a Kriging model can be used in environmental engineering to model how the degree of contamination of an unwanted substance varies over a given area of land (Cattle et al. 2002). In principle, a Kriging model is an interpolation based on a weighted average of surrounding data points. The weights are a function of their spatial covariance values, which themselves are assumed to be a function of distance. There are many good sources describing the actual calculation of Kriging weights; e.g., Wackernagel (1998), Diggle and Ribeiro (2001), and Stein (1999) to which the interested reader is referred so that the discussion here may be kept to a minimum. However, two basic effects are important to note. First, the closer a data point is to an estimation point, the larger its weight in the estimation becomes. Second, the closer two data points come to one another, the smaller the sum of their weights becomes, provided a third data point exists to which the extra weight can be transferred. This is consistent with the intuition that sampling twice at the same location does not add as much new information as sampling twice at different locations. The selection of the data points for interpolation of a given estimation point, referred to as clustering, plays an important role. The selection of the cluster size is a particular challenge. On one hand, the wider a cluster is, the more it includes points that have a small amount of predictive information. On the other hand, the narrower the cluster, the fewer observations are available on which to base a prediction. Some strengths of Kriging models are that they compensate for the effects of clustered data by assigning individual points within a close proximity less weight than isolated data points and thereby treat clusters more like single points, give estimates of estimation errors (Kriging variance) in addition to the estimates of the variables themselves, and provide a basis for stochastic simulation of potential realizations through the availability of estimation errors.

There are three main types of Kriging models: simple, ordinary, and Kriging with a trend. They primarily differ in their treatments of the trend component, $m(u)$ (e.g., assumption on a constant mean and nonbiasness over a random field). Within these types of Kriging models, they can be either univariate or multivariate, i.e., univariate Kriging models are used to estimate the value of an indicator at one geographical point using values of that indicators measured at neighboring points. In contrast, multivariate Kriging models are used to estimate the value of an indicator at one point by using not only values of that indicator measured at neighboring points but also values of other indicators measured at that points and neighboring points. More information can be found in Goovaerts (1997) and Wackernagel (1998). In principle, all types of Kriging models can be used to address the issue of missing data and determine the spatial correlation among data points.

Spatial Correlation in Kriging Models

Spatial correlation is expressed in Kriging models through a theoretical variogram. It expresses the functional relationship between distance and the variance of the difference between two random variables. Using the definitions of variance, covariance, and correlation one can express the relationship between the two random variables as follows:

$$\mathrm{Var}(a-b) = \sigma_a^2 + \sigma_b^2 - 2\sigma_a\sigma_b\rho_{a,b} \tag{1}$$

where σ and ρ = standard deviation and correlation coefficients, respectively. If one assumes that each random variable has the same variance, then Eq. (1) reduces to Eq. (2):

$$\mathrm{Var}(a-b) = 2\sigma_a^2(1-\rho_{a,b}) \tag{2}$$

This hypothesis is maintained here across all observations of the same indicator type and inspection campaign. A second maintained hypothesis is that the correlation between observations of any type is a stationary function of the distance h between the observations. Thus the theoretical variogram can be defined as

$$\mu(h) = \mathrm{Var}(h) = 2\sigma^2[1-\rho(h)] \tag{3}$$

and the higher the spatial correlation, the lower the variance of the difference. In contrast to correlation, variance is not a unitless number. Its absolute size is therefore, more difficult to interpret.

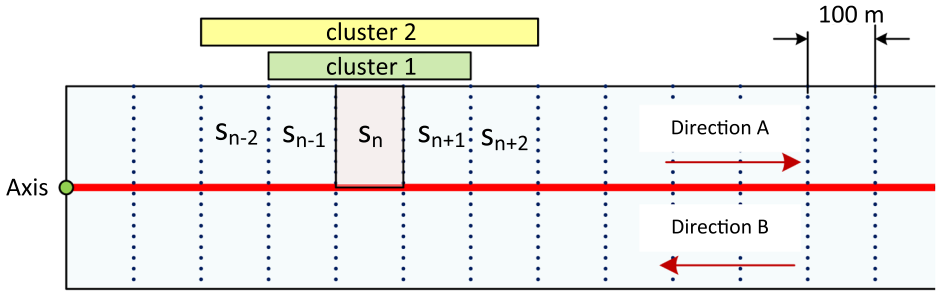

Fig. 2. Cluster sizes

Methodology

Steps

The methodology used in the investigation contains the following iterative steps.

Step 1
Define a set of clusters, i.e., the extent of the road section to be used to estimate the value of the target section, including the length of the target section. For example, in Fig. 2 where s_n is the target road section in the lane in which vehicles are traveling in Direction A, Cluster Size 1 is 300 m long, and Cluster Size 2 is 500 m long. This set needs to be selected from the smallest reasonable size, e.g., a 300-m-long road section, which encompasses two 100-m-long road sections on either side of the target road section, and reasonable steps, e.g., 200 m, to one where it is so large that it is not reasonable to believe that there would be any significant correlation between the values (e.g., 2,100 m).

Step 2
Define how the data will be selected from the cluster (e.g., the data selected from the cluster is all of the values of condition indicators attributed to the road section represented by the cluster, or only the values within specific ranges from the center of the road section represented by the cluster).

Step 3
Determine the correlation function. This is done first by plotting the experimental variograms for the condition indicators, whose values are to be estimated, i.e., the relationship between the distance from the center of the target section to each data point within the cluster (the lag differences) and the associated semivariance, selecting an appropriate form of the correlation function and then determining the values of the parameters of the correlation function to give the best fit with the experimental variogram, i.e., the theoretical variograms. The semivariance is calculated using the empirical semivariance function in Eq. (4):

$$\hat{\gamma}(h) = \frac{1}{2} \cdot \frac{1}{N(h)} \sum_{i=1}^{N(h)} [Z(u_\alpha + h) - Z(u_\alpha)]^2 \tag{4}$$

where $N(h)$ = number of data pairs at distance h (later herein referred as distance separation length).

Next, those estimates are fitted to a functional form to estimate the variogram. Some general functional forms for the correlation function that are often used are Cauchy, circular, cubic, gaussian, exponential, matern, and spherical (Wackernagel 1998; Diggle and Ribeiro 2001). These are referred to as the empirical and the theoretical variagrams, respectively. Theoretical variograms often contain a discontinuous jump at the origin and an upper bound where correlation between the observations has been reduced to zero, which are referred to as a sill (σ^2) and range parameter (ϕ). The correlation function, in which the covariance is prevented, can be written as

$$\mu(h) = 2\sigma^2 \left[1 - \exp\left(-\frac{h}{\phi}\right)\right] \tag{5}$$

Step 4
Derive the covariance function for each condition indicator from each inspection campaign for all road links in the network. Assuming variance is constant throughout the field, the covariance function has the following form:

$$C(h) = \sigma^2 \rho(h) \tag{6}$$

$$\rho(h) = \exp\left(-\frac{h}{\phi}\right) \tag{7}$$

Step 5
Determine the covariance matrix for condition indicator from each inspection campaign for all road links in the network. The covariance matrix expresses the covariance between each point in a cluster and every other point in the cluster. In this matrix cell i, j contains the value of this expression at the distance between points i and j. The elements on the diagonal are all equal to the variance of the detrended field. For the interest of reader, a rich literature describing various algorithms is given in the work of Wackernagel (1998).

Step 6
Determine the residuals between the values of each available condition indicator from each inspection campaign for all road links in the network estimated using the univariate Kriging model (described in the following subsection) and the values of the parameters corresponding to each condition indicator from each inspection campaign for all road links in the network and the actual values for each cluster size.

Step 7
Select the optimal cluster size, i.e., the cluster size that results in the lowest mean or standard deviation of residuals. If there is significant variation in the means of the residuals and little variation in the standard deviations of the residuals, then the cluster size that results in the lowest mean values is deemed the optimal cluster. If there is little variation in the means of the residuals but significant deviations in the standard deviations of the residuals, then the cluster size that results in the lowest standard deviations is deemed the optimal cluster. If it is not obvious which should be used, then weights need to be signed to both the mean and the standard deviation.

Table 1. Data Overview

I-value	Description	Number of data points 2000	2004	2009
I0	Surface damage without consideration of rut index	0	28,983	35,022
I2	Longitudinal unevenness	27,747	29,046	35,330
I3	Transversal unevenness	27,765	29,131	35,339

Step 8

Evaluate the ability of using spatial correlation to estimate the values of missing condition indicators.

Univariate Kriging Model

The univariate Kriging model used in this investigation, which is described in Goovaerts (1997) is of the simple types of Kriging models. In this model, the residual between the observed value and expected value is represented by $R(u) = Z(u) - m(u)$, where $Z(u)$ is treated as a random field with its expected value, $m(u)$. The residual between the expected value of the condition indicator at location u, $Z^*(u)$, is given by

$$Z^*(u) - m(u) = \sum_{\alpha=1}^{n(u)} \lambda_\alpha [Z(u_\alpha) - m(u_\alpha)] \tag{8}$$

where u, u_α = location vectors for the target point and one of the data points in a cluster, indexed by α; $n(u)$ = number of data points in the cluster that are used to estimate $Z^*(u)$; $m(u)$ and $m(u_a)$ = expected values of $Z(u)$ and $Z(u_\alpha)$; $\lambda_\alpha(u)$ = Kriging weights assigned to data points $z(u_\alpha)$ in the estimation of the target point u. The same data point will receive a different weight if it is used to estimate a different target point.

The objective of Kriging is to determine the weights, λ_α, that minimize the variance of the expected value of the condition indicator, $Z^*(u)$

$$\sigma_E^2(u) = \mathrm{Var}[Z^*(u) - Z(u)] \tag{9}$$

$$E\{Z^*(u) - Z(u)\} = 0 \tag{10}$$

After filtering out the trend, the residual random field $R(u)$ has a stationary mean of 0 and a stationary covariance function; that is, covariance is a function of the distance separation length h but not of position u.

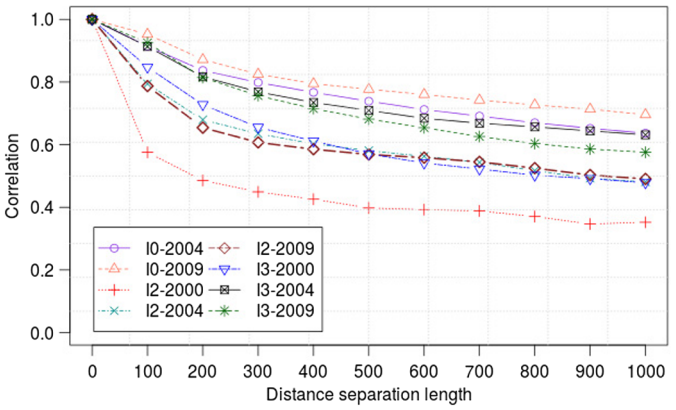

Fig. 3. Spatial correlation of serviceability indicators

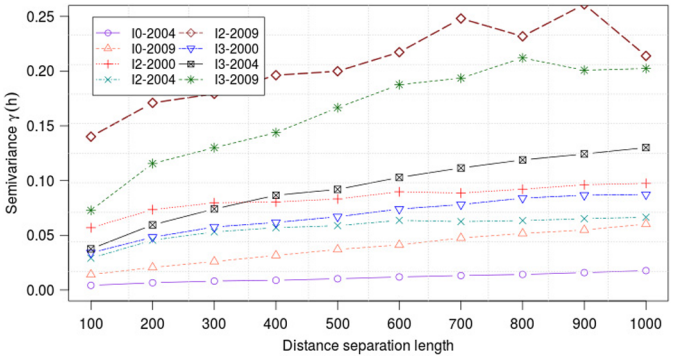

Fig. 4. Empirical variograms of condition indicators

Investigation

Background

In the investigation, the values of three condition indicators for surface defects, as well as longitudinal and transversal unevenness on the Swiss national highway network (Table 1) were estimated. Descriptions of I values in the table are defined in technical norm VSS 640925b (VSS 2003) in Switzerland. Starting with a complete data set, values were systematically removed and then reestimated using the methodology given in section "Methodology." The values of the condition indicators used were taken from three successive inspection campaigns conducted at five-year intervals. The values were determined using high-speed inspection vehicles with one value recorded to represent the average value for varying length of road

Table 2. Structure of the Database

Road identifier	Section identifier	Coordinates X	Y	Year	Inspection 1	Inspection 1 Inspection 2	…	Year	Inspection 1	Inspection 2 Inspection 2	…
1	1	x_1	y_1	2008	v_1^1	v_1^2	…	2012	…	…	…
1	2	x_2	y_2	2008	v_2^1	v_2^2	…	2012	…	…	…
1	3	x_3	y_3	2008	v_3^1	v_3^2	…	2012	…	…	…
2	1	…	…	2008	…	…	…	2012	…	…	…
2	2	…	…	2008	…	…	…	2012	…	…	…
2	3	…	…	2008	…	…	…	2012	…	…	…
2	4	…	…	2008	…	…	…	2012	…	…	…
…	…	…	…	2008	…	…	…	2012	…	…	…
…	…	…	…	2008	…	…	…	2012	…	…	…
R	…	…	…	2008	…	…	…	2012	…	…	…

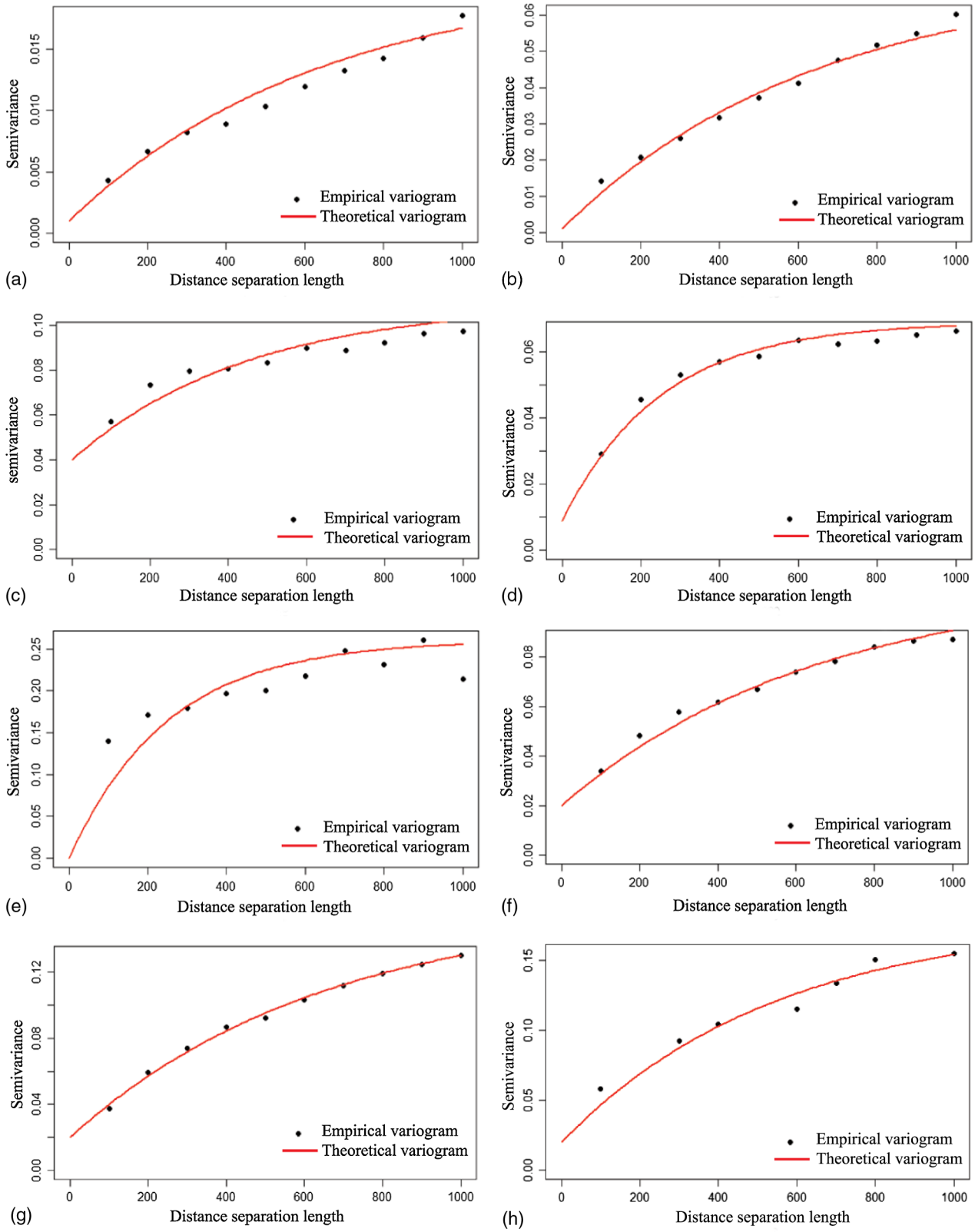

Fig. 5. Variograms for exponential model: (a) $\sigma^2 = 0.02$, $\phi = 650$; (b) $\sigma^2 = 0.07$, $\phi = 650$; (c) $\sigma^2 = 0.07$, $\phi = 450$; (d) $\sigma^2 = 0.06$, $\phi = 250$; (e) $\sigma^2 = 0.26$, $\phi = 250$; (f) $\sigma^2 = 0.09$, $\phi = 650$; (g) $\sigma^2 = 0.14$, $\phi = 650$; (h) $\sigma^2 = 0.16$, $\phi = 550$

section with different GIS shapes. These indicators are continuous within the range [0,5] where the best state is zero. Detailed standards exist on their measurement and yet achieving comparable measurements across inspection campaigns is not trivial. This is particularly true for the Indicator I0, which involves a visual assessment of the severity and extent of the damage to the road surface. The data were restructured as shown in Table 2 taking into consideration of reshaping the GIS shapes across inspection campaigns. The values are recorded by road direction, section, and lane (Fig. 2). Data elements for a section include the position

Table 3. Covariance Matrix ($s^2 = 0.02$ and $f = 650$)

h (m)	0	100	200	300	400	500	600	700	800	900	1,000
0	0.02000	0.01715	0.01470	0.01261	0.01081	0.00927	0.00795	0.00681	0.00584	0.00501	0.00429
100	0.01715	0.02000	0.01715	0.01470	0.01261	0.01081	0.00927	0.00795	0.00681	0.00584	0.00501
200	0.01470	0.01715	0.02000	0.01715	0.01470	0.01261	0.01081	0.00927	0.00795	0.00681	0.00584
300	0.01261	0.01470	0.01715	0.02000	0.01715	0.01470	0.01261	0.01081	0.00927	0.00795	0.00681
400	0.01081	0.01261	0.01470	0.01715	0.02000	0.01715	0.01470	0.01261	0.01081	0.00927	0.00795
500	0.00927	0.01081	0.01261	0.01470	0.01715	0.02000	0.01715	0.01470	0.01261	0.01081	0.00927
600	0.00795	0.00927	0.01081	0.01261	0.01470	0.01715	0.02000	0.01715	0.01470	0.01261	0.01081
700	0.00681	0.00795	0.00927	0.01081	0.01261	0.01470	0.01715	0.02000	0.01715	0.01470	0.01261
800	0.00584	0.00681	0.00795	0.00927	0.01081	0.01261	0.01470	0.01715	0.02000	0.01715	0.01470
900	0.00501	0.00584	0.00681	0.00795	0.00927	0.01081	0.01261	0.01470	0.01715	0.02000	0.01715
1,000	0.00429	0.00501	0.00584	0.00681	0.00795	0.00927	0.01081	0.01261	0.01470	0.01715	0.02000

(e.g., coordinates and axis distance), the date, and indicator values of different inspections. All road links were defined with distinct identification (e.g., link name). Each road section within a link was given a unique identification (e.g., road ID).

The spatial correlation of the values of the condition indicators is shown in Fig. 3, where the separation distance is the distance measured from the edge of the target road section to the edge of other road sections. The inspection campaign in the year 2000 did not include an observation for Indicator I0. The 2004 and 2009 campaigns measured the road sections using all three indicators. The lowest correlation values for both Indicators I2 and I3 occur in the year 2000. Barring these two observations, which probably reflect a less accurate inspection method, there is a rank order in spatial correlation where I0 > I3 > I2. That is, longitudinal unevenness, Indicator I2, appears to have the most spatial randomness, and surface defects, Indicator I0, the most spatial correlation. The values calculated at zero meters difference are deceptive since in fact, there is only one measurement at each point and the correlation with itself must be 1. If two independent inspections were executed at the same place, it is unlikely that the correlation would be 1. The evidence of spatial correlation declining in distance is nonetheless strong. In addition, the rate of decline levels off in all three cases after about 400 m and it remains clearly above zero even at a 1,000-m difference. Considering that the length of a typical intervention on a national highway is likely to be longer than 1,000 m, this result is to be expected due to shared histories of construction, maintenance, and use. An alternative, but not mutually exclusive cause, might be forward propagation of damage through the induced motion on passing vehicles.

Step 1: Define a Set of Cluster Sizes

The 10 selected cluster sizes had distance separation lengths of 100 to 1,000 m at 100-m intervals. A cluster with a length of 300 m and a 100-m-long target section has a distance separation length of 100 m.

Step 2: Define How the Data Will Be Selected from the Cluster

Two rules were used to select data from the clusters. In Scenario 1, all values of the condition indicators within a cluster were used. In Scenario 2, only the values that were farthest from the target section

Table 4. Scenario 1: Summary of the Standard Deviations and Means of the Residuals

	SD σ and means μ of residuals							
	I0		I2			I3		
h(m)	2004	2009	2000	2004	2009	2000	2004	2009
100	0.07795	0.14068	0.29075	0.21718	0.54079	0.21453	**0.22456**	0.30764
	$\mathbf{-8.94 \times 10^{-4}}$	1.64×10^{-3}	$\mathbf{-3.97 \times 10^{-4}}$	$\mathbf{-4.65} \times 10^{-4}$	-5.47×10^{-3}	-1.11×10^{-3}	-1.47×10^{-4}	$\mathbf{4.52 \times 10^{-3}}$
200	0.07726	0.13887	0.28355	0.21276	0.52899	0.21336	0.22777	0.31044
	-8.94×10^{-4}	1.59×10^{-3}	-1.41×10^{-3}	-4.70×10^{-4}	-5.60×10^{-3}	-9.71×10^{-4}	-1.27×10^{-4}	6.06×10^{-3}
300	0.07711	0.13847	0.2789	0.21147	0.51555	0.21248	0.22743	0.30682
	-9.06×10^{-4}	1.53×10^{-3}	-1.47×10^{-3}	-4.73×10^{-4}	-5.26×10^{-3}	-9.71×10^{-4}	-1.37×10^{-4}	6.64×10^{-3}
400	0.07706	0.13838	0.27686	0.21138	0.5159	0.21195	0.22698	**0.30431**
	-9.31×10^{-4}	1.52×10^{-3}	-1.48×10^{-3}	-4.75×10^{-4}	-4.86×10^{-3}	-9.71×10^{-4}	-1.17×10^{-4}	6.53×10^{-3}
500	**0.07705**	0.13837	0.27628	0.21047	0.5168	0.21156	0.22675	0.30557
	-9.39×10^{-4}	1.53×10^{-3}	-1.48×10^{-3}	-4.77×10^{-4}	-4.89×10^{-3}	-9.72×10^{-4}	-9.72×10^{-5}	6.30×10^{-3}
600	0.07709	0.1384	0.27542	0.20982	0.51407	**0.21146**	0.22651	0.30469
	-9.40×10^{-4}	1.55×10^{-3}	-1.50×10^{-3}	-1.37×10^{-3}	-4.86×10^{-3}	-9.72×10^{-4}	$\mathbf{-8.72 \times 10^{-5}}$	6.11×10^{-3}
700	0.07713	0.13842	0.27543	0.21031	0.51183	0.2117	0.22667	0.30612
	-9.61×10^{-4}	1.50×10^{-3}	-1.50×10^{-3}	-1.36×10^{-3}	$\mathbf{-4.74 \times 10^{-3}}$	-9.71×10^{-4}	-1.17×10^{-4}	6.47×10^{-3}
800	0.07711	0.13819	0.27552	0.20997	0.51427	0.21167	0.22668	0.30704
	-9.67×10^{-4}	1.45×10^{-3}	-1.50×10^{-3}	-1.36×10^{-3}	-4.78×10^{-3}	-9.71×10^{-4}	-1.27×10^{-4}	6.48×10^{-3}
900	0.07715	0.13824	0.27544	**0.2096**	0.51282	0.21151	0.22645	0.30671
	-9.78×10^{-4}	1.42×10^{-3}	-1.50×10^{-3}	-1.35×10^{-3}	-4.90×10^{-3}	-9.71×10^{-4}	-1.07×10^{-4}	6.53×10^{-3}
1,000	0.07716	**0.13817**	**0.27537**	0.20997	**0.51103**	0.21146	0.22656	0.30739
	-9.87×10^{-4}	$\mathbf{1.40 \times 10^{-3}}$	-1.48×10^{-3}	-1.35×10^{-3}	-4.95×10^{-3}	$\mathbf{-9.71 \times 10^{-4}}$	-1.37×10^{-4}	6.55×10^{-3}

Note: Values of standard deviations and means of residuals are shown in upper and lower part of each row, respectively. Values in bold are when they are converted to absolute values.

Table 5. Scenario 2: Summary of the Standard Deviations and Means of the Residuals

	Standard deviations σ and means μ of residuals							
	I0		I2			I3		
h (m)	2004	2009	2000	2004	2009	2000	2004	2009
100	**0.07795**	**0.14068**	**0.29075**	**0.21718**	**0.54079**	**0.21453**	**0.22456**	**0.30764**
	-8.94×10^{-4}	1.64×10^{-3}	-3.97×10^{-4}	-4.65×10^{-4}	-5.47×10^{-3}	-1.11×10^{-3}	-1.47×10^{-4}	4.52×10^{-3}
200	0.09842	0.19934	0.33684	0.26729	0.58848	0.26542	0.29054	0.41459
	-1.13×10^{-3}	1.65×10^{-3}	-1.69×10^{-3}	3.05×10^{-5}	-2.58×10^{-3}	$\mathbf{-1.05 \times 10^{-3}}$	1.40×10^{-4}	1.02×10^{-2}
300	0.10739	0.22484	0.34586	0.29056	0.57950	0.28863	0.32660	0.44583
	-1.24×10^{-3}	1.58×10^{-3}	-1.04×10^{-3}	5.78×10^{-5}	-5.01×10^{-4}	-1.31×10^{-3}	2.58×10^{-4}	9.72×10^{-3}
400	0.11649	0.23846	0.34815	0.29966	0.64494	0.30091	0.34765	0.48296
	-1.25×10^{-3}	1.64×10^{-3}	-1.40×10^{-3}	-9.08×10^{-5}	6.52×10^{-4}	-1.63×10^{-3}	-5.66×10^{-5}	7.71×10^{-3}
500	0.12422	0.25308	0.35496	0.29840	0.66850	0.31587	0.37111	0.51858
	-1.10×10^{-3}	1.39×10^{-3}	-1.44×10^{-3}	$\mathbf{-2.02 \times 10^{-6}}$	-8.76×10^{-4}	-1.68×10^{-3}	2.09×10^{-4}	4.89×10^{-3}
600	0.13485	0.26470	0.35504	0.30538	0.66808	0.33309	0.39533	0.52315
	-9.56×10^{-4}	$\mathbf{1.22 \times 10^{-3}}$	-1.68×10^{-3}	-4.66×10^{-6}	-1.79×10^{-3}	-1.74×10^{-3}	6.93×10^{-4}	3.63×10^{-3}
700	0.14333	0.27523	0.36930	0.31993	0.67440	0.35515	0.42239	0.54351
	-6.35×10^{-4}	1.31×10^{-3}	-1.73×10^{-3}	-1.53×10^{-4}	-2.73×10^{-3}	-1.65×10^{-3}	7.47×10^{-4}	$\mathbf{3.57 \times 10^{-3}}$
800	0.15000	0.28557	0.38290	0.32074	0.68903	0.36391	0.43648	0.56676
	-5.26×10^{-4}	1.40×10^{-3}	-1.48×10^{-3}	-3.78×10^{-4}	-3.44×10^{-3}	-1.39×10^{-3}	5.88×10^{-4}	3.99×10^{-3}
900	0.16131	0.29585	0.38907	0.32448	0.67490	0.36272	0.43648	0.58576
	-3.76×10^{-4}	1.34×10^{-3}	-1.38×10^{-3}	-5.77×10^{-4}	-4.04×10^{-3}	-1.52×10^{-3}	5.88×10^{-4}	4.39×10^{-3}
1,000	0.16908	0.30255	0.39459	0.33652	0.64653	0.36843	0.45262	0.60407
	-4.62×10^{-5}	1.57×10^{-3}	-1.16×10^{-3}	-7.46×10^{-4}	-4.60×10^{-3}	-1.36×10^{-3}	$\mathbf{-1.76 \times 10^{-5}}$	4.66×10^{-3}

Note: Values of standard deviations and means of residuals are shown in upper and lower part of each row, respectively. Values in bold are when they are converted to absolute values.

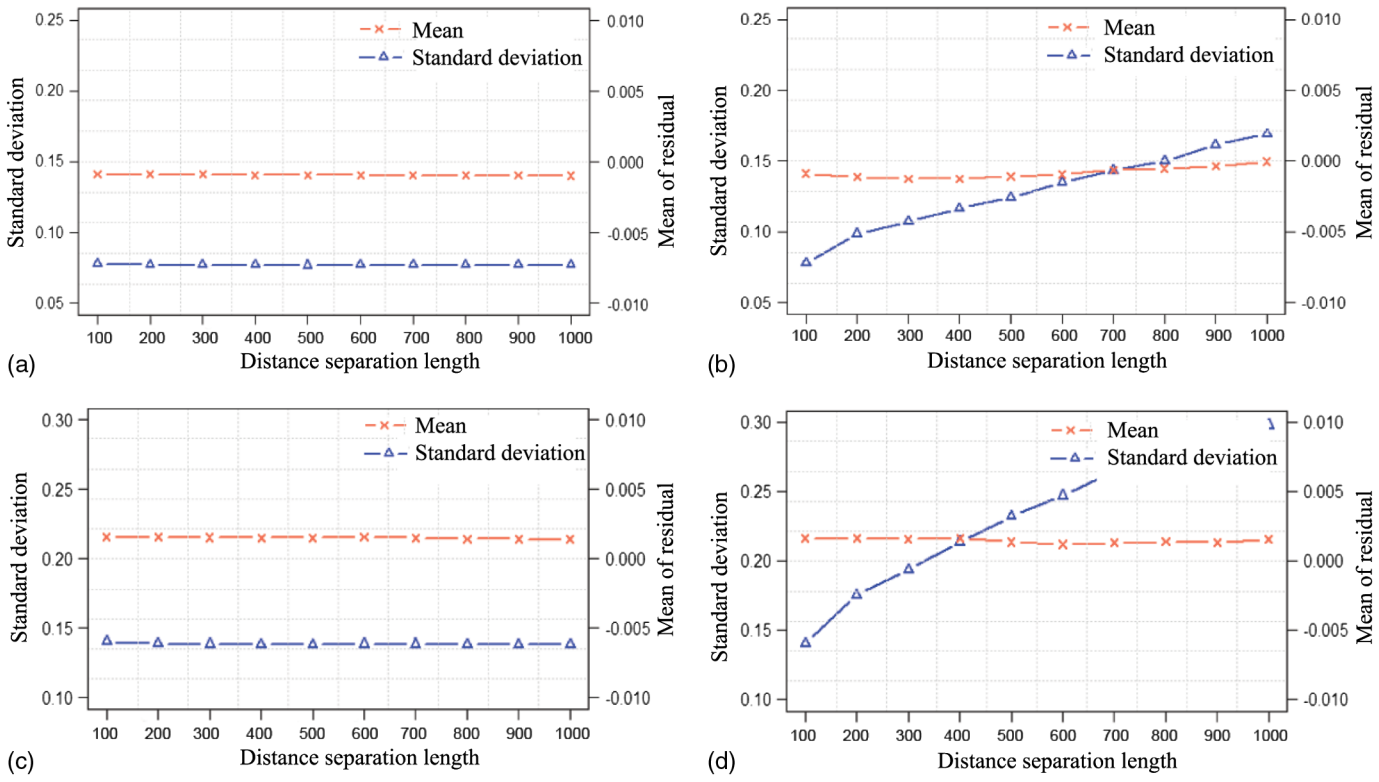

Fig. 6. Residuals-I0

were used. For example, with a cluster size of 500 m (a distance separation length of 200 m), in Scenario 1, the values of the condition indicators of the four road sections adjacent to the target road section are used, and in Scenario 2, only the values of the two road sections farthest from the target section were used.

Step 3: Determine the Correlation Function

The empirical variograms (Fig. 4) were drawn using all data for all condition indicators from all inspection campaigns using Eq. (4). These correspond to the correlation estimates in Fig. 3. It can be seen that the semivariance increases, and, therefore, the correlation

decreases, that larger the distance separation distance, until the observations are truly independent, which occurs once the experimental variograms are no longer increasing. The rank order in Fig. 4 is different than that in Fig. 3. This is because in Fig. 4, the correlation effect is scaled by the magnitude of the underlying variance of the random variables, whereas in Fig. 3 it is not.

Once the empirical variograms were determined, an exponential form for the correlation function was selected [Eq. (7)] based on a visual examination of the empirical variograms in Fig. 4 and the theoretical variograms were determined. The values of the parameters in each of the correlation functions for Condition Indicators I0, I2, and I3 for each inspection campaign are given, and the theoretical variograms are shown in Fig. 5.

Step 4: Derive the Covariance Function

The covariance function was then calculated for each road link in the network using Eq. (6). The values of the parameters are given in the figure for each condition indicator from each inspection campaign for all road links in the network.

Step 5: Determine the Covariance Matrix

The covariance matrix was determined for each condition indicator from each inspection campaign for all road links in the network using the correlation functions derived in Step 4. For example, for I2-2000, with its sill of $\sigma^2 = 0.02$ and range parameter of $\phi = 650$, the covariance matrix in Table 3 can be computed using Eq. (7).

Step 6: Determine the Residuals

Using the univariate Kriging model (subsection "Univariate Kriging model") together with the values of the parameters associated with each road link, the residuals were estimated for each condition indicator, for each road section on each road link using each of the clusters from the set of clusters defined in Step 1. In this example, the estimation was done using a standard optimization algorithm developed in *GeoR* package (Diggle and Ribeiro 2001, 2007). A summary of the means and standard deviations of the residuals for Scenarios 1 and 2 are given in Tables 4 and 5 and shown in Figs. 6–8. As can be seen, the variations in the mean values of residuals

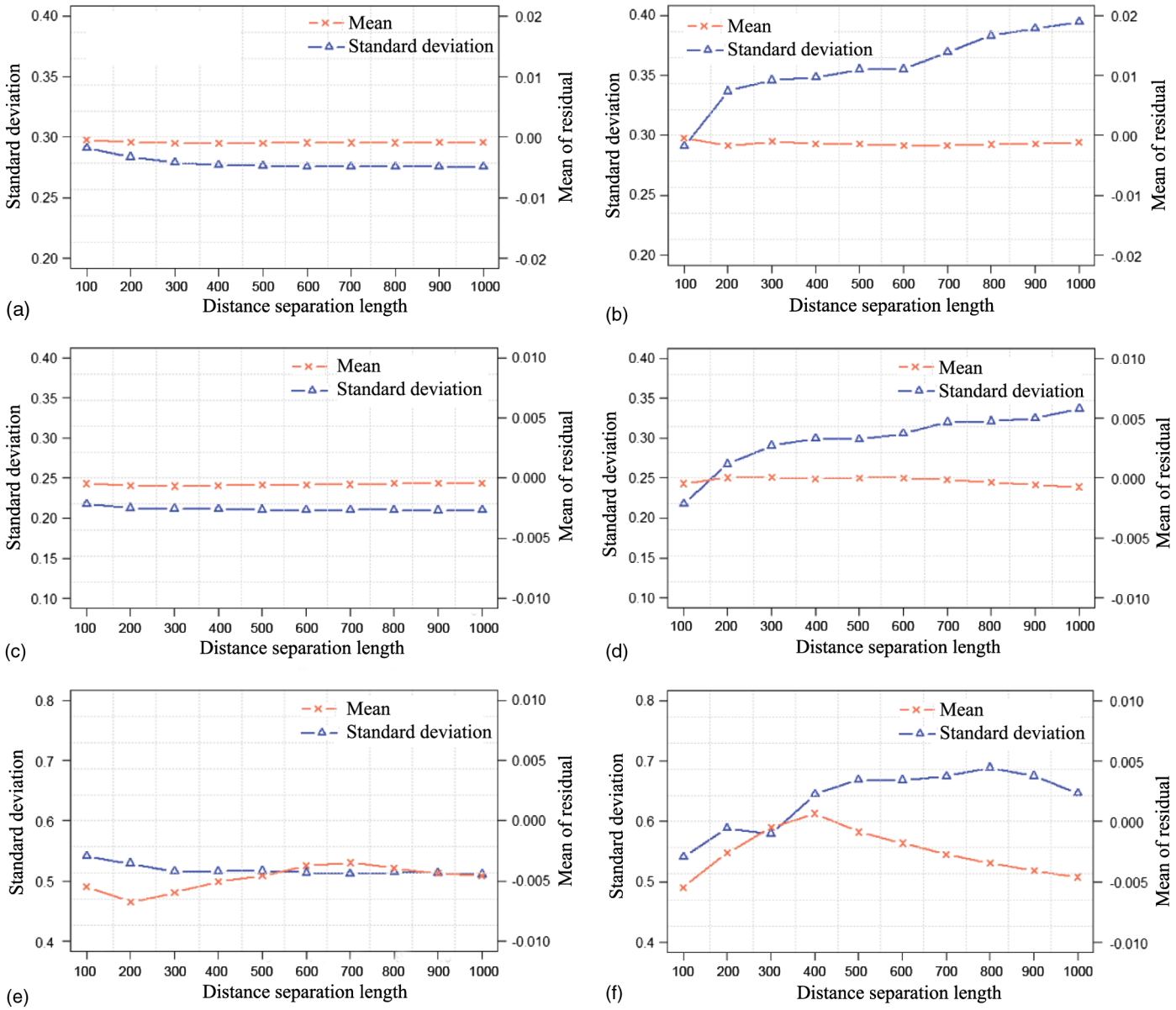

Fig. 7. Residuals-I2

in all cases are relatively small (<0.002), except for I2-2009 and I3-2009, but the values of standard deviations vary differently (e.g., value of standard deviation of residual of I2-2000 with distance separation length greater than 100 m varies about 10% compared to that with distance separation length of 100 m). For the purposes of discussion in a later section, values of standard deviations and means of residuals are shown in Tables 4 and 5 with their minimum values highlighted in bold.

Step 7: Select Optimal Cluster Size

The optimal cluster sizes if the lowest means of the residuals are used, and if the lowest standard deviations are used, are given in Tables 4 and 5 for Scenarios 1 and 2, respectively. They are not constant for all indicators, or for all inspection campaigns. For example, for Scenario 1, if the values of standard deviation of the residuals are used the optimal cluster sizes are 1,100 and 1,300 m for the inspection campaigns I0-2009 and I2-2000, respectively.

Step 8: Evaluate the Ability of Using Spatial Correlation to Estimate the Values of Missing Condition Indicators

Scenario 1: When All Data Points in the Clusters Were Used

When all data points in the clusters were used, it was found that the means of the residuals were close to zero and relatively constant. In this case, the standard deviation should be used to estimate the optimal cluster size.

The relationship between the standard deviations of the residuals and the cluster sizes (Figs. 6–8) can be grouped into two categories. The first category is one where there is little change in the standard deviations of the residuals as the cluster size changes, which is the case for I0 and I3 for all inspection campaigns. The second category is one where there is an initial decrease in the standard deviations of the residuals as the cluster size changes followed by a stabilization, which is the case for I2 for

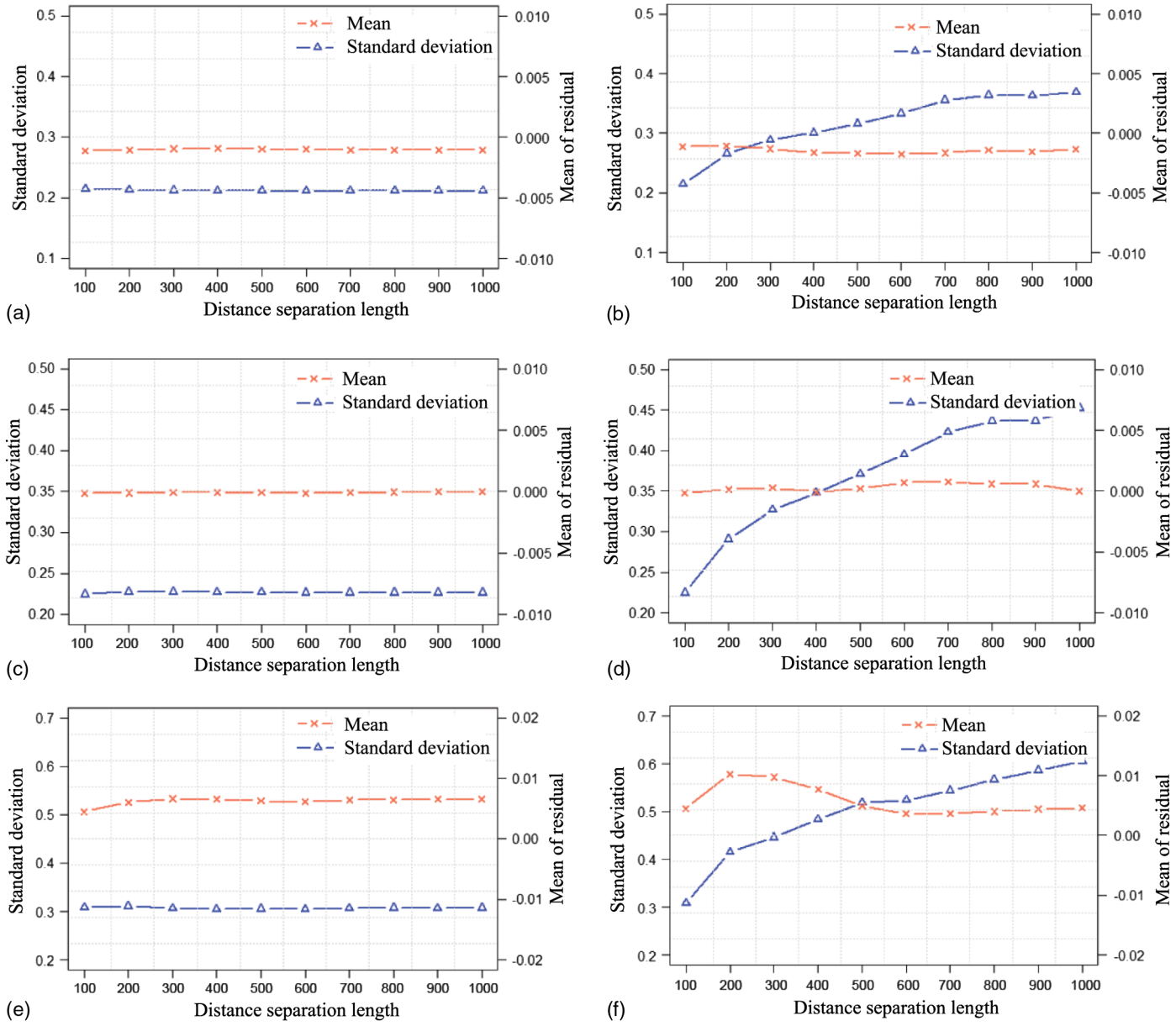

Fig. 8. Residuals-I3

all inspection campaigns (Table 4). In both categories, there is a decline in the value of standard deviation for indicators and inspection campaigns, with the exception for I3-2004, which are very slight (e.g., the value of standard deviation at a distance separation length of 100 m is 0.22456, which is smaller than the standard deviations corresponding to other distance separation lengths). The decrease of the standard deviations for I0-2004, I0-2009, I3-2000, and I3-2009 are within a range of 2%. This is, however, different for I2 for all inspection campaigns, which declines by 3 to 6% in all 3 years. The smallest standard deviations occur when the cluster sizes are between 900 and 1,300 m, which corresponds to distance separation lengths of 400 to 600 m, respectively.

By comparing the standard deviations of the residuals to the magnitude of the range of observed indicator values, it can be seen that 98% of all estimated values lie in the range 0 to 3, even though values of up to 5 were observed. Taking [0,3] as the effective range of values, a standard deviation of 0.3 is then exactly 10% of that range. In this case, only the standard deviations of I2-2009 are substantially above 0.3, having values near 0.5. Whether or not this is considered a good estimation depends, of course, on the application, but it is suspected to be better than simple interpolation and is certainly better than letting such missing values remain missing.

It can also be seen that significant differences exist in the standard deviations of the residuals for the same indicators for different inspection campaigns, e.g., Fig. 7 for Scenario 1. Given the sample size of nearly 30,000 measurements, it is unlikely this difference is due to sampling error. As the physical processes that cause the deterioration of the condition of the roads—and therefore the spatial correlation—are likely to have been relatively constant over the period of time in which the inspection campaigns were conducted, it is suspected that this variation is due to the changes in the inspection process itself. The exploration of the exact reason was beyond the scope of the work presented in this paper.

Scenario 2: When Only Edge Sectional Data Points Were Used

When only data from the edge sections of the cluster were used, the standard deviations of the residuals increase as the cluster size increases. This can be seen by the triangle marked lines in Figs. 6–8 (Scenario 2). This is not surprising because all intermediate values are dropped. The increasing standard deviation reflects the loss of correlation between measurements as the distance between them increases. The standard deviation of the residual increases by between 50 and 100% in all cases when the cluster size is increased from 300 and 2,100 m, which corresponds to the distance separation lengths from 100 to 1,000 m. It can also been seen that there is a degree of nonlinearity. The means of the residuals are largely unaffected by the distance between included observations, which confirms that the assumption that the means of the values of the indicators were unbiased in the use of the Kriging model was correct. In this case, the optimal cluster size is always 300 m.

Conclusion

In this paper, an investigation of the ability to estimate values of road section indicators based on their spatial correlation is presented. The investigation was done by estimating the values of condition indicators for surface defects, as well as longitudinal and transversal unevenness by exploiting the spatial correlation between them, on the Swiss national highway network. An univariate Kriging model was used. It is shown that the values of road section indicators can be estimated based on their spatial correlation with reasonably high levels of accuracy. The variation of the predictive ability per condition indicator is shown.

The results indicate that the use of univariate Kriging models is a viable way to estimate missing condition indicator values, which can be seen by the fact that the mean values of the residuals for all cluster sizes for all inspection campaigns for Scenario 1 were less than 0.0067 and those for Scenario 2 were less than 0.011. The standard deviations of the residuals, however, depending on the indicator and inspection campaign, can be relatively high.

As this investigation was limited to one kind of Kriging model, numerous extensions of the research are possible. Of particular interest is the investigation of the use of multivariate Kriging models that estimate values of one pavement condition indicator at a specific location using its own values in other locations and those of other pavement condition indicators. Additionally, future research should include investigations in the following areas:

- Determine the optimal distance between inspections.
- Evaluate the abilities of different inspection technologies.
- Identify potential measurement or data entry errors through a comparison with estimated values.

References

Al-Zoubi, M. M., Chang, C. M., Nazarian, S., and Kreinovich, V. (2015). "Systematic statistical approach to populate missing performance data in pavement management systems." *J. Infrastruct. Syst.*, 10.1061/(ASCE)IS.1943-555X.0000247, 04015002.

Anastasopoulos, P., and Mannering, F. (2015). "Analysis of pavement overlay and replacement performance using random parameters hazard-based duration models." *J. Infrastruct. Syst.*, 10.1061/(ASCE)IS.1943-555X.0000208, 04014024.

Ben-Akiva, M., Humplick, F., Madanat, S., and Ramaswamy, R. (1993). "Infrastructure management under uncertainty: Latent performance approach." *J. Transp. Eng.*, 10.1061/(ASCE)0733-947X(1993)119:1(43), 43–58.

Ben-Akiva, M., and Ramaswamy, R. (1995). "An approach for predicting latent infrastructure facility deterioration." *J. Infrastruct. Syst.*, 10.1061/(ASCE)1076-0342(1995)1:1(33), 33–43.

Cattle, J., McBratney, A., and Minasny, B. (2002). "Kriging method evaluation for assessing the spatial distribution of urban soil lead contamination." *J. Environ. Qual.*, 31(5), 1576–1588.

Chu, C., and Durango-Cohen, P. (2007). "Estimation of infrastructure performance models using state-space specifications of time series models." *Transp. Res. Part C*, 15(1), 17–32.

Chu, C., and Durango-Cohen, P. (2008). "Estimation of dynamic performance models for transportation infrastructure using panel data." *Transp. Res. Part B*, 42(1), 57–81.

Diggle, P. R., and Ribeiro, P. J. (2001). "GeoR: A package for geostatistical analysis." *R-News*, 1(2), 15–18.

Diggle, P. R., and Ribeiro, P. J. (2007). *Model-based geostatistics*, Springer, Berlin.

Farhan, J., and Fwa, T. (2013). "Airport pavement missing data management and imputation with stochastic multiple imputation model." *Transp. Res. Rec.*, 2336(1), 43–54.

Farhan, J., and Fwa, T. (2015). "Improved imputation of missing pavement performance data using auxiliary variables." *J. Transp. Eng.*, 10.1061/(ASCE)TE.1943-5436.0000725, 04014065.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*, Oxford University Press, NewYork.

Hong, F., and Prozzi, J. (2006). "Estimation of pavement performance deterioration using Bayesian approach." *J. Infrastruct. Syst.*, 10.1061/(ASCE)1076-0342(2006)12:2(77), 77–86.

Kestler, M., Harr, M. E., Berg, R. L., and Johnson, D. M. (1994). "Spatial variability of falling weight deflectometer data: A geostatistical analysis." *4th Int. Conf., Bearing Capacity of Roads and Airfields*, Minneapolis, 317–330.

Kobayashi, K., Kaito, K., and Lethanh, N. (2012). "A statistical deterioration forecasting method using hidden Markov model for infrastructure management." *Transp. Res. Part B: Methodol.*, 46(4), 544–561.

Lethanh, N., Kaito, K., and Kobayashi, K. (2015). "Infrastructure deterioration prediction with a poisson hidden Markov model on time series data." *J. Infrastruct. Syst.*, 10.1061/(ASCE)IS.1943-555X.0000242, 04014051.

Nakat, Z., and Madanat, S. (2008). "Stochastic duration modeling of pavement overlay crack initiation." *J. Infrastruct. Syst.*, 10.1061/(ASCE)1076-0342(2008)14:3(185), 185–192.

Paterson, W. (1986). "International roughness index: Relationship to other measures of roughness and riding quality." *Transp. Res. Rec.*, 1084, 49–59.

Reger, D., Christofa, E., Guler, I., and Madanat, S. (2013). "Estimation of pavement crack initiation models by combining experimental and field data." *J. Infrastruct. Syst.*, 10.1061/(ASCE)IS.1943-555X.0000148, 434–441.

Stampley, B., Miller, B., Smith, R., and Scullion, T. (1995). "Pavement management information system concepts, equations, and analysis models." *Rep. No. TX-96/1989-1*, Texas Transportation Institute, Texas Dept. of Transportation, TX.

Stein, M. (1999). *Interpolation of spatial data: Some theory for kriging*, Springer, Berlin.

VSS (Vereinigung Schweizerischer Strassenfachleute). (2003). "Swiss Standard SN-640925B-Erhaltungsmanagement der Fahrbahnen (EMF)." Zürich, Switzerland.

Wackernagel, H. (1998). *Multivariate geostatistics*, Springer, Berlin.