GeoDa Center for Geospatial Analysis and Computation

Arizona State University

Working Paper 2008-04

Space-Time Forecasting Using Soft Geostatistics: A Case Study in Forecasting Municipal Water Demand for Phoenix, Arizona

Seung-Jae Lee, Elizabeth A. Wentz, and Patricia Gober

Space-Time Forecasting Using Soft Geostatistics: A Case Study in Forecasting Municipal Water Demand for Phoenix, Arizona

Seung-Jae Lee[1*], Elizabeth A. Wentz[2], and Patricia Gober[2,3]

[1] Electric Systems Center, National Renewable Energy Laboratory, 1617 Cole Blvd., Golden, CO 80401, USA

[2] School of Geographical Sciences, Arizona State University, P.O. Box 870104, Tempe, AZ 85287-0104, USA

[3] Decision Center for a Desert City, School of Geographical Sciences and School of Sustainability, Arizona State University, P.O. Box 878209, Tempe, AZ 85287-8209, USA

Submitted to: *Geographical Analysis*

* Corresponding author: seungjae.lee@alumni.unc.edu

ABSTRACT: Managing environmental and social systems in the face of uncertainty requires the best possible forecasts of future conditions. We use space-time variability in historical data and projections of future population density to improve forecasting of residential water demand in the City of Phoenix, Arizona. Our future water estimates are derived using the first and second order statistical moments between a dependent variable, water use, and an independent variable, population density. The independent variable is projected at future points, and remains uncertain. We use adjusted statistical moments that cover projection errors in the independent variable, and propose a methodology to generate information-rich future estimates. These updated estimates are processed in Bayesian Maximum Entropy (BME), which produces maps of estimated water use to the year 2030. Integrating the uncertain estimates into the space-time forecasting process improves forecasting accuracy up to 43.9% over other space-time mapping methods that do not assimilate the uncertain estimates. Further validation studies reveal that BME is more accurate than co-kriging that integrates the error-free independent variable, but shows similar accuracy to kriging with measurement error that processes the uncertain estimates. Our proposed forecasting method benefits from the uncertain estimates of the future, provides up-to-date forecasts of water use, and can be adapted to other socio-economic and environmental applications.

Key Words: water use, forecasting, soft data, statistical moments, Bayesian Maximum Entropy

Introduction

Geographers and other scientists benefit from the conceptual foundations and subsequent implementation of tools to simulate and analyze space-time processes. The incorporation of space and time into studies of multidimensional and complex phenomena have been subjects of considerable theoretical, methodological, and applied research (MacEachren et al. 1998; MacEachren et al. 1999; Mennis and Peuquet 2000; Peuquet 2001, 2002, 2005; Bertolotto et al. 2007; Pebesma et al. 2007). Included are problems that address the rates, extents, and causes of tropical deforestation (Koffi et al. 1995); the anomalies related to vegetation and El Niño/Southern Oscillation events (Swetnam et al. 1999); and the results of forecasting urban growth (Ward et al. 2000). Two challenges in space-time analysis are interpolation and extrapolation. Interpolation involves estimating attribute values for locations within the spatial extent of the study area for which hard recorded data are not available. Extrapolation involves extending the spatial area or the temporal sequence beyond the scope of the observed data. Interpolation and extrapolation assume that observable patterns provide relevant information about the spatial and temporal dynamics of the phenomenon in question. While previous studies have mined these spatial and temporal dynamics separately, this study uses information about the dynamic interactions between space and time for future extrapolation.

There are numerous time-based approaches to extrapolate space-time phenomenon. Classical examples involve exponential smoothing, simple/weighted moving averages, adaptive/constant parameters, simple trend analysis, and regression techniques (Armstrong 1984; Adya and Collopy 1998; Gardner 2006). An extensive body of literature also covers more refined time-series models to account for autocorrelation in

regression errors (Wei 1990; Kedem 1993; Chatfield 2004). Commonly used models include the autoregressive (AR), moving average (MA), mixed AR-MA (ARMA), and integrated AR-MA (ARIMA). They first assume that estimated regression residuals in historical data are correlated, and then derive, after mathematical manipulations, ordinary regression models with an independent error term to use for forecasting.

Research addressing space-time (and not simply spatial or temporal) analysis uses a generalized regression technique that provides probabilistic outputs that vary with distance to data points. Geostatistical methods cope with non-stationary properties inherent in environmental data while accounting for spatial autocorrelations (Araghinejad et al. 2006); they were initially implemented for purely spatial estimation. Later, a more generalized space-time approach was developed by adding time as an additional dimension of space (Kyriakidis and Journel 1999). More advanced space-time geostatistical approaches were developed to account for causal dependencies in the composite space-time metric (Christakos 1992; Kyriakidis and Journel 1999; Christakos 2000), and their applications are increasingly found in environmental sciences (Vyas and Christakos 1997; Kyriakidis and Journel 2001a; Kyriakidis and Journel 2001b; Goovaerts et al. 2006) and land cover modeling (Boucher et al. 2006).

Space-time geostatistics are limited because they rely on complete and error-free measurements (i.e., hard data), which can be sparse. In addition, they use linear estimation procedures despite the non-linear dynamics of many biophysical and human systems. The goal of this paper is to demonstrate that it is possible to make estimates of the future from statistical moments between dependent and independent variables. This process creates soft data, using the relationship between the variables (first order

moments) and the level of uncertainty in the relationship (second order moments). Using soft data improves forecasting accuracy by providing a larger database and by integrating what we know about uncertainty into the modeling process.

We used the Bayesian Maximum Entropy (BME) approach of geostatistics (Christakos 1990, 2000; Christakos et al. 2002) to process soft data in a non-linear way. Unlike kriging methods, which assume Gaussian distributions (i.e., integrating up to second order statistical moments), BME can incorporate higher-order statistical moments. It can therefore cope with non-Gaussian conditions. BME has been used in applications dealing with urban sustainability (Brazel et al. 2007), climatology (Lee et al. 2008), hydrology (Serre et al. 2003a; Lee and Wentz 2008), exposure and health mapping (Lee 2005; Akita et al. 2007; Puangthongthub et al. 2007), risk assessment (Serre et al. 2003b; Choi et al. 2007), and geographical epidemiology (Law et al. 2004). These applications show that BME is a promising estimator, but there are, to our knowledge, no studies that use BME to forecast the distribution of a variable in space and time.

We demonstrate the development of a space-time forecasting model that takes advantage of soft data using a case study of water demand in Phoenix Arizona. This case study is relevant to urban planners and policy makers because of the growing urban population and the need to plan for residential water usage in a rapidly growing desert city. It also fulfills the requirement of our study because there is a dependent variable (water consumption) that requires forecasting; and an independent variable that can be measured in the present and projected into the future (population density). There is the additional requirement of sufficient geographic locations where there are observations for both the dependent and independent variable. The statistical moments derived between

the water usage (dependent variable) and population density (independent variable) in the present is applied to the projected independent variable to generate soft data of future water use.

Forecasting water demand in Phoenix, Arizona

The study area we used to develop a space-time extrapolation technique is the City of Phoenix, located in Maricopa County, Arizona. Phoenix is at the northern edge of the Sonoran Desert where summer temperatures can average 42°C or higher with rainfall averaging 20 mm per year. For the year 2000, the Maricopa Association of Governments (MAG) counted 3.7 million residents in Maricopa County with 1.5 million of those in the City of Phoenix. MAG estimates that the population of Phoenix will grow to 2.2 million by the year 2030 (MAG 2003). Population growth of this magnitude in an arid environment requires credible estimates of water demand for land planners and water managers.

Data

We used three sources of data to forecast water consumption for the City of Phoenix. The dependent variable is residential water consumption by census tract, which we acquired from the City of Phoenix Water Services Department for the years 1995-2004. These hard data, based on monthly billing records, were aggregated to the census tract level to protect the confidentiality of the city's individual water customers. The independent variable is population density per census tract derived from the 2000 US Census. The independent variable for the future is the projected population density, which we obtained

from MAG, the regional planning authority, for the years: 2010, 2020, 2025, and 2030. The specifics of each data source are described below.

Residential water consumption (*RWC*) data are derived from monthly billing records for water users in City of Phoenix for the year period 1995-2004. The monthly records were available as volumetric values in liters aggregated to census tracts (RWC_c) and summarized by user types (e.g., single family, multi-family, office, industry, retailer, public use, and mixed use). We extracted only the residential water records (both single family and multi-family) to develop a residential water duty (*RWD*) per census tract per year with:

$$RWD_{c,t} = \frac{RWC_{c,t}}{Area_c} \quad (1)$$

where *RWC* denotes the amount of residential water consumption (in liters) from single- or multi-family users for a particular census tract (*c*) and a particular year (*t*) within the period 1995-2004. *Area*, in km^2 is the total area of a census tract (*c*). Several years had missing *RWC* and therefore there are missing *RWD* for some tracts leading to different sample sizes per year (*n*=315 for 1995, *n*=304 for 1996-2001, *n*=305 for 2002-2003, and *n*=307 for 2004). Figure 1 illustrates *RWD* for *t*=1996, 2000, and 2004.

Population density for the 304 census tracts in the City of Phoenix was derived from Summary File 1 of Census 2000 (Figure 2). Because these data are based on an enumeration of residents for the year 2000, we assume these are hard data and there is no associated uncertainty. Although the problem of Census undercounts has been well documented in the demography literature, we are assuming that it is relatively small and unlikely to substantially affect our results. We extracted the total population for 304

census tracts in the City of Phoenix where *RWD* exist for the year 2000 and calculated the population density (people/km^2) for each census tract.

We obtained future population density from population projections for the years 2010, 2020, 2025, and 2030, provided by interim socioeconomic projections from MAG (2003). MAG followed the projection protocols developed by the Arizona Department of Economic Security, and allocated resident population for the future years by Municipal Planning Area (MPA), Regional Analysis Zone (RAZ), and Socioeconomic Analysis Zone (SAZ). The population projections by SAZ were used in our study because each observation represents the smallest area (in some cases, the same area as a census tract) and therefore provide the largest number of space-time points for future populations. We then developed future population density (people/km^2) for the 607 SAZs for each of the years 2010, 2020, 2025, and 2030.

BME Forecasting Approach

The modeling process involves three primary steps, explained in detail in this section. The first step involves generating an initial probability density function (pdf) of water consumption given the general knowledge base G, which uses the mean and covariance functions from the observed data. The second step builds the site-specific knowledge base, S consisting of hard and soft data on water use. Hard data are historical measurements at the census tract level and soft data are generated at all SAZ locations using a regression model. The final step updates the initial pdf with the S, which leads to the posterior pdf, which we use to map water consumption for all desired future time periods and the measures of uncertainty for each.

(i) Generating the prior pdf given the general knowledge base, G

We first introduce a space-time random field (STRF) to represent the space-time dynamics of residential water duties as log-*RWD*. We take the logarithm for estimation purposes because 1) water consumption is non-negative, and 2) such log-transformation are more likely to work with the Gaussian assumption. The STRF is defined as $X(\boldsymbol{p}_{\text{map}})$, denoting a random variable log-*RWD* in 3 dimensions (2-d for space and 1-d for time), where $\boldsymbol{p}_{\text{map}}$ consists of data points $\boldsymbol{p}_{\text{data}}$ and estimation points $\boldsymbol{p}_{\text{k}}$. In our case $\boldsymbol{p}_{\text{data}}$ are the points for observed residential water duties at census tracts over time (log-$RWD_{c,t}$) and soft data described below. The $\boldsymbol{p}_{\text{k}}$ are the points for future water consumption estimates. (see below for a description of $\boldsymbol{p}_{\text{k}}$). The STRF effectively reflects space-time variability and data uncertainty in residential water consumption through a joint probability density function (pdf) $f_X(\boldsymbol{?}_{\text{map}})$ where $\boldsymbol{?}_{\text{map}}$ are all possible realizations of the STRF $X(\boldsymbol{p}_{\text{map}})$ at $\boldsymbol{p}_{\text{map}}$. The pdf is used to describe the probability of a given $\boldsymbol{?}_{\text{map}}$:

$$f_X(\boldsymbol{?}_{\text{map}})\, d\boldsymbol{?}_{\text{map}} = Prob[\boldsymbol{?}_{\text{map}} < X(\boldsymbol{p}_{\text{map}}) < \boldsymbol{?}_{\text{map}} + d\boldsymbol{?}_{\text{map}}], \quad (2)$$

where *Prob*[.] is probability operator. This step constructs the prior pdf $f_G(\boldsymbol{?}_{\text{map}})$ that represents the initial probability of $X(\boldsymbol{p}_{\text{map}})$ over space and time, provided by the general knowledge base G. The general knowledge base G consists of the mean trend $m_X(\boldsymbol{p}_{\text{map}})$ and covariance functions $c_X(\boldsymbol{p}_{\text{map}}, \boldsymbol{p}_{\text{map}}')$ of water consumption (see Christakos 2000 for more details). Given the general knowledge base G, we derived a Gaussian-type prior pdf.

The mean trend function we used was an additive space-time trend model that applies space-time exponential filters (i.e., spatial range for exponential fitter=5km, temporal range for exponential filter=4years) to measured log-*RWD*. To obtain the covariance function for this study we calculated covariances at a series of spatial (r) and temporal (t) lags. We fitted these values to the experimental covariances. The fitted covariance is

separable and constructed by two exponential functions each of which is parameterized by space-time sills (c_{01} and c_{02}) and ranges (a_{r1}, a_{r2}, a_{t1}, and a_{t2}):

$$c_X(r,t) = c_{01}\exp\left(\frac{-3r}{a_{r1}}\right)\exp\left(\frac{-3t}{a_{t1}}\right) + c_{02}\exp\left(\frac{-3r}{a_{r2}}\right)\exp\left(\frac{-3t}{a_{t2}}\right), \quad (3)$$

where c_{01}= 1.53 (log-liters/km^2)^2, c_{02}=2.29 (log-liters/km^2)^2, a_{r1} = 1 km, a_{r2}= 9 km, a_{t1} = 1.6 years, and a_{t2} = 75 years.

(ii) Characterizing site-specific knowledge base, S

There is second type of knowledge base for BME called the site-specific knowledge base S, which represents error-free measurements (hard data) c_{hard} and uncertain data (soft data) c_{soft} of log-*RWD*. The output includes realizations of the STRF c_{data}=(c_{hard}, c_{soft}) at data points $\boldsymbol{p}_{\text{data}}$=($\boldsymbol{p}_{\text{hard}}$, $\boldsymbol{p}_{\text{soft}}$). Specifically in this study c_{hard} correspond to log-*RWD* measured at census tracts for 1995-2004, and $\boldsymbol{p}_{\text{hard}}$ represent the centroid of a census tract (c) and a particular year (t) during the period 1995-2004. The following equality holds between a STRF $X(\boldsymbol{p}_{\text{hard}})$ and its realization c_{hard}:

$$Prob[\, X(\boldsymbol{p}_{\text{hard}}) = c_{\text{hard}}] = 1, \quad (4)$$

The c_{soft} correspond to the estimated log-*RWD* using a linear regression model at the centroids ($\boldsymbol{p}_{\text{soft}}$) of the SAZ boundaries. The soft data are derived by applying regression results between log-*RWD* for the year 2000 (c_c in equation 5) and log-population density (y_c in equation 5) for the year 2000 centered at census tracts to log-population density at the centroids of SAZs in the future (y_i in equation 6).

The c_i in equation (6) of future log-*RWD* are expected values m_1 denoting first order statistical moments. This regressed relationship includes uncertainty that is characterized by standard errors equivalent to second order statistical moments m_2. The m_1 and m_2 builds

Gaussian soft data. To quantify m_1 and m_2, we used a quadratic relationship between c_c and y_c. We selected a non-linear relationship because BME is a non-linear estimator so it is more efficient than a linear estimator when dealing with non-linear properties. This procedure is based on the following equation:

$$c_c = b_0 + b_1 y_c + b_2 y_c^2 + e_c \quad (5)$$

where b_0, b_1, and b_2 are coefficients, e_c is an uncorrelated random error with zero mean and common variance. The regression theory leads to least squares parameters (b_0, b_1, and b_2) by minimizing the sum of the squares of the vertical distance between predicted and observed values. For a given measurement of the log-population density by SAZ (y_i) the regression results predict a non-linear estimate of log-*RWD* (c_i) per SAZ given y_i ($m_1[c_i|y_i]$), and its associated uncertainty ($m_2[c_i|y_i]$) through the following equation:

$$c_i = b_0 + b_1 y_i + b_2 y_i^2. \quad (6)$$

Gaussian soft data c_{soft} at $\boldsymbol{p}_{\text{soft}}$ for each centroid of the 607 SAZs for the years 2010, 2020, 2025, and 2030 are then described by a conditional probability density function $f_S(c_i|y_i)$:

$$?_{\text{soft}} = f_S(c_i|y_i) = N(m_1[c_i|y_i], m_2[c_i|y_i]). \quad (7)$$

In the case where y_i is a fixed value of the log of population density by SAZ with no uncertainty from projection errors, the $m_1[c_i|y_i]$ and $m_2[c_i|y_i]$ are equivalent to $b_0+b_1y_i+b_2y_i^2$ and $s_X^2(d^T(D^TD)^{-1}d)$ respectively where $d=[1 \quad y_i \quad y_i^2]^T$, D is a design matrix consisting of the first column with 304 series of 1, the second with the series of y_c, and the third with the series of y_c^2, and s_X^2 is an unbiased estimate for the common variance. It is more appropriate to use the square of the standard prediction error of c_i,

rather than the standard error of c_i ($m_2[c_i]$) when predicting a single (or independent) variable is important (Montgomery and Runger 2003). Equation (7) then becomes:

$$f_S(c_i|y_i) = N\ (b_0+b_1 y_i+b_2 y_i^2,\ s_X^2(d^T(D^T D)^{-1}d+1)), \quad (8)$$

When y_i is treated as random variable biased by the errors, equation (8) should be expanded. In cases where a regression parameter b_j (i.e., b_0, b_1, or b_2) and y_i are independent and the regression parameters are mutually dependent, different forms of statistical moments relative to those in equation (8) can be obtained. We, therefore, build probabilistic soft data $f_S(c_i)$ characterized by new moments:

$$f_S(c_i) = N\ (b_0+b_1 y_i+b_2(m_2[y_i]+ y_i^2),\ s_X^2(d^T(D^T D)^{-1}d)+f+ s_X^2), \quad (9)$$

where f is a function of b_j, y_i, $m_2[b_j]$, $m_2[y_i]$, and covariance matrix between b_j. We note that equation (8) is just a special case of equation (9) because equation (9) directly reduces to equation (8) under the condition that the uncertainty source is negligible (i.e., $m_2[y_i]=0$).

Up to this point we have two types of soft data (equation 8 and equation 9). Everything is known except for $m_2[y_i]$ in equation (9), which we need to approximate. A simple way is to use nugget covariance analysis (Lee 2005). We first equate a projection error-free Spatial Random Field (SRF) $Z(\boldsymbol{s})$ to projection field $Z'(\boldsymbol{s})$ (i.e., population density by SAZ for a year) times multiplicative projection errors e ($\boldsymbol{s}$). Taking the logarithm on both sides leads to the following relationship:

$$Y'(\boldsymbol{s}) = Y(\boldsymbol{s}) - \log e(\boldsymbol{s}), \quad (10)$$

where $Y'(\boldsymbol{s})$ =log-$Z'(\boldsymbol{s})$, and $Y(\boldsymbol{s})$=log-$Z(\boldsymbol{s})$. With the assumptions of 1) independence between $Y(\boldsymbol{s})$ and log $e(\boldsymbol{s})$, and 2) log $e(\boldsymbol{s})$ with a pure nugget covariance function, equation (10) is rewritten as:

$$c_{Y'}(r) = c_Y(r) + \sigma_{\log e}^2 \, \delta(r), \tag{11}$$

where $c_{Y'}(r)$ and $c_Y(r)$ are respectively covariances of $Y'(\boldsymbol{s})$ (log-population density by SAZs) and $Y(\boldsymbol{s})$ as a function of spatial lag r, and $\delta(r)$ is the Dirac delta function. In the case of zero lag, equation (11) is simplified to:

$$m_2[Y'] = m_2[Y] + m_2[\log e^2]. \tag{12}$$

We approximate $m_2[\log e^2]$ (i.e., random projection errors) because $m_2[Y']$ and $m_2[Y]$ are obtained from modeling experimental covariance of the realizations y_i for $Y'(\boldsymbol{s})$. As a result an expected value and variance of $Y(\boldsymbol{s})$ is derived when y_i is given, together with equation (10) and properties of log-normal distribution. Thus $Y(\boldsymbol{s})$ given y_i has a normal distribution N:

$$N\left(y_i - m_2[\log e^2]/2, \; m_2[\log e^2]\right). \tag{13}$$

Finally, the y_i and $m_2[y_i]$ in equation (9) are substituted by the $y_i - m_2[\log e^2]/2$ and $m_2[\log e^2]$ in equation (13) respectively.

(iii) Forecasting water consumption, combining knowledge bases G and S

To map future water consumption, we created a grid of 3721 points ($\boldsymbol{p}_k$) across Phoenix for 26 annual time periods from 2005 and 2030. Each grid intersection becomes an estimation point $\boldsymbol{p}_k$ representing the centroid of an undefined area. We did not account for the varying support of the data because the study of Lee and Wentz (2008) already addressed the support changes for Phoenix's water use. The area of the undefined

polygons is then assumed to be similar to those of c_{data} (mean of 4.01 km^2 and standard deviation of 14.17 km^2).

This step produces the posterior pdf $f_K(?_k)$ at any estimation point $\boldsymbol{p}_k$ by using both G and S knowledge bases through a Bayesian conditioning, i.e., $f_K(?_k)=f_G(?_k|\ \boldsymbol{?}_{hard},\boldsymbol{?}_{soft})$. This step provides the final probability of water consumption by the posterior pdf $f_K(?_k)$ at $\boldsymbol{p}_k$:

$$f_K(?_k) = A^{-1}\int d\boldsymbol{?}_{soft}\, f_S(\boldsymbol{?}_{soft})\, f_G(\boldsymbol{?}_{hard}, \boldsymbol{?}_{soft}, ?_k), \qquad (14)$$

where A is a normalization coefficient. The first order statistical moment of the posterior pdf is the estimate $?_k$ in our study while the second order statistical moment of the posterior pdf is the estimation uncertainty affected by the presence of data around the estimation point, and uncertainty in the soft data.

Validation

We validated our soft-data-based approach to forecasting water use in two ways. We first wanted to demonstrate that incorporating soft data improves forecasting capabilities. We therefore compared a soft data approach (BME with soft data) with two simple space-time forecasting methods (soft-data-free methods), which differ based on the type of hard data used. Our second validation method compares our approach to processing soft data with two other methods that use independent data in the forecasting process: (1) by means of cross-covariance (co-kriging), and (2) by means of soft data (kriging with measurement error).

Our first validation effort compares BME to two types of simple space-time kriging. The first simple space-time kriging approach uses only c_{hard}. The second uses both c_{hard} and $c_{hardened}$ (definition defined below). Five different cases are compared to understand

the impact soft data have on space-time forecasting. We used observed water use data (log-*RWD*, as hard data c_{hard}) from the period 1995-1999 to forecasted log-*RWD* for the year 2000, where we have observed *RWD* data. Each case consists of water use forecasting estimates for the year 2000 using c_{hardened} (see its definition below) and c_{soft}, and a different subset of c_{hard} where

Case 1 uses c_{hard} from 1995-1999

Case 2 uses c_{hard} from 1995-1998

Case 3 uses c_{hard} from 1995-1997

Case 4 uses c_{hard} from 1995-1996

Case 5 uses c_{hard} from 1995.

Using the measured values of *log-RWD* for the year 2000 (c_c) and *log*-population density (y_i) provides us with the information we need to calculate the regression parameters found in equation (6) and their uncertainty with which we can generate the soft data c_{soft} in equation (8) at the locations of y_i. The soft data account for data uncertainty from the extrapolation processes. If we neglect the uncertainty source then the soft data are hard data, identical to the first order moments of the soft data. We define these as hardened data, c_{hardened} to differentiate from the error-free measurements (c_{hard}). Using the five cases of c_{hard} we compared BME (using c_{hard} and c_{soft}) to two simple space-time kriging methods, one that uses c_{hard} alone and the second that uses both c_{hard} and c_{hardened} (Figure 4). We compared the year 2000 estimated extrapolation values from the three methods to the year 2000 observed values. These observed and predicted values of log-*RWD* for the year 2000 lead to the mean square errors (MSE) that are compared.

For the second validation we compared the BME approach of soft data processing to co-kriging and kriging with measurement error. We performed the analysis on the population density for the year 2000 because it is the only year where we have observed dependent and independent data. For each method we performed the following steps, 1) we randomly identified 60% of the census tracts (n=182) as the measured values for both dependent and independent variables, and assigned the independent variable alone (population density) to the remaining 40%; 2) with the dependent and independent variables available, we gained cross-covariance for co-kriging; 3) for the 40% sample, we obtained soft data for BME and kriging with measurement error; 4) for the 60% sample, we applied cross-validation to derive interpolation estimates of water use; and 5) using the interpolated estimates and the observed water use values for the 60% sample, we calculated the MSE to measure the accuracy of the interpolation. We performed this exercise 1000 times per method. For each iteration, we used a different random set of input data. We utilized a one-way ANOVA to compare the MSE results of the co-kriging and kriging with measurement error to BME.

Results

Validation

The MSE of BME (using c_{hard} and c_{soft}) and the two space-time kriging methods (using c_{hard} alone and both c_{hard} and c_{hardened}) were plotted for the five different cases for the year 2000 (Figure 4). From Case 1 to Case 5, forecasting accuracy tends to decrease regardless of the method because there are fewer hard data points as model inputs. Extrapolation based on historical data alone leads to inaccurate estimation as estimates are made

beyond the temporal scope of the observed data. To overcome the disadvantage in space-time forecasting, we develop a framework that benefits from independent data. According to our validation results, BME reduces the MSE of space-time kriging using c_{hard} and c_{hardened} by 43.9% in Case 5 which has the least error-free hard data. Space-time kriging using c_{hard} and c_{hardened} is less accurate than BME because the independent data are assimilated without accounting for data uncertainty in the extrapolation. The independent information, therefore, could lead to up-to-date extrapolation estimates only when its associated uncertainty is rigorously and simultaneously incorporated. There are also accuracy improvements when comparing BME to space-time kriging using c_{hard} data across the five cases, ranging 24.1% to 26.4%. We attribute these improvements to the incorporation of soft data into the forecasting procedure.

For the second validation we compared the BME approach to soft data processing to co-kriging and kriging with measurement error. Figures 5(a) and 5(c) shows the covariance for the dependent (log-*RWD*) and independent variables (log-population density) respectively, and Figure 5(b) indicates cross-covariance between the two variables. While each circle indicates experimental covariance, each plain curve denotes modeled covariance. The modeled covariance consists of two composite exponential functions:

$$c(r) = c^*_{01} \exp\left(\frac{-3r}{a^*_{r1}}\right) + c^*_{02} \exp\left(\frac{-3r}{a^*_{r2}}\right), \qquad (15)$$

where spatial ranges a^*_{r1} = 3.5 km and a^*_{r2}= 25 km for all models, c^*_{01}=1.6062 (log-liters/km^2)^2 and c^*_{02}=0.0328 (log-liters/km^2)^2 for log-*RWD*, c^*_{01}=0.7235 (log-poeple/km^2)^2 and c^*_{02}=0.0462 (log-poeple/km^2)^2 for log-population density, and

c^*_{01}=0.7464 (log-liters/km^2×log-poeple/km^2) and c^*_{02}=0.0562 (log-liters/km^2×log-poeple/km^2) for cross-covariance. Kriging with measurement error and BME depend on the covariance for log-*RWD* whereas co-kriging relies on the complete covariance matrix in Figure 5. As shown in Figure 6, we compute percent reduction in MSE from co-kriging to BME (Figure 6a) and kriging with measurement error to BME (Figure 6b) based on 1000 MSE iterations of each method. If any two methods that are compared result in an identical MSE, the percent reduction between the two methods is zero, shown as a horizontal line in the figure. The kriging methods are more accurate than BME above the line and less accurate than BME below the line.

Table 1 reports the ANOVA results with MSE as dependent variable and each method as independent variable. As demonstrated by Table 1 BME produces similar results to kriging with measurement error but better results than co-kriging. In fact BME reduces least squares mean MSE by 12.2% over co-kriging. This reduction leads us to reason that, when integrating additional data (population density) BME with soft data produces more accurate results than co-kriging, which relies on cross-correlation. The similarity in accuracy between BME and kriging with measurement error is because we used Gaussian soft data describing up to the second order statistical moments. Although we did not derive non-Gaussian soft data here, if non-Gaussian soft data are ready for estimation, BME is the only method that incorporates non-Gaussian soft data. We expect that BME would reduce MSE over kriging with measurement error, as demonstrated in the study of Serre and Christakos (1999).

Estimating future water use in Phoenix

To obtain $m_2[\log e^2]$ in equation (13) for each year 2010, 2020, 2025, and 2030, we calculated experimental covariances of $Y'(\boldsymbol{s})$ (in our study log-population density by SAZs) at certain spatial lags (circles in Figure 7), and fit the covariances with an exponential model (solid curve in Figure 7). The first circle at zero of spatial lag denotes variance ($m_2[Y']$ in equation 12). To approximate the projection error ($m_2[\log e^2]$ in equation 13), we initially calculated covariances at first two spatial lags that are close to the zero lag (second and third circles in the figure), then $m_2[Y]$ in equation (12) through linear extrapolation using the second and third circles, and finally $m_2[\log e^2]$ by equation (12). The $m_2[\log e^2]$ is interpreted as an experimental nugget of the covariance model and shown as a thick vertical line at zero of spatial lag (Figure 7; Table 2). Each value coincides with $m_2[\log e^2]$ in equation (13) representing an average projection error in the SAZ data for a given year. We then compute $y_i - m_2[\log e^2]/2$ in equation (13) using the predicted $m_2[\log e^2]$. The values of y_i and $m_2[y_i]$ equation (9) are respectively substituted by $y_i - m_2[\log e^2]/2$ and $m_2[\log e^2]$ that characterizes uncertainty from the projections.

We construct the relationship between c_c and y_c by calculating the least squares parameters in equation (6):

$$c_i = 8.8567 + 1.8694 y_i - 0.0604 y_i^2 \qquad (16)$$

Figure 8 shows this relationship (solid curve), 95% prediction interval (dotted curve), and c_c against y_c (dots). For extrapolation purposes, this relationship is applied to all future log-population density y_i to produce the soft pdf $f_S(c_i)$ at the space-time points covering Phoenix. The values of y_i, however, remain uncertain due to the projection errors embedded in y_i. If the projection errors are inevitably neglected (i.e., $m_2[y_i]=0$), soft data

generation relies on equation (8) as used in the validation study. Our proposed framework generates soft data using equation (9) while accounting for the data uncertainty sources from the projection error in addition to temporal extrapolation.

BME was used to map future water duties for the City of Phoenix to the year 2030. The measured *log-RWD* for all years available (1995-2004) is now assigned as hard data c_{hard} (n=3056). To derive soft data c_{soft} we maintain c_c and y_c to determine the regression parameters, and apply the regressed results to y_i representing projected population density by SAZs for the years 2010, 2020, 2025, and 2030. Since this y_i contains projection errors, the c_{soft} (n=2428) is generated by equation (9) rather than equation (8).

BME processes the c_{hard} and c_{soft} and resulting estimation is a series of the posterior pdf at the estimation points across Phoenix and all years between 2005 and 2030. For illustration purposes we represent maps of Phoenix's water duties (Figure 9) in 2005, 2010, 2015, 2020, 2025, and 2030 by extracting mean values of the posterior pdfs for these years. These results illustrate that Phoenix's residential water use peaks between 2012 and 2017, and afterward gradually decreases by 2030. The up-and-down behavior reflects the changing balance between densification which increases water use and conservation which reduces it. Increasing conservation is reflected in our historical series of water use; household water demand has, in fact, declined over time in Phoenix. Densities have steadily increased in Phoenix, and that trend is expected to continue as revealed in the population density projections. Forecasts of increasing water use in the 2010 and 2015 maps reflect rapidly increasing densities aligned with minor increases in conservation. In the 2020, 2025, and 2030, conservation effects begin to outweigh density gains, and consumption declines overall.

Discussion

This paper demonstrates the potential of BME to forecast of water use for Phoenix Arizona. Water use in the future remains uncertain, however, we can use knowledge about the relationship between water use and population density and estimates of future population density patterns to infer the space-time dynamics of water use in Phoenix. Soft data generated by the regression results between water consumption and population density provides a reasonable approximation of future patterns. In an evaluation exercise, we showed that our space-time geostatistical approach is promising because it processes 1) space-time dependencies in historical data, and 2) an independent variable for future points pertinent to the application through soft data detailing uncertain water use in the future.

An important component of water conservation policy development and infrastructure management is having most accurate forecasting model of future residential water demand. Water demand in Phoenix is affected by uncertain climate, rapid population growth, an urban heat island effect, and the use of pools and irrigated landscapes (Brazel et al. 2007; Guhathakurta and Gober 2007; Wentz and Gober 2007). In response to the long-term risk from water scarcity, numerous conservation strategies have been implemented by local and state governments, leading to a gradual decrease of per capita annual water use in Phoenix (Balling and Gober 2007). Our developed model provides credible forecasts of future water demand and considers on-going conservation policy and population growth.

Soft data for forecasting should be more informative by considering 1) any interaction terms neglected in equation (9) to avoid overestimating corresponding variances, 2) multiple independent variables rather than one variable (i.e., population density in our study), 3) point-specific projection errors and higher order statistical moments derived from a mathematical framework or a measurement error model. We will pursue these points in future publications.

Conclusion

The BME approach demonstrated here for Phoenix water consumption takes advantage of composite space-time dynamics to project future water use. We use statistical moments to generate future patterns of water use that include uncertainty (i.e., extrapolation and projection error) but nevertheless improve upon uncertainty-free estimations. This method of forecasting can be adapted to a wide range of socio-economic and environmental applications, including land use/land cover change, small-area population forecasting, energy and water demand, and modeling the spread of disease.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. SES-0345945, Decision Center for a Desert City (DCDC). Any opinions, findings and conclusions or recommendation expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

References

Akita, Y., G. Carter, and M.L. Serre. (2007). "Spatiotemporal Non-Attainment Assessment of Surface Water Tetrachloroethene in New Jersey." *Journal of Environmental Quality* 36(2), 508-520.

Araghinejad, S., D. H. Burn, and M. Karamouz. (2006). "Long-Lead Probabilistic Forecasting of Streamflow Using Ocean-Atmospheric and Hydrological Predictors." *Water Resources Research* 42, W03431, doi:10.1029/2004WR003853.

Adya, M, and F. Collopy. (1998). "How Effective Are Neural Networks at Forecasting and Prediction? A Review and Evaluation." *Journal of Forecasting* 17, 481-495.

Armstrong, J. S. (1984). "Forecasting by Extrapolation: Conclusions from 25 Years of Research." *Interfaces* 14, 52-66.

Balling, R. C., and P. Gober. (2007). "Climate Variability and Residential Water Use in the City of Phoenix, Arizona." *Journal of Applied Meteorology and Climatology* 46, 1130-1137.

Bertolotto, M., S. Di Martino, F. Ferrucci, and T. Kechadi. (2007). "Towards a Framework for Mining and Analysing Spatio-Temporal Datasets." *International Journal of Geographical Information Science* 21(8), 895-906.

Boucher, A, K. C. Seto, and A. Journel. (2006). "A Novel Method for Mapping Land Cover Changes: Incorporating Time and Space with Geostatistics." *IEEE Transactions on Geoscience and Remote Sensing* 44(11), 3427-3435.

Brazel, A., P. Gober, S. J. Lee, S. Grossman-Clarke, J. Zehnder, B. Hedquist, and E. Comparri. (2007). "Determinants of Changes in the Regional Urban Heat Island in Metropolitan Phoenix (Arizona, USA) between 1990 and 2004." *Climate Research* 33, 171-182.

Chatfield, C. (2004). *The Analysis of Time Series*. Boca Raton, FL: Chapman & Hall/CRC.

Choi, K. M., H. L. Yu, and M. L. Wilson. (2007). "Spatiotemporal Statistical Analysis of Influenza Mortality Risk in the State of California during the Period 1997-2001." *Stochastic Environmental Research and Risk Assessment*, doi:10.1007/s00477-007-0168-4.

Christakos, G. (1990). "A Bayesian/Maximum-Entropy View to the Spatial Estimation Problem." *Mathematical Geology* 22(7), 763-776.

Christakos, G. (1992). *Random Field Models in Earth Sciences*. Mineola, NY: Dover Publications.

Christakos, G. (2000). *Modern Spatiotemporal Geostatistics*. New York, NY: Oxford University Press.

Christakos, G., P. Bogaert, and M. L. Serre. (2002). *Advanced Functions of Temporal GIS*. New York, NY: Springer-Verlag.

Gardner, E. S. (2006). "Exponential Smoothing: The State of the Art-Part II." *International Journal of Forecasting* 22, 637-666.

Goovaerts, P., A. Auchincloss, and A. V. Diez-Roux. (2006). "Performance Comparison of Spatial and Space-Time Interpolation Techniques for Prediction of Air Pollutant Concentrations in the Los Angeles Area." *Society for Mathematical Geology XIth International Congress*, S13-11.

Guhathakurta, S., and P. Gober. (2007). "The Impact of the Phoenix Urban Heat Island on Residential Water Use." *Journal of the American Planning Association* 73(3), 317-329.

Kedem, B. (1993). *Time-Series Analysis by Higher Order Crossings*. New York, NY: IEEE Press.

Koffi B., J.M. Gregorie., G. Mahe, and J.P. Lacaux. (1995). "Remote-Sensing of Bush Fire Dynamics in Central-Africa from 1984 to 1988 – Analysis in Relation to Regional Vegetation and Pluviometric Patterns." *Atmospheric Research* 39(1-3), 179-200.

Kyriakidis, P. C., and A. G. Journel. (1999). "Geostatistical Space-Time Models: A Review." *Mathematical Geology* 31(6), 651-684.

Kyriakidis, P. C., and A. G. Journel. (2001a). "Stochastic Modeling of Atmospheric Pollution: A Spatial Time-Series Framework. Part I: Methodology." *Atmospheric Environment* 35, 2331-2337.

Kyriakidis, P. C., and A. G. Journel. (2001b). "Stochastic Modeling of Atmospheric Pollution: A Spatial Time-Series Framework. Part II: Application to Monitoring Monthly Sulfate Deposition over Europe." *Atmospheric Environment* 35, 2339-2348.

Law, D.C.G., M.L. Serre, G. Christakos, P.A. Leone, and W.C. Miller. (2004). "Spatial Analysis and Mapping of Sexually Transmitted Disease to Optimize Intervention and Prevention Strategies." *Sexually Transmitted Infections* 80, 294-299.

Lee, S. J. 2005. "Models of Soft Data in Geostatistics and Their Application in Environmental and Health Mapping." Ph.D. dissertation, Department of

Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, NC, U.S.A.

Lee, S. J., and E. Wentz. (2008). "Applying Bayesian Maximum Entropy to Extrapolating Local-Scale Water Consumption in Maricopa County, Arizona." *Water Resources Research* 43, W01401, doi:10.1029/2007WR006101.

Lee, S. J., R. Balling, and P. Gober. (2008). "Bayesian Maximum Entropy Mapping and Soft Data Problem in Urban Climate Research." *Annals of the Association of American Geographers* 98(2), 309-322.

MacEachren, A. M., F. Boscoe, D. Haug, and L. Pickle. (1998). "Geographic Visualization: Designing Manipulable Maps for Exploring Temporally Varying Georeferenced Statistics." *IEEE Information Visualization Symposium*, 87-94.

MacEachren, A. M., M. Wachowicz, R. Edsall, D. Haug, and R. Masters. (1999). "Constructing Knowledge from Multivariate Spatio-Temporal Data: Integrating Geographic Visualization (GVis) with Knowledge Discovery in Databases (KDD)." *International Journal of Geographic Information Science* 13(4), 311-334.

Maricopa Association of Governments. (2003). "Interim Socioeconomic Projections Documentation, Phoenix, Arizona."

Mennis, J., and D. J. Peuquet. (2000). "A Conceptual Framework for Incorporating Cognitive Principles into Geographical Database Presentation." *International Journal of Geographical Information Science* 14(6), 501-520.

Montgomery, D. C, and G. C. Runger. (2003). *Applied Statistics and Probability for Engineers*. New York, NY: John Wiley & Sons, Inc.

Pebesma, E. J., K. de Jong, and D. Briggs. (2007). "Interactive Visualization of Uncertain Spatial and Spatio-Temporal Data under Different Scenarios: An Air Quality Example." *International Journal of Geographical Information Science* 21(5), 515-527.

Peuquet, D. J. (2001). "Making Space for Time: Issues in Space-Time Representation." *Geoinformatica* 5(1), 11-32.

Peuquet, D. J. (2002). *Representations of Space and Time*. New York, NY: Guilford Press.

Peuquet, D. J. (2005). "Theme Section on Advances in Spatio-Temporal Analysis and Representation." *ISPRS Journal of Photogrammetry and Remote Sensing* 60(1), 1-2.

Puangthongthub, S., S. Wangwongwatana, R.M. Kamens, M.L. Serre. (2007). "Modeling the Space/Time Distribution of Particulate Matter in Thailand and Optimizing its Monitoring Network." *Atmospheric Environment* 41, 7788-7805.

Serre, M. L., G. Christakos. (1999). "Modern Geostatistics: Computational BME Analysis in the Light of Uncertainty Physical Knowledeg – the Equus Beds Study." *Stochastic Environmental Research and Risk Assessment* 13, 1-26.

Serre, M. L, G. Christakos, H. Li, and C. T. Miller. (2003a). "A BME Solution of the Inverse Problem for Saturated Groundwater Flow." *Stochastic Environmental Research and Risk Assessment* 17, 354-369.

Serre, M. L., A. Kolovos, G. Christakos, and K. Modis. (2003b). "An Application of the Holistochastic Human Exposure Methodology to Naturally Occurring Arsenic in Bangladesh Drinking Water." *Risk Analysis* 23(3), 515-528.

Swetnam T. W., C. D. Allen, and J. L. Betancourt. (1999). "Applied Historical Ecology: Using the Past to Manage for the Future." *Ecological Applications* 9(4), 1189-1206.

Vyas, V. M., and G. Christakos. (1997). "Spatiotemporal Analysis and Mapping of Sulfate Deposition Data Eastern U.S.A." *Atmospheric Environment* 31(21), 3623-3633.

Ward D, S. R. Phinn SR, and A. T. Murray. (2000). "Monitoring Growth in Rapidly Urbanizing Areas Using Remotely Sensed Data." *Professional Geographer* 52(3), 371-386.

Wei, W. W. S. (1990). *Time Series Analysis*. New York, NY: Addison-Wesley Publishing Company, Inc.

Wentz, E. A., and P. Gober. (2007). "Determinants of Small-Area Water Consumption for the City of Phoenix, Arizona." *Water Resources Management* 21(11), 1849-1863.

Figure captions

Figure 1: Residential water duty data (liters/km^2) for the years 1996, 2000, and 2004.

Figure 2: Population density data (People/km^2) for the census tracts of the year 2000.

Figure 3: Maricopa Association of Governments (MAG) population projections (People/km^2) by Socioeconomic Analysis Zone (SAZ) for the years 2010, 2020, 2025, and 2030.

Figure 4: Mean square estimation errors of three space-time geostatistical methods over 5 different cases representing various forecasting situations.

Figure 5: A matrix of experimental (circles) and modeled (plain curve) covariances used for the second validation study: (a) covariance for log-*RWD*, (b) cross-covariance between log-*RWD* and log-population density, and (c) covariance for log-population density.

Figure 6: 1000 sets of percent reduction in MSE (each dot) (a) from Co-kriging to BME and (b) from Kriging with measurement error to BME.

Figure 7: Experimental covariances (circles), an exponential covariance model (solid curve), an experimental nugget indicating an average of projection errors (thick vertical

line) for SAZ log-population density of the years (a) 2010, (b) 2020, (c) 2025, and (d) 2030.

Figure 8: log-*RWD* versus log-population density observed for the year 2000 (dots), first (solid curve) and second (dotted curve) order statistical moments.

Figure 9: BME processes historical hard data and future soft data (equation 9) to produce its forecasting maps of Phoenix's water duties (liters/km^2) in between 2005 and 2030. Among the 26 snapshots created, we show only six maps of the years 2005, 2010, 2015, 2020, 2025, and 2030 for illustration purposes.

Tables

Table 1: One-way ANOVA output to test method effects on MSE

Method	Least Squares Mean MSE	Standard Error	Pr>| t |
Co-kriging	1.5234E16	7.4612E13	<.0001
Kriging with measurement error	1.3371E16	7.4612E13	<.0001
BME	1.3373E16	7.4612E13	<.0001
Least Squares Means for effect Method Pr>| t | for H_0: LSMean(i)=LSMean(j)			
i/j	Co-kriging	Kriging with measurement error	BME
Co-kriging		<.0001	<.0001
Kriging with Measurement error	<.0001		0.9879
BME	<.0001	0.9879	

Table 2: Experimental nugget by year

Year	Experimental Nugget	See Figure
2010	2.8515	7a
2020	1.5566	7b
2025	1.0526	7c
2030	1.0067	7d

Figures

Figure 1

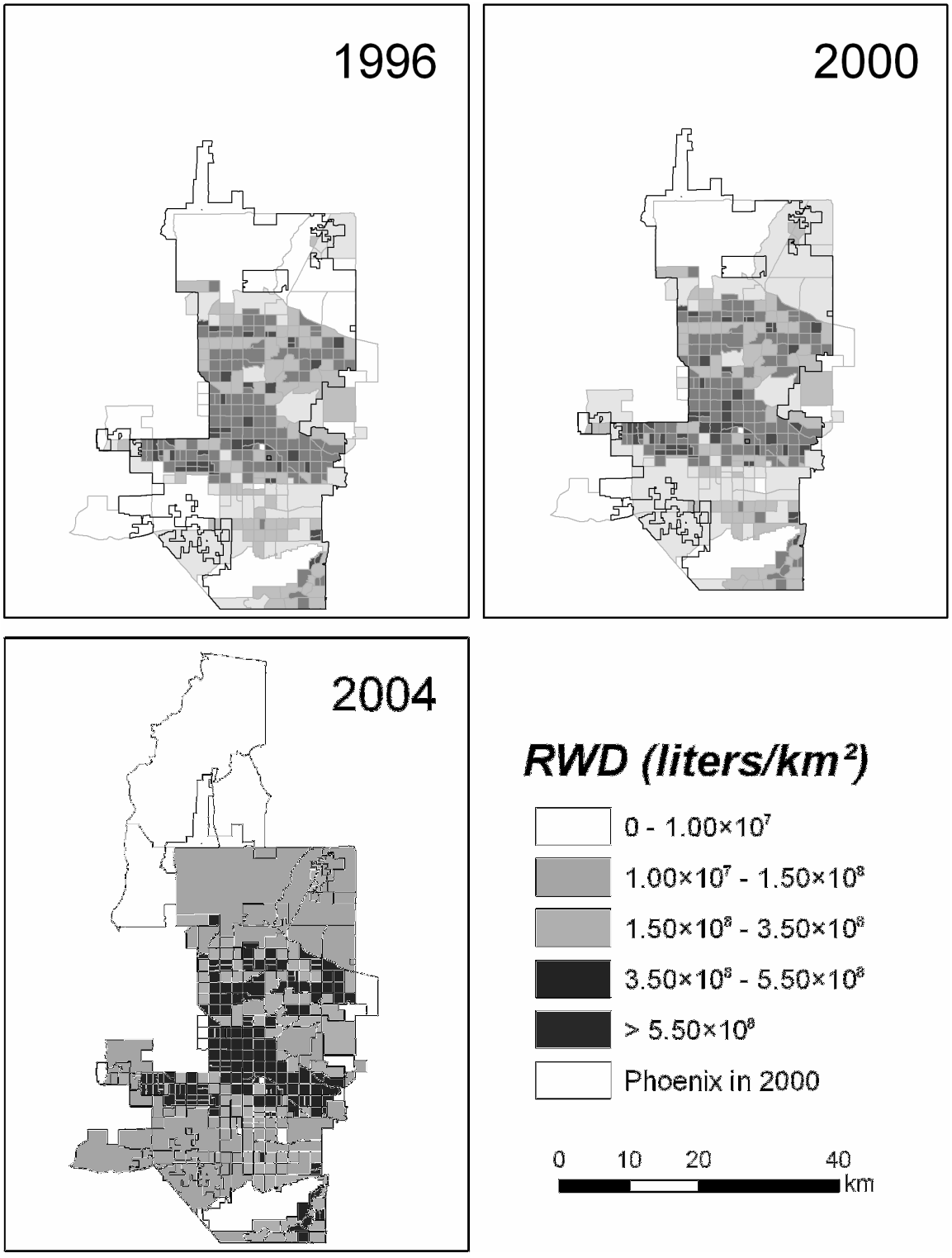


Figure 2

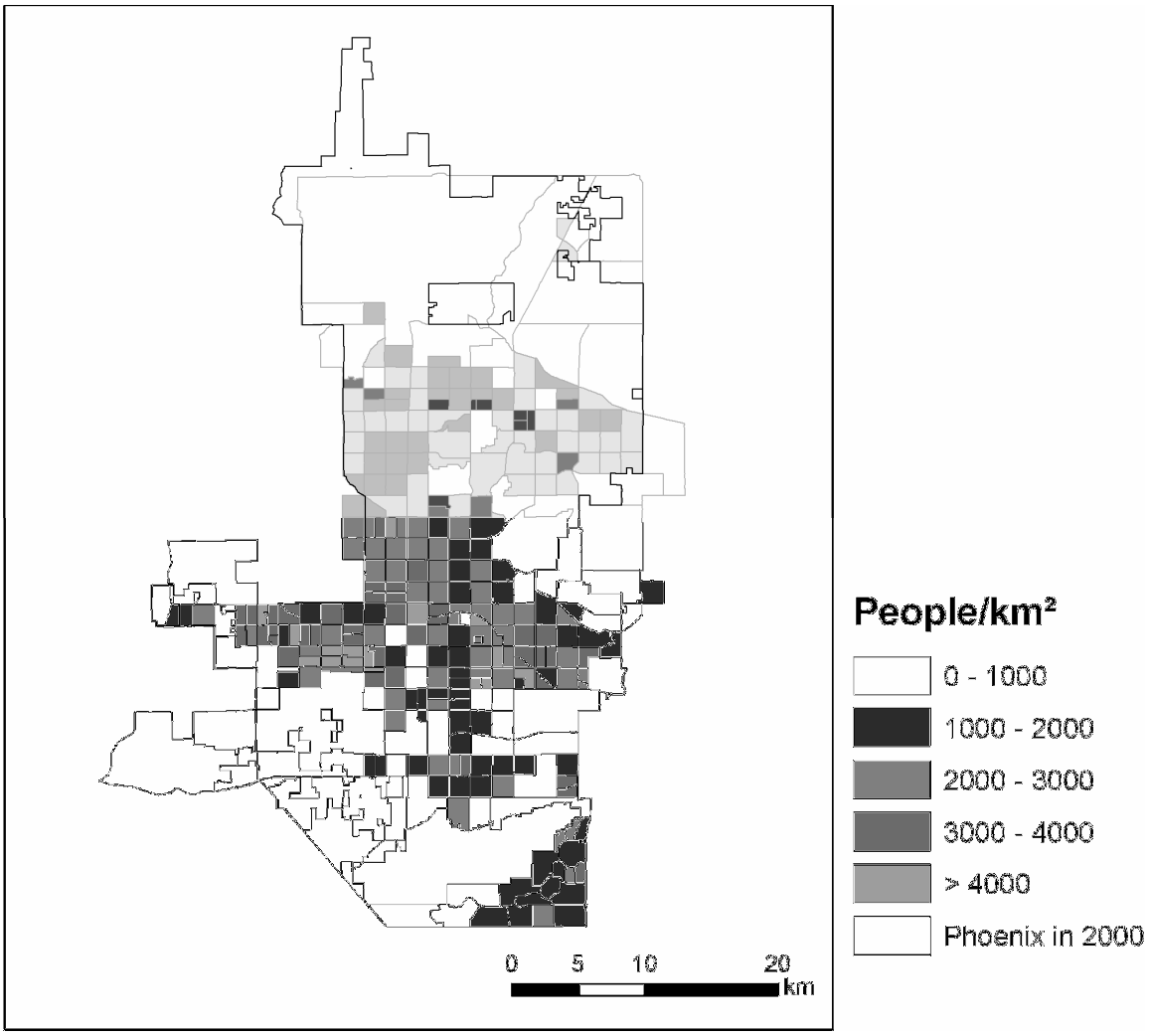

People/km²
0 - 1000
1000 - 2000
2000 - 3000
3000 - 4000
> 4000
Phoenix in 2000
0 5 10 20
km

Figure 3

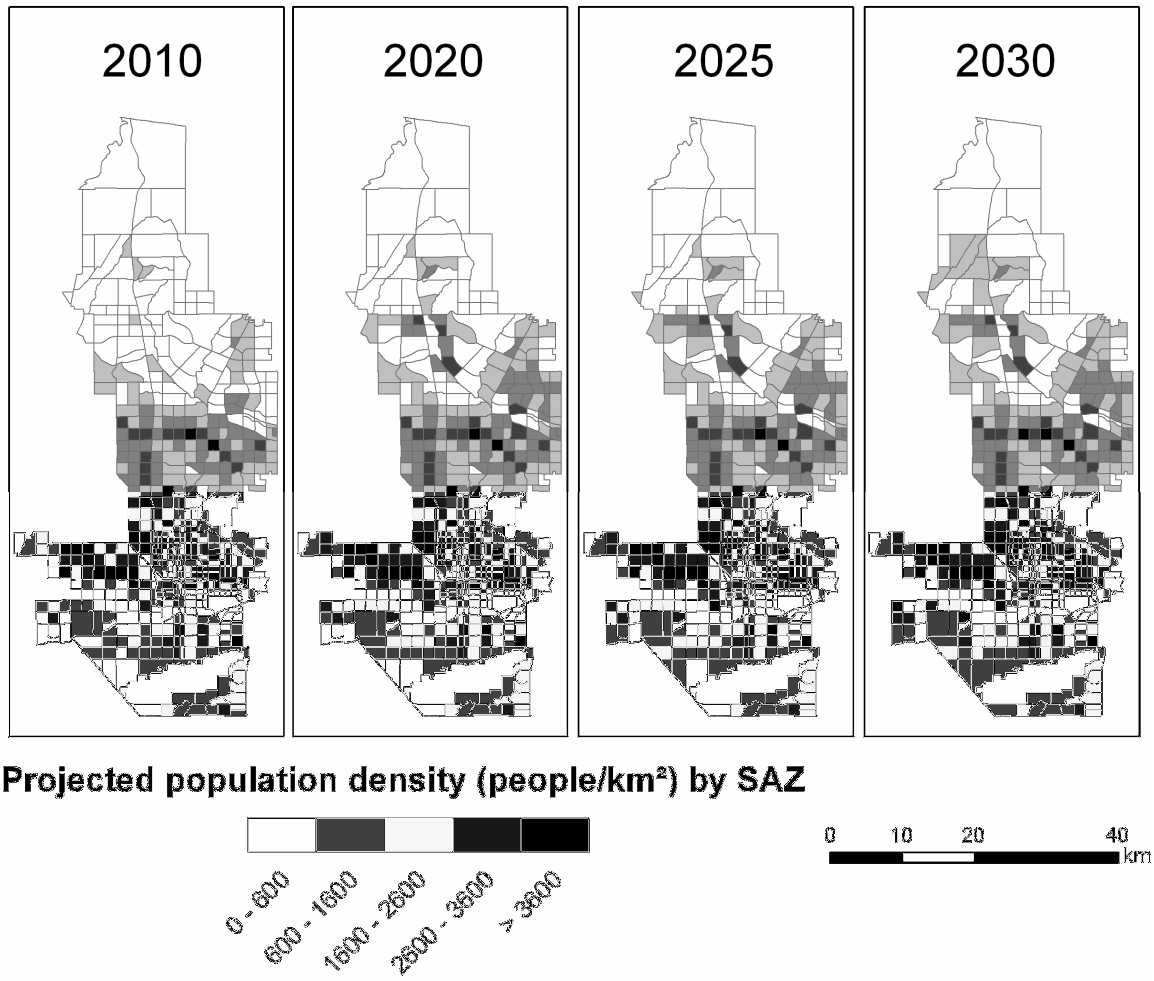

2010
2020
2025
2030
Projected population density (people/km²) by SAZ
0 - 600
600 - 1600
1600 - 2600
2600 - 3600
> 3600
0
10
20
40
km

Figure 4

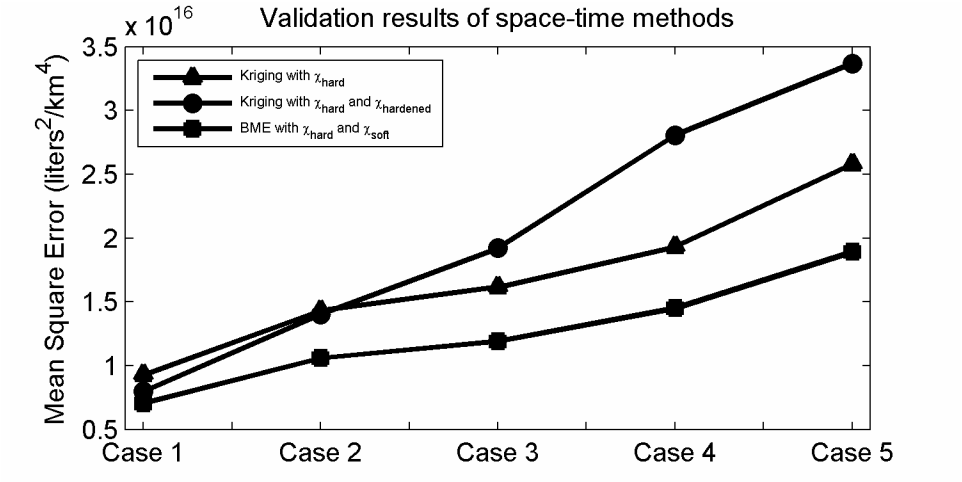

Validation results of space-time methods
x 10^16
Mean Square Error (liters^2/km^4)
3.5
3
2.5
2
1.5
1
0.5
Kriging with χhard
Kriging with χhard and χhardened
BME with χhard and χsoft
Case 1
Case 2
Case 3
Case 4
Case 5

Figure 5

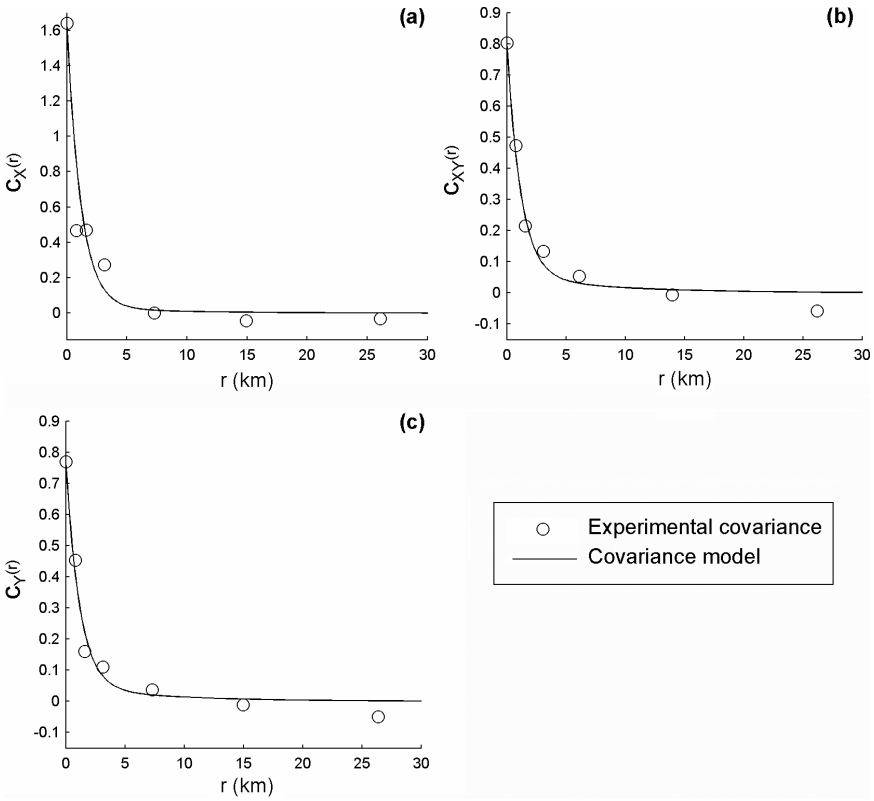

(a)
(b)
(c)
$C_X(r)$
$C_{XY}(r)$
$C_Y(r)$
r (km)
Experimental covariance
Covariance model

Figure 6

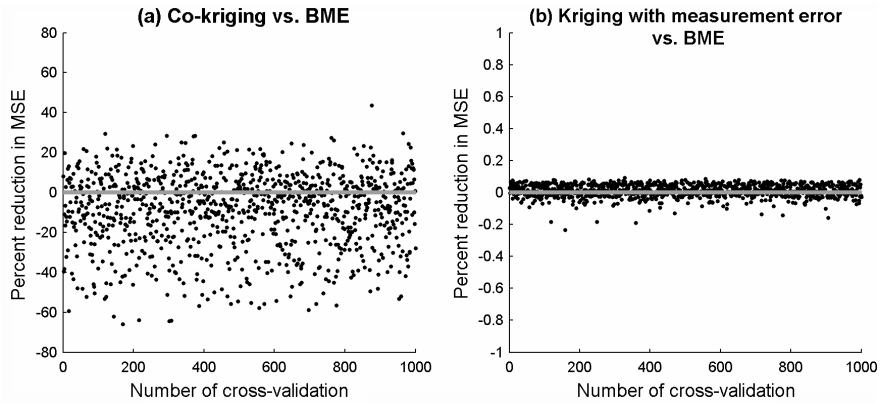

(a) Co-kriging vs. BME
Percent reduction in MSE
Number of cross-validation
(b) Kriging with measurement error vs. BME
Percent reduction in MSE
Number of cross-validation

Figure 7

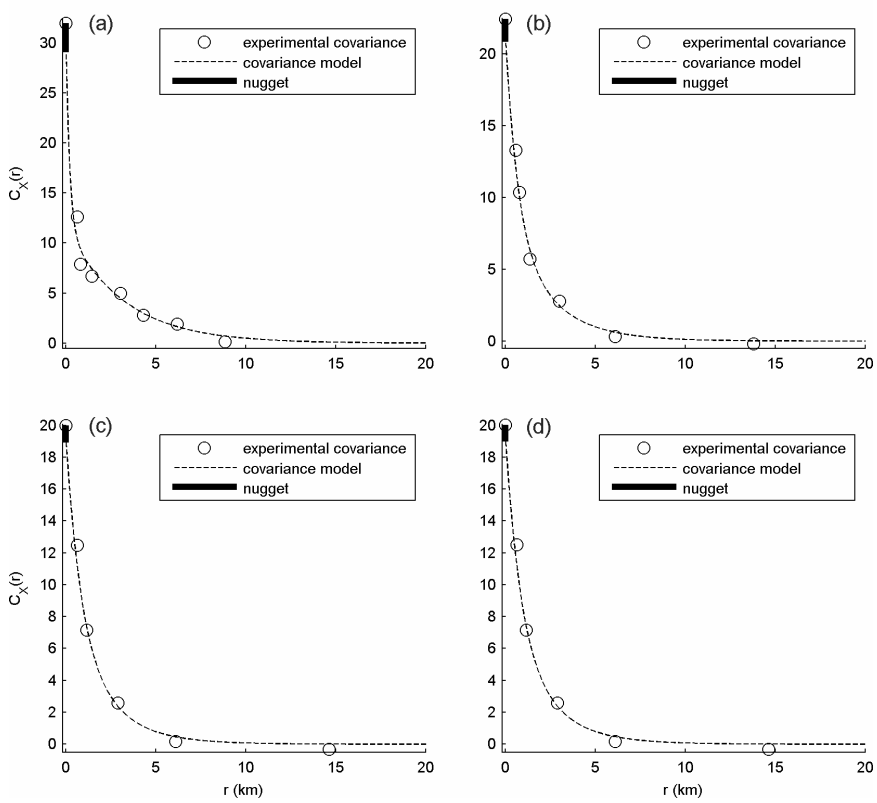

(a)
(b)
(c)
(d)
experimental covariance
covariance model
nugget
$C_X(r)$
r (km)

Figure 8

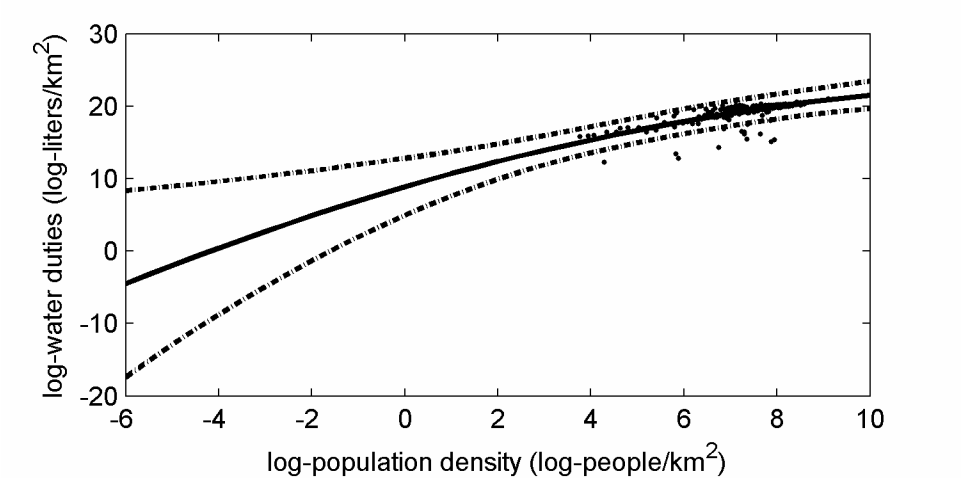
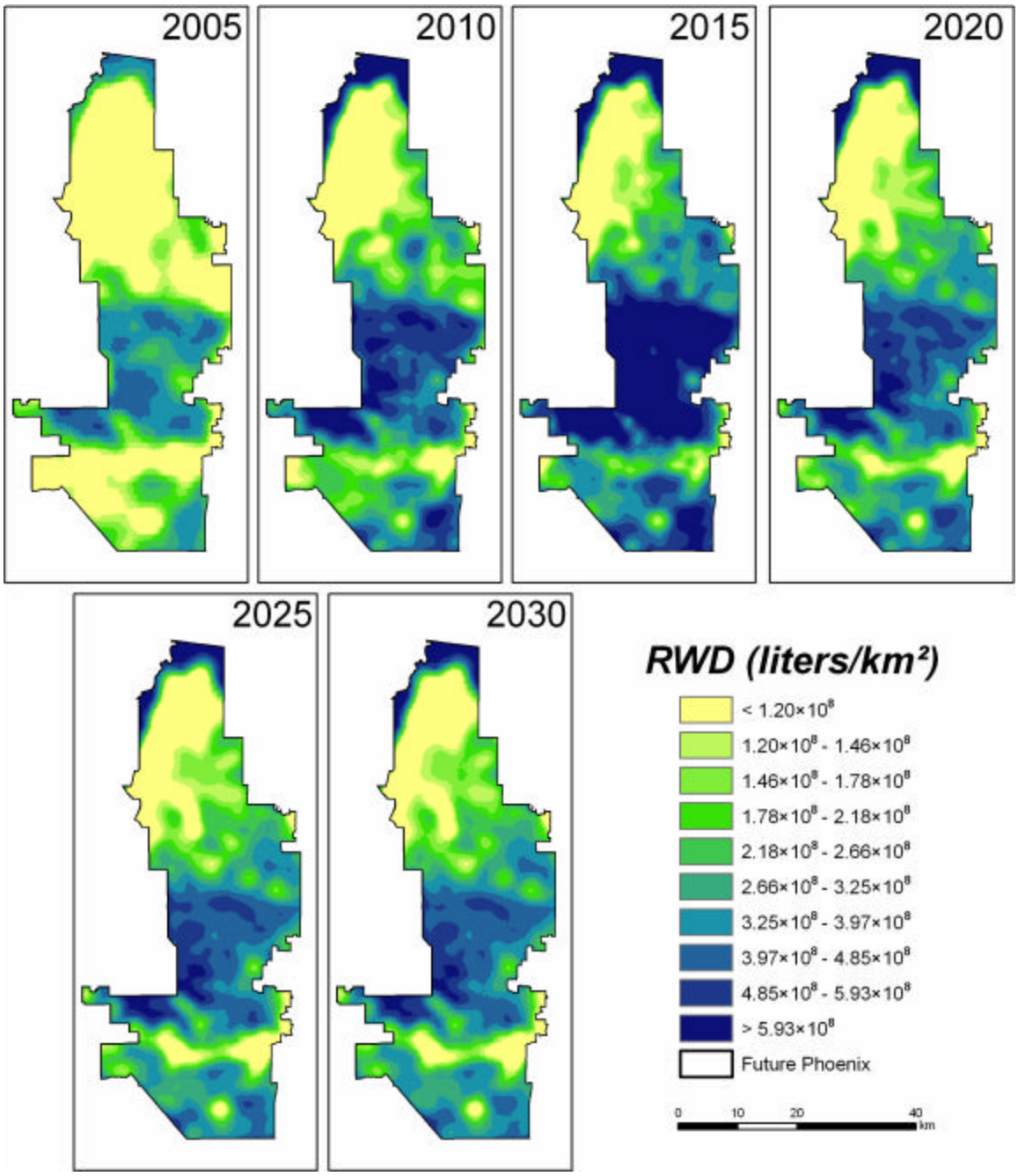

log-water duties (log-liters/km^2)
30
20
10
0
-10
-20
-6
-4
-2
0
2
4
6
8
10
log-population density (log-people/km^2)

Figure 9


2005
2010
2015
2020
2025
2030
RWD (liters/km²)
< 1.20×10⁸
1.20×10⁸ - 1.46×10⁸
1.46×10⁸ - 1.78×10⁸
1.78×10⁸ - 2.18×10⁸
2.18×10⁸ - 2.66×10⁸
2.66×10⁸ - 3.25×10⁸
3.25×10⁸ - 3.97×10⁸
3.97×10⁸ - 4.85×10⁸
4.85×10⁸ - 5.93×10⁸
> 5.93×10⁸
Future Phoenix
0
10
20
40
km