A NEW MULTI-SCALE MODELING APPROACH FOR SPACE/TIME RANDOM FIELD ESTIMATION

Kyung-Mee Choi [a, *], George Christakos [b], Mark L. Wilson [a]

[a]Dept. of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029, USA – kmchoi@umich.edu, wilsonml@umich.edu
[b]Dept. of Env. Sci. and Eng., University of North Carolina, Chapel Hill, NC, USA - george_christakos@unc.edu

Commission IV, WG IV/1

KEY WORDS: Multiple scale, Local scale, Measurement scale, random field, BME, Spatiotemporal, Estimate

ABSTRACT:

Modern geostatistical mapping methods are being applied to various types of data to produce more realistic and flexible characterizations of a natural random process. The Bayesian Maximum Entropy (BME) is a well-known geostatistical estimation method, especially for the use of soft knowledge as well as exact measurement data. Although development in geostatistical methods helps us to solve limitations on the format of available data, real studies always present *in situ* problems. Spatial scale of mapping (grid) points in a mapping model is usually not considered at the spatial scale of measurement data, especially in the studies that involve health-related data. Moreover, the spatial scale of measurement data may not be uniform, but varies among different measurements. For example, in studies of epidemiology or environmental health exposure, spatial scale of available measurement data is often limited and becomes different from the interesting spatial scale that is sought for in the estimation of the unknown random fields. It may be difficult and unrealistic to obtain measurement data at the scale of interest. Most current geostatistical methods have difficulty explaining physical phenomenon of unknown random fields over a continuous mapping domain at a scale smaller than that available from measurement data. This study explores how we can define these different scales in a geostatistical mapping model, and attempts to generate a meaningful spatiotemporal map of estimates of unknown random fields at the scale of interest. The estimation process of this study is based on the BME method to allow the probabilistic type of "soft" data, which are not actually observed, but simulated at the local scale to the measurement scale. This new modelling approach has been called multi-scale or local-scale mapping model. With actual mortality data collected over the 58 counties of the state of California, USA, we applied this multi-scale modelling approach, and obtained more accurate and realistic spatiotemporal maps of mortality rate estimates over California. We compared these estimates with those found by another approach that did not account for multiple scales on the same data. It was verified by actual mortality data obtained at the zip-code scale. These estimates found by the multi-scale approach were considered to be more accurate than those from the other modelling approach.

1. INTRODUCTION

Spatiotemporal mapping analysis had been originally introduced by D. G. Krige in 1951, but not until 1971 was it actively developed by G. Matheron. Today, geostatistical analysis is applied to various spatial problems, although it initially addressed the area of mining engineering. Recently, applications to health-related studies have been made, for example kriging estimation of epidemiological data on influenza-like illness in France (Carrat and Valleron, 1992) and spatio-temporal mapping based on BME (Bayesian Maximum Entropy) of mortality data in the state of California in USA (Choi et al., 2001a). When mapping analysis is applied to spatial data, especially related to disease patterns, we search for transmission mechanisms, possible risk factors, and other useful information for health risk management (Christakos and Hristopoulos, 1998). With data collected at different locations for a natural process, we can find variations and uncertainties in these natural processes. For example, a stochastic process can be characterized by the spatial and temporal variations that we find based on available data for that process. We can obtain insights that might inform natural variation in processes by understanding the spatial distribution in available data of these processes.

Different stochastic estimation methods may provide different estimation results depending on the assumptions and limitations of the methods. BME method allows more flexible data to be used in the estimation of a stochastic process than other mapping methods. In this procedure of BME estimation, we can use various types of uncertain knowledge called "soft" data, such as interval, probabilistic or functional types of available data (Christakos, 2000; Choi et al. 1998). Often, meaningful characterizations of a natural process may involve the appreciation of its spatiotemporal variation at multiple scales, rather than a fixed scale over a mapping domain of space. Rigorous description of the situation depends on the scale at which phenomena are considered, rather than the scale at which measurements are taken (Choi et al., 2001b). In epidemiological studies, we often are interested in the spatial distribution of epidemiological variation at a smaller spatial scale than that of the measurements which were made. However data often are collected at a larger scale in most surveillance systems or health studies. Moreover, observed data

* Corresponding author.

may be available only at arbitrary discrete locations rather than over a continuous spatial domain.

In this study, we introduce a recent development of a new modeling approach called multi-scale modeling (Choi, 2001) which accounts for multiple scales in the estimation of spatiotemporal random fields considering those data that were at limited scales of different discrete spatial areas. The multi-scale modeling approach is developed based on the BME estimation method, and simulates soft data at the scale of mapping interest not using the data actually observed at the limited scales from the natural field. From these features, the new approach can explain more realistic stochastic variation and find more accurate estimates for the random processes than other approaches that do not account for the multiple scales. We applied the multiple-scale approach to a real-world study that involves mortality data collected at the spatial scale of counties in the State of California, USA. We generated a map of estimated mortality rates using simulated data at the local scale.

First, we review the Spatiotemporal Random Field (S/TRF) representation of a natural process and the framework of BME estimation. Then we present the multi-scale modelling approach that considers phenomena at multiple scales. Finally, we apply this approach to mortality data obtained at the scale of counties of the State of California, USA. Realistic and meaningful maps of estimates are generated. This approach is verified using mortality data collected at an actually local scale to the counties but which was not used for the estimation. The multi-scale approach produced more accurate maps than those found by other approaches.

2. SPATIOTEMPORAL RANDOM FIELDS AND OVERVIEW OF THE BME METHOD

2.1 S/T Random Fields

To describe randomness of a natural process, we define a spatiotemporal (s/t) random field $X(\mathbf{s},t)$ that takes real values at points $\mathbf{s}=(s_x,s_y)\in R^2$ and $t\in T\subseteq R^1_{0,+}$ in a domain of two-dimensional space and time. In most studies, a random field is assumed isotropic and characterized by mean and isotropic covariance functions. Stochastic trend of the isotropic random field is explained by a mean function $m_x=E[X(\mathbf{s},t)]$ at any point in the domain of interest. Stochastic variation is explained by an isotropic spatiotemporal covariance function $c_x(r,\tau)=E[X(\mathbf{p})-m_x][X(\mathbf{p}')-m_x]$ that assumes correlation between random fields at any pair of points, $\mathbf{p}=(\mathbf{s},t)$ and $\mathbf{p}'=(\mathbf{s}',t')$ for $\mathbf{p},\mathbf{p}'\in R^2\times T$ depends only on the spatial and temporal lag distances $r=|\mathbf{p}-\mathbf{p}'|$ and $\tau=|t-t'|$ between the pair of points. In this study, Greek letter χ is used to represent for the value of a random variable, and small letter x is used to denote a random variable for the random field. Bold letter $\mathbf{x}$ is used to denote a vector of variables or values.

2.2 Overview of the BME Method

The BME estimation (Christakos, 1992) consists of three main stages: prior, case-specific, and posterior. Each has corresponding knowledge. Prior stage processes general knowledge ($\mathcal{G}$), such as statistical moments or underlying physical laws that we can obtain from past experience and studies about the process. In prior stage, a prior probabilistic density function $f_{\mathcal{G}}(\chi_{\text{map}})$, $\chi_{\text{map}}=(\chi_{\text{data}},\chi_k)$ is formulated based on available prior knowledge. The case-specific stage (also termed the specificatory stage) involves certain and uncertain types of data. A unique advantageous feature of the BME method is uncertain knowledge (called "soft" data) χ_{soft} which is considered in the estimation procedure as is data from exact measurements, or "hard" data χ_{hard}. Soft data can be any uncertain type of knowledge about the process such as probabilistic distribution functions or intervals. We denote case-specific or specificatory knowledge ($\mathcal{S}$) in terms of both hard and soft data as

$$\mathcal{S}\;:\;\chi_{\text{data}}=(\chi_{\text{hard}},\chi_{\text{soft}})=(\chi_1,\ldots,\chi_m).$$

Below are several typical types of soft data χ_{soft} formulated for the BME estimation.

Interval type : $\chi_{\text{soft}}\in\mathbf{I}=(I_{m_h+1},\ldots,I_m)$

Probabilistic type : $P_{\mathcal{S}}[\mathbf{x}_{\text{soft}}\leq\zeta]=F_{\mathcal{S}}(\zeta)$

Functional type : $P_{\mathcal{S}}[\mathbf{x}_{\text{soft}}\leq\zeta,x_{\text{k}}\leq\zeta_{\text{k}}]=F_{\mathcal{S}}(\zeta,\zeta_{\text{k}})$

In the posterior stage, both general and case-specific knowledge are combined to generate a posterior probabilistic density function $f_{\mathcal{K}}(\chi_k)$ as

$$f_{\mathcal{K}}(\chi_k)=A^{-1}\int_D d\Xi(\chi_{\text{soft}})f_{\mathcal{G}}(\chi_{\text{map}})$$

where $A=\int_D d\Xi(\chi_{\text{soft}})\int_D d\chi_k f_{\mathcal{G}}(\chi_{\text{map}})$ and $\chi_{\text{map}}=(\chi_{\text{hard}},\chi_{\text{soft}},\chi_k)$.

Table 1 shows integration domain **D** and operator $\Xi(\chi_{\text{soft}})$ in the posterior function according to specificatory knowledge. BME estimate is found by various selection rules such as mode or mean estimate from the posterior function $f_{\mathcal{K}}(\chi_k)$. Estimation uncertainty for BME estimate is explained by a mean squared error based on the posterior function.

$\mathcal{S}$	**D**	$\Xi(\chi_{\text{soft}})$
Interval	$\mathbf{I}$	χ_{soft}
Probabilistic	$\mathbf{I}$	$F_{\mathcal{S}}(\chi_{\text{soft}})$
Functional	$\mathbf{I}\cup\mathbf{I}_k$	$F_{\mathcal{S}}(\chi_{\text{soft}},\chi_k)$

Table 1. Integration domain **D** and operator $\Xi(\chi_{\text{soft}})$

3. MULTI-SCALE MODELING APPROACH

When we estimate unknown random fields, stochastic variation of the random fields is derived from available measurement data at the scale that the measurements have been made rather than at the scale that is estimated. The scale of estimation points is taken uniformly and at a sufficiently fine scale that we can consider those estimation points to cover a continuous mapping domain. However, regardless of the scale we are interested in for the estimation of a process, the spatial scale of measurement data is exclusively dependent on the scale of available data. The scale of measurement data is usually large compared with the scale of estimation points and may not be uniform among the data. For example in the mortality study over the state of California (Choi et al., 2001a), mortality data were collected at the scale of counties that are different from one another in their sizes and shapes. The new multi-scale modelling approach (Choi, 2001) takes it into consideration not only the scale of estimation points, but also the scale of measurement data so that it can account for the multiple scales for the grid points of estimation and data of a natural process over a continuous mapping domain. Rigorous and more realistic description of stochastic variation becomes obtainable at the scale that phenomena are considered for the estimation of random field by the multi-scale modelling approach.

Let us denote a random field defined at a measurement scale as $Z(\mathbf{s},t)$, and a random field defined at the local scale to the measurement scale as $X(\mathbf{s},t)$ for a natural process of interest. These two random fields can be explained by each other as in Eq. (1).

$$Z(\mathbf{s},t) = \| D(\mathbf{s}) \|^{-1} \int_{\mathbf{u}\in D(\mathbf{s})} d\mathbf{u} X(\mathbf{u},t) \qquad (1)$$

where $\| D(\mathbf{s}) \|$ is the size of $D(\mathbf{s})$. Assuming that the random fields are isotropic, we can easily derive a same mean function $m_z = m_x$ for the random fields as each other from the equation above. Let c_z and c_x be denoted for isotropic s/t covariance functions respectively for the random fields defined at measurement scale and at its local scale. These covariance functions are also related to each other as shown below in Eq. (2), which is derived from Eq. (1) and the definition of isotropic covariance functions. For spatial and temporal lags $r = | \mathbf{s} - \mathbf{s}' |$ and $\tau = t - t'$, it is obtained

$$c_z(r,\tau) = \| D(\mathbf{s}) \|^{-1} \| D(\mathbf{s}') \|^{-1} \sum_{\mathbf{u}\in D(\mathbf{0})} \int_{\mathbf{u}'\in D(r)} d\mathbf{u} d\mathbf{u}' c_x(|\mathbf{u}-\mathbf{u}'|,\tau) . \qquad (2)$$

We cannot find the covariance c_x directly from measurement data since c_x represents local stochastic variation for random field $X(\mathbf{s},t)$. With an initial guess for c_x, we obtain c_z from the relation in Eq. (2). And c_z can approach the experimental covariance at the measurement scale, which we obtain from the measurement data. Change models for c_x by trial-and-error so that c_z obtained by Eq. (2) from c_x becomes close enough to the experimental variation obtained from the measurement data. Once we have found the isotropic characteristics for the local random field $X(\mathbf{s},t)$, we simulate a set of local scale random fields $\{ X(\mathbf{u},t), \forall \mathbf{u} \in D(\mathbf{s}) \}$ to the measurement scale of $Z(\mathbf{s},t)$. By making a large enough number of simulations, we apply the central limit theorem so that any initial probabilistic distribution can be accepted for the simulation of $X(\mathbf{u},t)$. With the isotropic property of $X(\mathbf{s},t)$, the multi-scale approach takes local data that are distance-dependent between estimation and data points from the simulated random fields $\{ X(\mathbf{u},t) \ \forall \mathbf{u} \in D(\mathbf{s}) \}$ rather than taking one of the simulated values at a certain point (e.g., a centroid). Then, we find an estimate by applying the BME method with the knowledge of covariance and simulated data (not observed) at the local random fields. Local scale data can be of any type that the BME method can process (as explained earlier). In the next section, we apply this multi-scale approach to real mortality data and generate a map of estimates over the state of California. Verification of the estimated results follows.

4. THE CALIFORNIA MORTALITY STUDY

We applied the new multi-scale approach for mapping analysis to mortality data from the Statistics Health Department of the State of California. Mortality is a more certain indicator variable that can be useful in studies of environmental exposure or epidemiological risk. The spatial scale of mortality data was available only at the resolution of California counties (N=58). We aimed to obtain a map of estimated mortality rates at a uniform and smaller scale by applying the multi-scale approach. Mortality data collected at the scale of zip-code were not used for the estimation, but allowed us to later verify the estimation results based on the county scale.

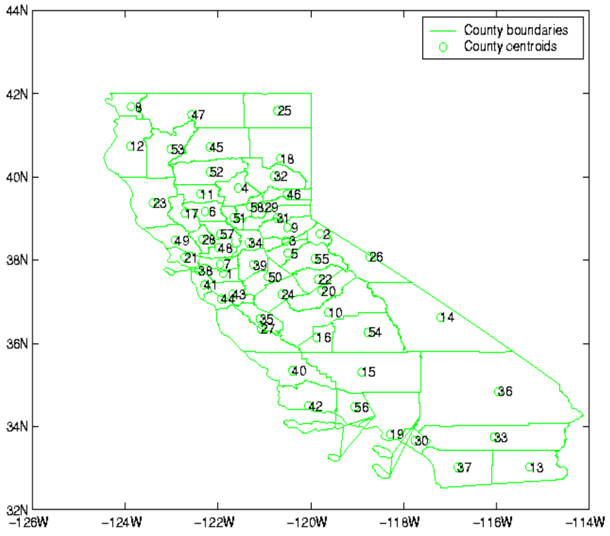

Figure 1. County map with centroids (circles) and identity numbers according to the order of county names.

4.1 Mortality Data

From the data set of death counts collected over the 58 counties of California during 1989, we selected death records only for California residents that were identified by ICD-9 (International

Classification of Death) code under 800. The total number of selected records for this study was 219,182. The map of California in Figure 1 shows county boundaries and county identity numbers on the county centroids according to the order of county names.

We obtained mortality rates by dividing the number of deaths by the population of each county, and used them as measurement data to generate estimated values. Using a smoothing technique (Choi, 2001; Choi et al., 2001a), we obtained space/time trends over these mortality rates. By leaving the trends out, we obtained a residual process of mortality rate $Z(\mathbf{s},t)$ at the county scale, which was assumed to be isotropic. Figure 2 shows the temporal trend (lined) over the mortality rate data (with circles) during the study year in the unit of days.

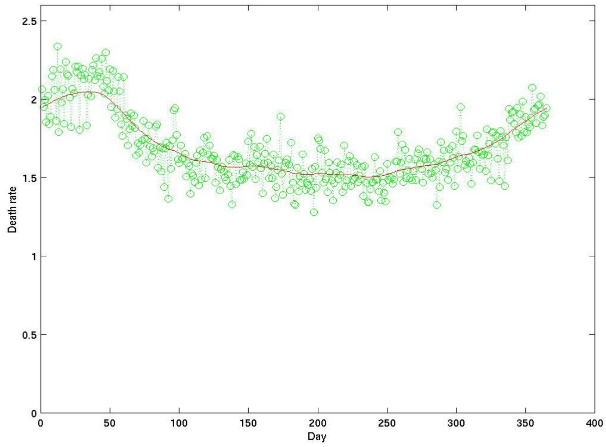

Figure 2. Mortality rates (with circles) and the temporal trend (lined) for the counties of 500,000 residents in California (deaths per 100,000 people per day).

4.2 Knowledge at the Local Scale

By applying the multi-scale modelling approach for the estimation of mortality rates over California, we first obtained experimental covariance at the county scale, shown in Figure 3 with circles. We approached the experimental county scale covariance with the covariance model $c_z(r,\tau)$ at the county scale obtained from various covariance models $c_x(r,\tau)$ at the local scale by trial-and-error as explained above. The nested exponential model in Eq. (3) is an optimal covariance function for $c_x(r,\tau)$ that we finally obtained at the local scale:

$$c_x(r,\tau)=\sum_{i=1}^{2} c_i \exp(-\frac{3r}{a_{r,i}})\exp(-\frac{3\tau}{a_{t,i}}) \qquad (3)$$

where c_1=8.09, $a_{r,1}$=1.14, $a_{t,1}$=43.5, c_2=4.37, $a_{r,2}$=51, $a_{t,2}$=990. From the values of c_1 and c_2, we can see that the covariance model $c_x(r,\tau)$ in the above is characterized more by the first nested structure that by the other. The second nested structure of $c_x(r,\tau)$ has a wider range both in space and time than the first nested structure. There was an additional third term in finding the $c_x(r,\tau)$, which was so small and ignored in Eq. (3).

Figure 3 shows the optimal covariance model $c_x(r,\tau)$ at the local scale and experimental covariance together with its fitting model $c_z(r,\tau)$ at the county scale. In the Figure 3, different covariance functions are shown on different y-axes. The left y-axis is used for the covariance at the county scale, and the right y-axis is for the covariance $c_x(r,\tau)$ at the local scale.

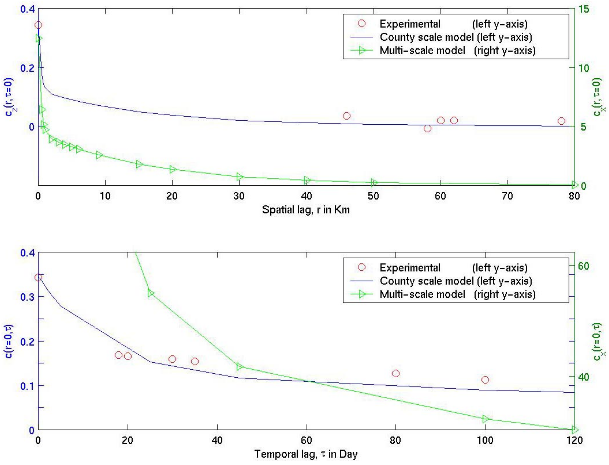

Figure 3. Isotropic spatiotemporal covariance of mortality rate (deaths per 100,000 people per day)^2 as the functions of spatial lag (top plot) and temporal lag (bottom plot).

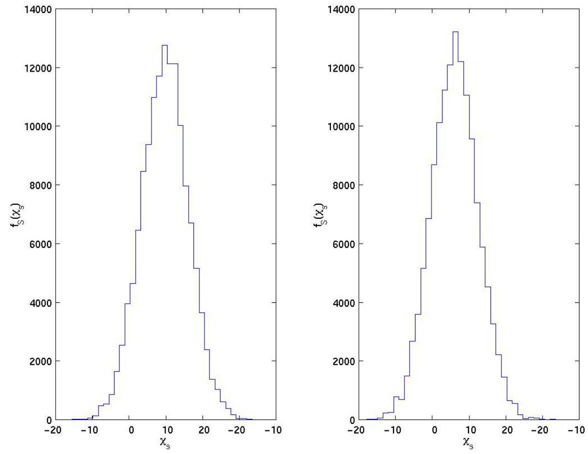

Figure 4. Probabilistic Soft data (in deaths per 100,000 people per day) at zero-lag of local distance between estimation and local data points to (a) county Alameda and (b) county Amador on January 1, 1989.

We simulated 10,000 local-scale random fields to each county and took the probabilistic type of soft data of the distance-dependent simulated values. Figure 4 shows the soft data for the residual process of mortality rates per 100,000 people per day at zero-lag of local distance between estimation and data points to county Alameda and county Amador on January 1, 1989. Since the soft data in Figure 4 is for the residual process of mortality rates, it ranges from negative to positive. With the soft data and covariance obtained at the local scale, we applied the BME estimation method and found mortality rate estimates throughout California.

4.3 Estimation Result and Verification

Using this multi-scale approach, we obtained estimates of mortality rates over a continuous spatial domain of California at any day during the study year. Figure 5 shows the spatial distribution of estimated mortality rates on Jan. 1, 1989. In this map, same mortality rates are shown on a same contour line. We compared these estimates with those obtained by a different modelling approach not accounting for the multiple scales. Figure 6 shows the spatial distribution of the estimates of mortality rate by a measurement scale (county scale) approach based on the same data set (Choi et al., 2001a). Figure 5 appears to show a more realistic spatial variability than in Figure 6 obtained by the measurement scale approach.

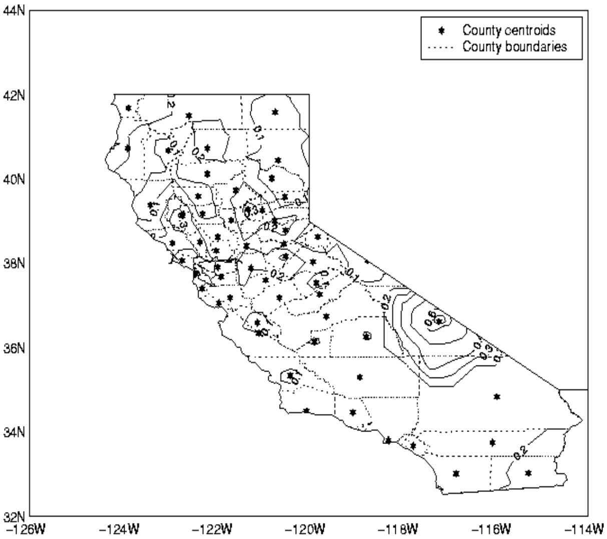

Figure 5. Spatial distribution of mortality rate estimates over California on Jan. 1, 1989 obtained by the multi-scale modelling approach (deaths per 100,000 people per day).

Then, actual mortality data were used at the zip-code scale to determine which modelling approach found closer estimates. To exclude computational error caused by mortality rates not relating to any modelling approach, we calculated the average of estimation errors that were above a threshold value θ representing the computational error. We compared in several ways the estimated mortality rates with others based on the average error as a function of θ, which were obtained by two different modelling approaches at the county scale and the corresponding local scale. We calculated the percentage of reduction in average error obtained by the multi-scale approach with respect to the average error obtained by the measurement scale approach. In Figure 7 (top), we see the values are all negative, suggesting as much reduction by the multi-scale approach as the percentage value. As we can see (Figure 7), the reduction increases as the threshold θ increases. This tells us that, as the average estimation error becomes larger, we can expect more reduction of the average estimation error by applying the multi-scale approach.

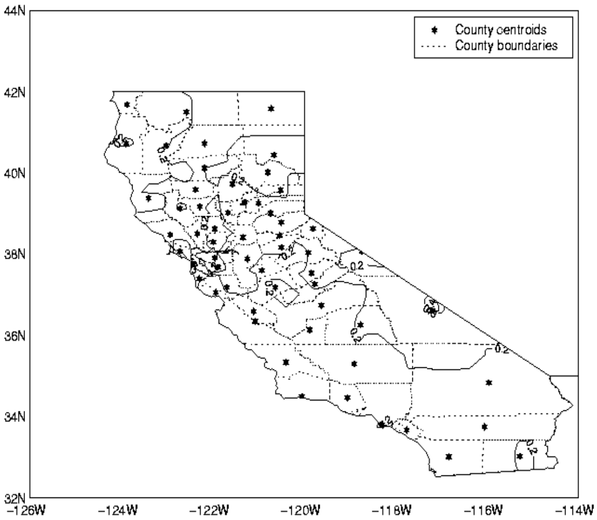

Figure 6. Spatial distribution of mortality rate estimates over California on Jan. 1, 1989 obtained by the measurement scale modelling approach (deaths per 100,000 people per day).

Figure 7 (bottom) shows that the ratio of the counts for these approaches have smaller average errors than that by the other approaches. The numerator of the ratio is the count that the multi-scale approach has smaller error than that by the measurement scale approach, and denominator is the other case for the measurement scale approach. The ratios are above 1 for almost all θ, which means that the multi-scale approach found estimates with less estimation errors more often than the measurement-scale approach. We conclude based on these comparisons that the multi-scale approach found estimates more accurately (i.e., closer to the actually local observations) than did the measurement scale approach .

5. CONCLUSIONS

To obtain a meaningful spatial distribution of a natural process from a limited scale of measurement data, we have introduced the recent development of multi-scale modelling. Stochastic variation of a random field was estimated, we believe realistically, at a scale local to the measurement scale which is usually limited by availability of data. Spatial analysis was studied conceptually over a continuous domain of space, although data were not observed in the continuous domain. Indeed, the scale of measurement data is usually not available at a uniform scale. By describing different scales of random fields for a natural process, we can represent more realistic variation for the process at the local scale to the available scale of actual measurement data. This multi-scale modelling approach seems to present more accurate spatial distribution of random field estimates than any other modelling approaches that do not account for multiple scales. This new modelling approach finds estimates based on the BME method so that realistic simulated soft data can be used. By collecting mortality rates at a local scale (California counties), it was verified that the estimates found by multi-scale approach were more accurate than those by county scale modelling approach accounting for only the measurement scale. Depending on available data and study purposes, the multi-scale modelling approach can be used to find geostatistical estimates at any scale of interest without being restricted by the scale of observed data.

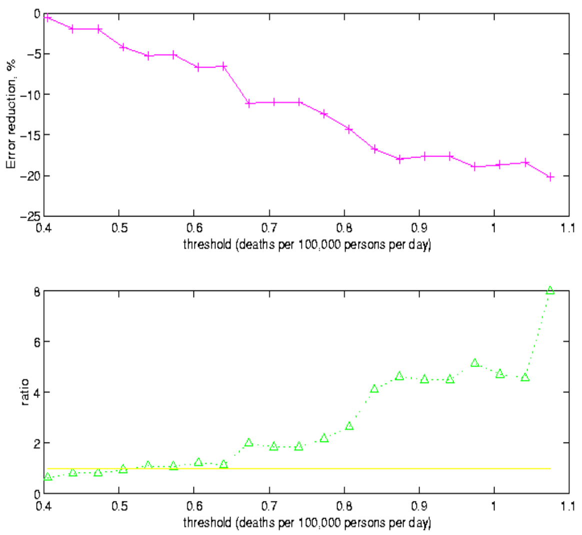

Figure 7. Top plot is the reduction percentage obtained by the multi-scale approach with respect to the average estimation error by the county scale approach. Bottom plot is the ratio of counts with smaller average estimation errors by the approaches. If the ratio is greater than 1, the multi-scale approach found estimates more often with smaller estimation errors than by the other approach.

6. REFERENCES

Carrat, F. and A.-J. Valleron, 1992. Epidemiologic mapping using the ‘kriging’ method: Application to an influenza-like illness epidemic in France. In: *Amer. Jour. of Epidemiology,* 135(11), pp. 1293-1300.

Choi, K-M, 2001. *BME-based spatiotemporal local scale mapping and filtering with mortality data: The California study.* Ph.D. dissertation, Dept. of Environmental Sciences and Eng., Univ. of North Carolina at Chapel Hill, Chapel Hill, NC.

Choi, K-M, G. Christakos, and M. Serre, 1998. Recent developments in vectorial and multi-point Bayesian maximum entropy analysis. In: *IAMG 98, Proceeding of 4th Annul Confer. of the International Assoc. for Mathematical Geology* – Vol. 1, Buccianti A., G. Nardi and R. Potenza (eds.), De Frede Editore, Naples, Italy, pp. 91-96.

Choi, K-M, G. Christakos, and M. Serre, 2001a. Space/time BME analysis and mapping of mortality in California. In: *The 53rd Session of the International Stat. Institute*, Seoul, Korea pp. 22-29.

Choi, K-M, M. Serre, and G. Christakos, 2001b. Spatiotemporal BME mapping of mortality fields at different spatial scales. In: *Jour. of Exposure Analysis and Environ. Exposure* (submitted).

Christakos, G., 1992. *Random field models in Earth Sciences.* Academic Press, San Diego, CA. 474 p.

Christakos, G., 2000. A view of Modern Geostatistics. Oxford Univ. Press, New York, NY, 304 p.

Christakos, G. and Hristopulos, D.T., 1998. *Spatiotemporal environmental health modelling: A tarctatus stocahsticus.* Kluwer Academic Publ., Boston, 424 p.

DSMF (Death Statistical Master Files), 1989. Center for Health Statistics, Sacramento, CA.

7. ACKNOWLEGEMENTS

This work has been supported by a grant from the Office of the Vice President for Research, The University of Michigan.