

A Comparison of Thiessen-polygon, Kriging, and Spline Models of UV Exposure

Zaria Tatalovich

GIS Research Laboratory, Department of Geography, College of Letters, Arts and Sciences, University of Southern California, 3620 South Vermont Avenue, Kaprielian Hall 416, Los Angeles, CA 90089-0255
Tel: 213.821.1313, Fax: 213.740.0056 e-mail: tatalovi@usc.edu

Abstract

The performance of Thiessen-polygon and Kriging procedures from the standard GIS package was evaluated based on the magnitude and distribution of errors, and the results compared with Australian National University Splines (ANUSPLIN) routine that runs outside typical GIS through a series of C++ and FORTRAN commands. The objective is to identify model that produces least amount of uncertainties in predictions, and utilize that model to generate UV exposure estimates at unsampled locations across United States. Input data include average global radiation measures (AVGALO) computed from hourly data for period 1961-1990 for 215 stations in the U.S. (from National Solar Radiation Database), latitude, longitude, and elevation from 30 arc-second Digital Elevation Model (from National Oceanic and Atmospheric Administration). ANUSPLIN model produced results with the smallest mean absolute error, smallest variance of error, smallest root mean square error, the highest correlation coefficient between the predicted and observed values and the smallest correlation between the errors and observed values. The ANUSPLIN routine was used to generate solar radiation exposure estimates per square kilometer across U.S. that will further serve to calculate individual UV exposure and melanoma risk.

Author Keywords: UV exposure, GIS, modeling, interpolation, splines.

1. Introduction

1.1 Scope and Objectives

Melanoma is one of the most rapidly increasing cancers among the white population in the United States (Krickler and Armstrong 1996; Beddingfield 2003). Studies consistently point to ultra violet (UV) exposure as the most important risk factor for melanoma in those individuals with a phenotypic susceptibility and, to a lesser extent ozone depletion (Elwood 1989; Elwood and Koh 1994; Elwood and Jopson 1997; De Fabo et al. 2004). Yet, the evidence has often been considered weak and conflicting due to a lack of accurate UV exposure assessment. The results varied from study to study likely because the methods used to estimate UV exposure were often varied and imprecise (Fears et al. 2002). One way to approach this problem is to assess the performance of different spatial interpolation techniques in estimating UV exposure.

The aim of present research is to evaluate performance of Thiessen-polygon and Kriging procedures from the standard Geographic Information Systems (GIS) package based on the magnitude and distribution of errors, and compare results with Australian National University Splines (ANUSPLIN) routine that runs outside typical GIS through a series of C++ and FORTRAN commands. The idea to step outside the typical GIS toolbox and use Spline routine for comparison is stimulated by the success of this procedure in predicting climate data, principally precipitation and temperature (Hutchinson 1991b, 1993, 1995, 1998). It is expected that Spline interpolation may prove as effective in the analysis of solar radiation exposure.

The errors of prediction are evaluated using fundamental statistical parameters such as mean absolute error, root mean square error, variance of errors, etc. GIS technology is particularly well suited to communicate the results of this type of analysis because it enables visualization of geographical distribution of errors via maps, calculation of statistical parameters,

and their comparison through charts and graphs. The paper concludes by identifying interpolation method that produces minimal prediction error and delivers UV exposure estimates that would further be utilized for assessment of melanoma risk.

1.2 Studies of UV Exposure

The earliest epidemiologic evidence that solar radiation might play a role in the etiology of melanoma resulted from ecological studies that demonstrated the inverse relationship between latitude of residence and incidence and mortality rates for melanoma in predominantly white populations (Lancaster and Nelson 1957; Fears et al. 1976; Holman et al. 1980; Scotto et al. 1982). Lancaster and Nelson (1957) linked the general latitude association to melanoma mortality with the variations in sunlight and UV radiation and concluded that the distribution of melanoma among fair-skinned populations was consistent with a hypothesis of excess sunlight as an important predisposing cause of melanoma. However, the ecological studies have been subject to many assumptions, and thus the results must be interpreted in light of the many limitations inherent in the study design. Specifically, the lack of individual exposure measures in these studies made it impossible to establish a causal association between melanoma and UV radiation (Elwood and Jopson 1997).

The results from the ecological studies of melanoma and latitude led to case-control studies, which further investigated the hypothesis that exposure to UV radiation is causally associated with melanoma using estimates of cumulative sun exposure (Lee 1989). However, in these studies individual exposure has been particularly difficult to quantify, since both timing and magnitude were thought to be important, and neither was easily documented (Fears et al. 2002). The estimates of sun exposure were defined somewhat differently from study to study, but all were based on individual interviews or questionnaires and, as such, provided an assessment of

individual exposure (Longstreth 1987). The conclusions generated based on data from interviews were tempered by the possibility of systematic recall bias and random misclassification (Cockburn et al. 2001).

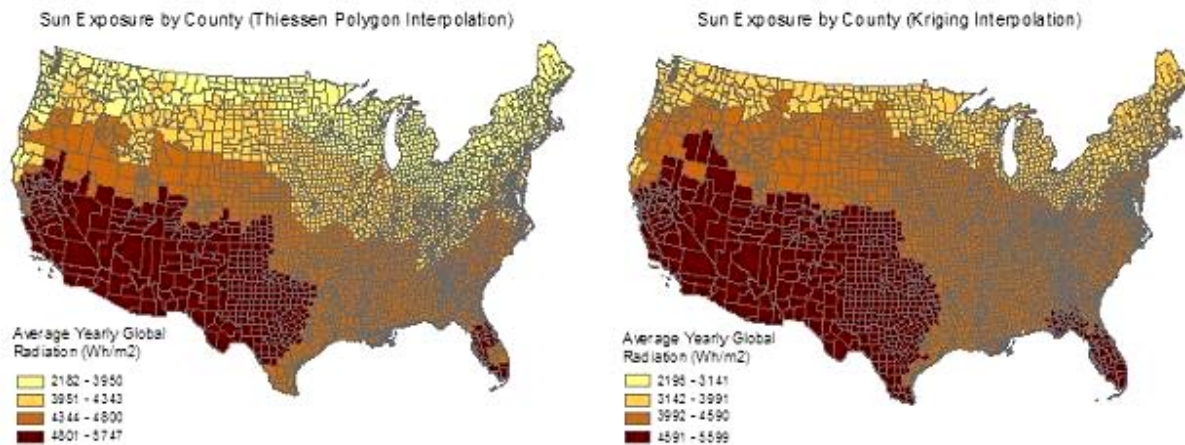
To avoid complications associated with recall bias Armstrong (1984) designed a different approach, lead by the hypothesis that individual risk for melanoma is associated with lifetime residential history and the strength of sunlight at those residences. The measure of cumulative sun exposure was based on location and duration at each residence and mean annual hours of bright sunshine at each location, resulting in an estimate for lifetime exposure at home rather than an actual estimate of time spent in the sun. Fears and colleagues (2002) have applied the same approach using the average annual UVB intensity at places of residence to estimate lifetime exposure. Exposure estimates were derived from 11 years of ground level measurements from RB meters at 30 stations using regression equations.

The review of research studies on UV exposure leads to conclusion that although some progress has been made towards new and improved methods for individual UV exposure assessment, little attention has been given to the issue of *uncertainty* in UV estimates, from which measures of individual exposure are derived. Consequently, the results generated in recent studies of individual UV exposure, although possibly free of recall bias, may still be tempered by the lack of accuracy of exposure estimates. This leads to question regarding how well can we model UV exposure given the methods and measurement network available to us at this time. A way to address this question may be to examine the performance of different interpolation methods for UV exposure assessment.

1.3 Spatial Interpolation Methods

Spatial interpolation is a procedure of estimating the values of properties at unsampled locations based on the set of observed values at known locations (Burrough 1986). A large number of interpolation methods have been developed for use with point, line, and area data. Spatial interpolation techniques are reviewed in detail by Lam (1983), Burrough (1986), Meyers (1994), Burrough and McDonnell 2000), and Cressie (2003). No matter which interpolation technique is used, the derived values are only estimates of what the real values should be at a particular location. The quality of any analysis that relies on interpolation of observed data is, therefore, subject to a degree of uncertainty (Chiles and Delfiner 1999). Different interpolation methods can therefore generate different predictions at same locations. As an example, two solar radiation exposure maps (Figure 1) resulted from two different interpolations performed on the same dataset: Thiessen polygon and kriging.

Figure 1. Thiessen Polygon and Kriging Exposure Models



The two maps are obviously *different*, and it is impossible to tell which is more accurate. Assessment of uncertainty (error) associated with interpolation techniques available in most GIS packages has been done previously, and the results suggest that Kriging, IDW, Thiessen polygons and TIN interpolations performed almost on the same level (Siska and Hung 2005). Of

interest to present research is to compare two interpolation techniques commonly used in GIS – Thiessen polygons and Kriging – and see how they perform against Spline interpolation that is run outside the typical GIS toolbox.

Thiessen polygons, also referred to as the Dirichlet Tessellations or the Voronoi Diagrams, are an exact method of interpolation that assumes that the values of unsampled locations are equal to the value of the nearest sampled point. This method is commonly used in the analysis of climatic data when the local observations are not available, and so the data from the nearest weather stations are used. Thiessen polygons define the individual 'regions of influence' around each of a set of points such that any location within particular polygon is nearer to that polygon's point than to any other point, and therefore, has the same value (Heywood et al. 1998). Thiessen polygon interpolation is available within common GIS packages, such as for example ESRI's Arc GISⁱ. It is relatively simple procedure that requires point coverage for input and is executed using the ArcGIS toolbox command. A major difficulty with Thiessen-polygon approach is that the measures are assumed to be more homogenous within units (polygons) and to change values only at the boundaries. Since there is only one observation per polygon no within-area variation can be estimated (Gold 1991).

Kriging is a geostatistical method that uses known values and a semivariogram to predict the values at some unmeasured locations. Semivariance is a measure of the degree of spatial dependence between samples. The magnitude of the semivariance between points depends on the distance between the points. With kriging therefore, predicted values are not the same as the “source” point (like in Thiessen approach) but rather vary depending on their proximity to the source. The semivariogram model that best fits the data is developed to produce the optimum weights for interpolation (Burrough and McDonnell 2000). The derivation of semivariogram in

kriging is discussed in greater detail in Cressie (2003). Most GIS packages offer several kriging models: simple, ordinary, universal, indicator, disjunctive, and probability. Each type has its counterpart in cokriging – a multivariate equivalent to kriging. Of interest to this analysis are the first three types of kriging that all assume normal distribution of data, but make different assumptions about the mean value of the variable under study. Simple kriging requires a constant known mean for input to model, ordinary kriging assumes constant but unknown mean, universal kriging assumes a varying mean over space (Krivoruchko and Gotway 2004). Johnston et al. (2001) suggest that it is best to stay with ordinary kriging unless there is a good reason to stray. If, for example there is a need to account for the trends observed in the exploratory data analysis, then universal kriging may be used. Another option available within GIS environment is to perform *detrending* with kriging. This is useful when there is a need to remove a surface trend from the data (e.g. spatial dependence of values on elevation) to better understand the real surface values, and then use kriging or cokriging on the residual (detrended) data.

Splines technique has been described by Wahba (1980) and computationally developed by Hutchinson (1991a,) for use with climate data principally in mind. Thin plate smoothing splines can be viewed as a generalization of standard multivariate linear regression, in which the parametric model is replaced by a suitably smooth non-parametric function. The degree of smoothness, or inversely the degree of complexity, of the fitted function is usually determined automatically from the data by minimizing a measure of predictive error of the fitted surface given by the generalized cross validation (GCV) (Craven and Wahba 1979; Hutchinson and Gessler 1994). A comparisons of spline technique with kriging are given by Hutchinson (1991b, 1993, 1995), Hutchinson and Gessler (1994), and Laslett (1994). Whereas thin plate splines are defined by minimizing the roughness of the interpolated surface with prescribed residuals from

the data, kriged surfaces are defined by minimizing the variance of the error of estimation, which normally depends on the preliminary semi-variogram analysis (Hutchinson and Gessler, 1994). The latter is seen as primary disadvantage of Kriging: “This structure can be difficult to estimate and validate”... “The method is hampered by ad hoc assumptions about the form that the variogram should take and the computational difficulties in assessing the merit of different functional forms” (Hutchinson, 1991b, p. 106)

The Australian National University Spline (ANUSPLIN) package has been designed to provide a facility for transparent analysis and interpolation of noisy multi-variate data using thin plate smoothing splines (Hutchinson, 2003). The package supports this process by providing comprehensive statistical analyses, data diagnostics and spatially distributed standard errors. It also supports flexible data input and surface interrogation procedures. ANUSPLIN package is made up of eight programs: SPLINA, SPLINB, SELNOT, ADDNOT, DELNOT, GCV GML, LAPPNT, and LAPGRD. Detail description of the programs is given by Hutchinson (2003).

Some advantages of ANUSPLIN in precipitation modeling are summarized by Custer et al. (1996) as follows:

1. The approach is simple and computer storage requirements are comparatively modest, although the routine is computationally complex and requires workstation.
2. The routine does not depend on prior climatologically extracted covariances and does not assume that the spatial covariance is stationary.
3. ANUSPLIN does not depend on subjective decisions about weighting functions because the program minimizes the generalized cross-validation (GVC) automatically. Regions of similar or related topography do not have to be identified by operator. Similarly, no radius of influence of an input precipitation (solar radiation) station is required, and the spatial distribution of the points may be irregular. Kriging has these requirements.

4. ANUSPLIN works well in areas where foreknowledge about study region is limited.
5. ANUSPLIN allows inclusion of both partial and complete records of precipitation as input.
6. The predicted precipitation is calculated based on existing precipitation records, latitude, longitude, and elevation. The elevation component is particularly important in mountainous terrain where precipitation is produced as air masses lift over mountains.”

The advantages of ANUSPLIN in precipitation modeling make it seem equally suitable for modeling solar radiation. Specifically, the last remark about the importance of elevation component is applicable to solar radiation that also varies with differences in elevation and different features of the receiving terrain (Wilson and Gallant 2000). High elevations are generally exposed to more solar radiation and with it, more UV radiation than low elevations because higher altitudes have lower atmospheric pressure and lower amounts of air molecules and dust, which altogether causes less scattering of UV than at lower elevations. In general, each 1 km (kilometer) increase in altitude increases the ultraviolet irradiance by about 6% (Blumthaler et al. 1997). “ The incorporation of dependences on elevation and landscape shape has played a major role in developing accurate spatial representation of environmental variables such as surface climate and in assessing spatially detailed impacts of projected climate change” (Hutchinson 1995).

Thiessen polygon, kriging and spline procedures are evidently very different not only statistically, but also in terms of computational complexity, and their ability to incorporate additional variables, all of which may differentially affect the predictions. One way to assess the performance of different exposure models would be to examine the magnitude and distribution of prediction errors.

2. Data and Methods

2.1 Solar Radiation Data

National Solar Radiation Database (NSRAD) produced by National Renewable Energy Laboratory (NREL) under the Department of Energy's (DOE) Resource Assessment Programⁱⁱ served as a source of solar radiation data for input in each interpolation procedure. NSRAD contains statistical summaries computed from hourly data for the 30 years from 1961 to 1990 for 239 radiation stations across the United States. The statistics include the average and standard deviation of the daily total solar energy (direct normal, diffuse horizontal, and global horizontal) for each station month, year, and 30-year averages. NSRAD database is available for download from the NREL websiteⁱⁱⁱ.

This project used a measure of 30-year Average daily total Global solar radiation (AVGLO), defined as total amount of direct and diffuse solar radiation in Wh/m² (Watt hours per square meter) received on a horizontal surface. The decision to use 30-year average was based on the preliminary analysis of temporal variability that showed no statistically significant difference in AVGLO measures between three 10-year periods for each radiation station.

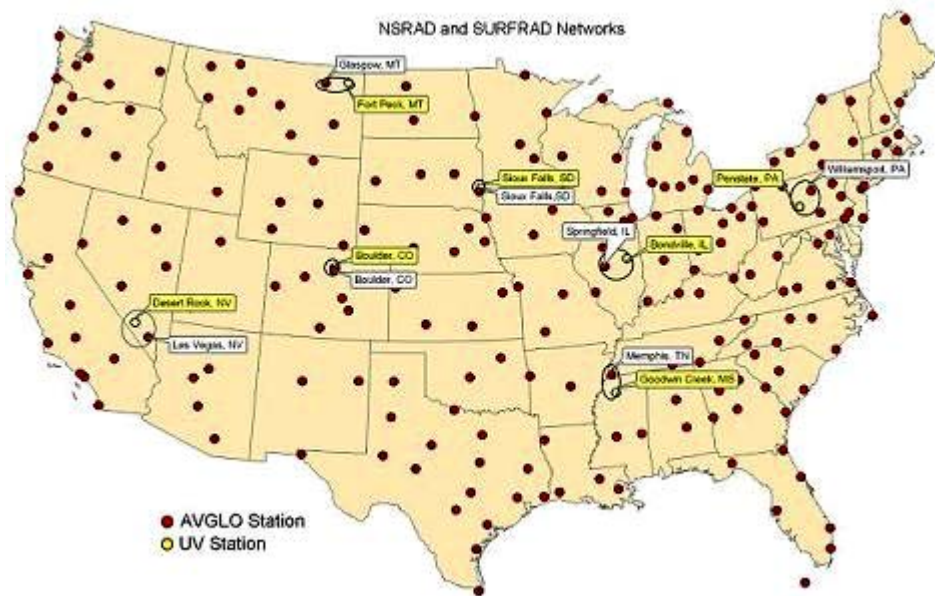
2.1.2 AVGLO as Proxy for UV

A possibility to use AVGLO measures as a proxy for UV was examined, because the quantity of AVGLO data was larger and more complete than any UV database available at the time. Monthly UVB data from the Surface Radiation Budget Network (SURFRAD) stations of the National Oceanic and Atmospheric Administration (NOAA) were used for this type of analysis. Detailed description of SURFRAD database is given at NOAA web site^{iv}. In summary, UVB flux is given as total UVB convoluted with the erythemal action spectrum, (i.e. that part of the UVB spectrum responsible for sun burns on human skin (erythema) and DNA damage). The

erythemal UVB irradiance is computed for 300 Dobson units of ozone. This is done because the ozone value over the stations is unknown during the near real-time processing.

To use AVGLO data as a proxy for UV it was necessary to demonstrate that their relationship is strong. UVB measures downloaded for this project include monthly UVB averages in mW/m^2 (milli Watts per square meter), covering the period 1995-2004 recorded at 7 SURFRAD stations that are operating in climatologically diverse regions: Montana, Colorado, Illinois, Mississippi, Pennsylvania, Nevada and South Dakota. A sample of 7 stations that measure global radiation was selected based on proximity to UV stations (Figure 2).

Figure 2: Sampling NSRAD Stations Based on Proximity to UV Stations



The sample of 7 stations was tested with “signed test for median” and found to be representative of population of values covered by 215 NSRAD stations ($p = 0.57$). Correlation analysis was performed to evaluate relationship between the monthly UVB values recorded by SURFRAD network and the AVGLO values recorded at 7 closest representative stations. 7 correlation coefficients were generated (one for each pair of records at matched locations) and a

corresponding r^2 coefficients. For every pair of stations $r^2 > 0.97$, which indicates that measures of global radiation (AVGLO) correspond to measures of UV at similar locations and could, therefore, be used as a proxy measures for UV radiation in this research.

2.2 Digital Elevation Model Data

30 arc-second Digital Elevation Models (DEMs) for four quadrangles covering North America was obtained via web site^v from the National Geophysical Data Center (NGDC) of the National Oceanic and Atmospheric Administration (NOAA). The DEMs were converted from image to grid, projected to geographic projection WGS84, edge-matched, and then joined in ARC/INFO GRID module to create a single DEM. The DEM was further clipped to cover only the Continental United States with 9600 cells on x-axis, 3120 on y-axis, and 4256 records of elevation in Kilometers (Km). This DEM was generated for use with ANUSPLIN (LAPGRD program).

2.3 Methods

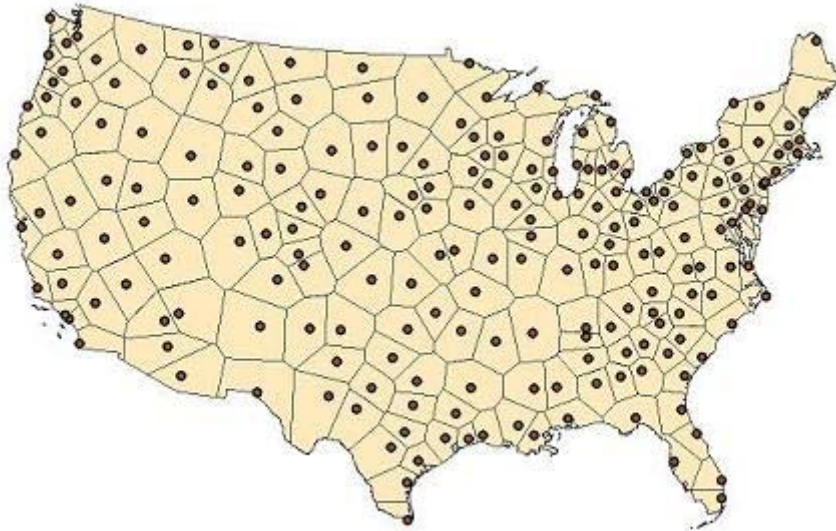
The three interpolation procedures used the same radiation station dataset comprised of 215 radiation stations in the Continental United States - longitude, latitude, and elevation at station location - to generate a set of predicted values at known locations. In each instance the predicted values were generated by systematically removing the input data for one radiation station, then calculating radiation for that station based on other station(s). The routine used to accomplish this task differs from one method to another.

2.3.1 Thiessen-polygons

Thiessen-polygons procedure was run using a Thiessen command within ArcGIS9 toolbox on point coverage with 215 radiation stations. The resulting coverage was clipped to the shape of continental United States. Thiessen map (Figure 3) suggests that global radiation at all

locations within each polygon “belongs” to a particular radiation station within that polygon and therefore, has the same value.

Figure 3. Thiessen Map



To test the performance of Thiessen-polygon approach the predictions were generated for known locations by removing one data point at a time and calculating radiation for that station using the nearest station. The idea was to determine how well the neighboring station estimates the missing value. 215 point-coverages were generated, each with a different data point removed, then Thiessen-polygons procedure was performed 215 times on all coverages to generate predictions of each missing value. The predicted values were then recorded in the attribute table that contains the matching observed values for all data points in the original coverage. The actual, absolute, and percent differences between observed and predicted values (error) were calculated within ArcGIS for further statistical analysis described later in this section.

2.3.2 Kriging

Kriging procedure was performed using the Geostatistical analyst extension of ESRI's ArcGIS9. The input data included point coverage with 215 average global radiation (AVGLO) values, latitude, longitude, and elevation at 215 locations. As earlier noted, kriging procedure requires a lot of decision making in terms of kriging model type, and in case of detrending, neighborhood parameter adjustments. What helps the decision-making is to examine the data set in order to understand the distribution and any large-scale trends. The exploratory data analysis suggested that dataset is normally distributed and that there is a trend in x and y directions. It is likely that this trend is a result of differences in elevation between the data points, as discussed in the introductory section. Six types of models were run consecutively each with different parameter adjustments to arrive to the best model prediction for each: Ordinary kriging, Universal kriging, Ordinary detrended kriging, Universal detrended kriging, Ordinary cokriging (with elevation), and Universal cokriging. Detrending was not part of all model runs because, generally it is best to keep the kriging (cokriging) model as simple as possible and only to remove trend if significantly improves results (Krivoruchko and Gotway 2004).

The Geostatistical analyst provided an indication of which model produces minimal prediction error through cross-validation. Cross-validation uses all of the data to estimate trend and autocorrelation models, and then removes each data location, one at a time, and predicts the associated data value. Cross-validation matrix provides measures of error, absolute error, and standard error of prediction. Generally, the best model is the one that has standardized mean nearest to zero, the smallest root mean square prediction error, the average standard error nearest to root mean square prediction error, and the standardized root mean square prediction error

nearest to one (Johnston et al. 2001). Table 1 summarizes the best prediction statistics for each of the six models.

Table 1: 6 Kriging Models: Prediction Statistics

STATISTICS	KRIGING		DETRENDENT KRIGING		COKRIGING	
	Ordinary	Universal	Ordinary	Universal	Ordinary	Universal
Root Mean Square Error	104	104	108	116	112	106
Average Standard Error	159	112	118	201	163	113
Mean Standardized	-0.005	-0.009	-0.005	-0.03	-0.003	-0.006
Root Mean Square Error Standardized	0.63	0.89	0.88	0.56	0.67	0.9

A comparison of error statistics for 6 kriging models suggests that Universal kriging and Universal cokriging performed at the same level, generating the best statistics. Having the slightly lower root mean square error, Universal kriging model was selected for comparison with Thiessen and ANUSPLIN models.

2.3.3 ANUSPLIN

ANUSPLIN routine utilized FORTRAN program SPLINA to calculate predictions of global solar radiation using partial thin plate smoothing splines on 215 data points. This program is suitable when there are less than 350 data points per data set, in contrast to SPLIB that is used for larger datasets. The SPLINA program required two input files. The first user-directive file contained the number of independent spline variables (2) – latitude and longitude, number of independent covariates (1) – elevation, lower and upper limits for longitude and latitude of the area covered by radiation stations, elevation in kilometers, order of spline (2), number of surfaces (1) – radiation values, number of data points (215), and input-output parameters specified in the directive. The second ASCII file contained the annual radiation means for the period 1961-1990 (Wh/m²), the station locations in decimal degrees of longitude and latitude, and elevation (all derived from NSRAD).

SPLINA generated numerous diagnostics in addition to an ASCII file containing the surface coefficients summarizing the relationship between mean global solar radiation, latitude/longitude and elevation. The surface diagnostics include generalized cross validation (GCV) estimate, a mean square error (MSE) of the smoothed data values, a mean square residual (MSR), a mean relative error variance estimate (VAR), and their square roots. The GCV is a measure of the predictive error of the fitted surface, calculated by removing each data point in turn and summing the square of discrepancy of each omitted data point from a surface fitted to all the other data points (Hutchinson 2003).

2.4 Model Evaluation

The accuracy of model predictions generated by Thiessen-polygon, Kriging and Splines procedures in this research was assessed based on the magnitude and distribution of errors – the difference between observed values and model predicted values - in four ways:

- (1) The root mean square error (RMSE) was calculated for each model prediction using the formula:

$$RMSE = \sqrt{\frac{SSE_i^2}{n}}$$

where SSE is sum of squared errors (observed - estimated values) and n is the number of pairs (errors). RMSE is frequently used as

an important parameter that indicates the accuracy of spatial analysis in GIS and remote sensing (Siska and Hung 2005).

- (2) The mean absolute error (MAE), the average absolute difference between observed and predicted values for the 215 data points was computed along with variance of errors (VE). Large MAE values and comparably large VE indicate larger discrepancies between predicted and observed values.

(3) Correlation coefficients were calculated between observed and predicted values and between errors and observed values. Better model performance is indicated by higher coefficients between observed and predicted, and lower coefficients between errors and observed values.

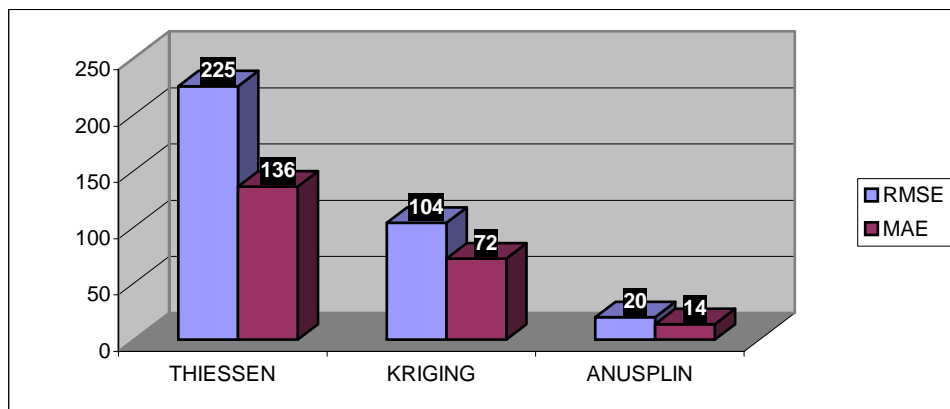
(4) The model predictions were also checked for the differences in spatial distribution of absolute errors. This analysis was important to identify the regions where the largest discrepancies occurred. For each model the absolute differences between observed and predicted values were plotted on a separate map for visual comparison.

3. Results

3.1 Model Predictions

(1) The results based on RMSE indicate that the interpolation using ANUSPLIN yielded the smallest errors while the Thiessen polygon indicated the highest RMSE. ANUSPLIN interpolation was 11.25 times more accurate than the Thiessen polygon and 2.2 times than Universal kriging method (Figure 4).

Figure 4: Differences in Model Predictions



(2) The mean absolute error values also indicate that ANUSPLIN performed better than any other interpolation method on this data. On average the ANUSPLIN error was 14.2 Wh/m² (or 0.34%). The Universal kriging results generated the mean absolute error of 72.5 Wh/m² (1.68%), while Thiessen polygon interpolation indicated the mean absolute error of 136 Wh/m² (3.16%). Additionally, ANUSPLIN model generated the smallest variance of errors, whereas Thiessen polygon predictions had the highest variance of error. Both measures taken together confirm that ANUSPLIN produced the smallest uncertainties in predicting values at known locations (Table 2).

(3) ANUSPLIN routine also produced highest correlation coefficient between observed and predicted values, and lowest correlation coefficients between errors and observed values.

Table 2: Error Statistics

	RMSE	MAE	VE	Correlation OBS/PRED	Correlation OBS/ERR
THIESSEN	225	136	50906	0.90	0.40
KRIGING	104	72	10885	0.98	0.31
ANUSPLIN	20	14	384	0.99	0.15

(4) Plotting the distribution of absolute errors on a map was helpful to understand their geographical distribution. In Thiessen polygon interpolation the largest prediction errors occurred for those radiation stations that are located in areas with sharp differences in elevation. The absolute errors range between 0.5% and 16% (Figure 5). In Universal kriging model, the greatest errors are distributed in similar locations as with Thiessen-polygon model, that is, in areas with sharp elevation gradient, although kriging yield generally better predictions, ranging from 0 to 13% (Figure 6). In contrast, errors produced by ANUSPLIN model are relatively uniformly distributed across the region and do not exceed 1.7% (Figure 7).

Figure 5. Thiessen Polygon Model: Distribution of Error

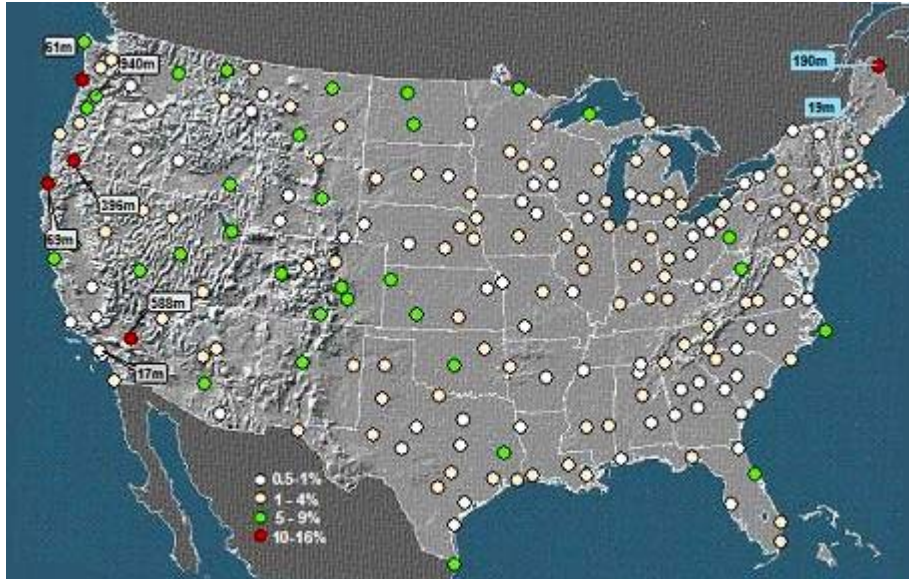


Figure 6. Kriging Model: Distribution of Error

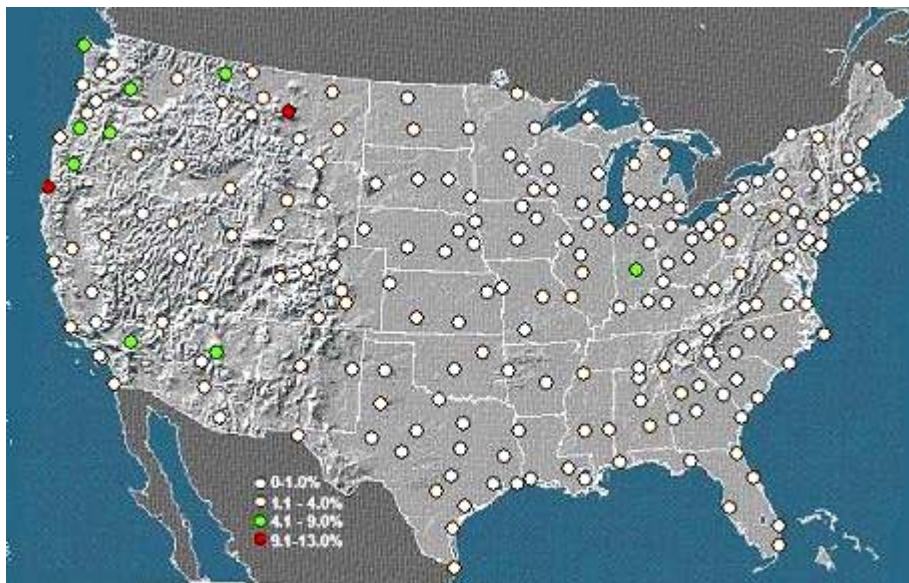
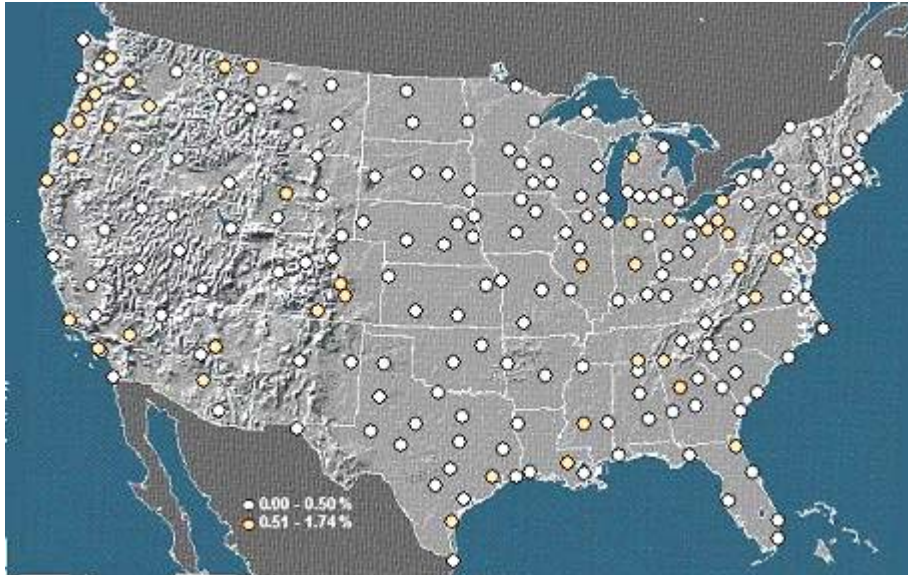


Figure 7. ANUSPLIN Model: Distribution of Error



3.2 Final Exposure Model

Having been identified as superior over Thiessen polygon and Kriging interpolation, ANUSPLIN was used once again, to generate predictions across United States at unsampled locations. This time the routine utilized FORTRAN program LAPGRD that uses the SPLINA output surface coefficients file and DEM data to interpolate mean annual global radiation across geographic and elevation gradient.

The LAPGRD program required three input files: the surface coefficients file output from SPLINA, the 1km² DEM in ASCII format, and a user-directive file containing, elevation and location bounds, grid cell size, special value options for cells with no data, and input-output file parameters. LAPGRD combines the surface coefficients with DEM to estimate solar radiation values at each DEM grid node. This program generated an ASCII file with elevation and solar radiation estimates that was transferred to ARC/INFO and converted to grid with ASCIIGRID command. This grid was used to generate a map of solar radiation exposure per 1km² (Figure 8).

Figure 8. ANUSPLIN Exposure Model



Selected in a square in Figure 8 is the Island of Santa Catalina in California, which was chosen to exemplify, on a 3D larger-scale map, a subtle changes in exposure that are not as obvious on the previous map (Figure 9). Variations in average annual solar radiation values on Santa Catalina range from 4812 to 4875 Wh/m² (a difference of about 1%).

Figure 8: Santa Catalina 3D Exposure Model



4. Conclusion

In this paper the performance of Thiessen-polygon and Kriging procedures from the standard GIS package (ArcGIS9) was evaluated based on the analysis of errors, and the results compared with ANUSPLIN spline routine that runs outside typical GIS toolbox. As the statistical analysis indicated, kriging performed slightly better than Thiessen polygons, whereas ANUSPLIN proved by far superior among the three routines in predicting values in unsampled locations on a non-uniform (elevation-dependent) surface such as radiation. ANUSPLIN model produced results with the smallest mean absolute error, smallest variance of error, smallest root mean square error, the highest correlation coefficient between the predicted and observed values and the smallest correlation between the errors and observed values. All of these measures

indicate the least amount of uncertainty associated with the model predictions. In addition, ANUSPLIN proved to be relatively simple and computationally efficient procedure for interpolating solar radiation values at such high resolution (of 1 km²) for the region as large as Continental United States.

The value of the resulting digital exposure map is in ability to uncover subtle changes in solar radiation exposure. This may be useful for practical and other research applications, such as for example, for assessment of individual cumulative sun exposure. Further research by author will use the ANUSPLIN model predictions to calculate individual historical UV exposure measures in the case-control setting based on residential histories. The research will link interdisciplinary expertise to help advance our understanding of the role of UV exposure and other biological and lifestyle factors in melanoma etiology.

Acknowledgements:

This work would not have been possible without generous support of my adviser John P Wilson who provided not only advisement and numerous innovative ideas, but the opportunity and funding for this research, including purchase of ANUSPLIN software. The research also greatly benefited from discussions with Myles Cockburn who offered insight into epidemiological side of UV exposure research and suggested appropriate references.

References:

- Armstrong B K 1984 Melanoma of the skin. *British Medical Bulletin* 40: 346-50
- Beddingfield F C III 2003 Melanoma and cutaneous malignancies. *The Oncologist* 8: 459-65
- Blumthaler M, Ambach W, Ellinger R 1997 Increases in solar UV radiation with altitude. *Journal of Photochemistry and Photobiology B: Biology* 39: 130-4
- Burrough P A 1986 *Principals of Geographical Information Systems for Land Resources Assessment*. Oxford, Clarendon Press
- Burrough P A, McDonnell R 2000 *Principles of Geographical Information Systems*. New York, Oxford University Press
- Chiles J P and Delfiner P 1999 *Geostatistics: Modeling Spatial Uncertainty*. New York, John Wiley and Sons
- Cockburn M, Black W, McKelvey W, Mack T 2001 Determinants of melanoma in a case-control study of twins (United States). *Cancer Cases and Control* 12: 615-25
- Craven P and Wahba G 1979 Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numerical Mathematics* 3: 377-403
- Cressie N 2003 *Statistics for Spatial Data*. Revised Edition, New York, John Wiley & Sons
- Custer S G, Farnes P, Wilson J P, Snyder R D 1996 A comparison of hand-and spline-drawn precipitation maps for mountainous Montana. *Water Resources Bulletin* 2: 393-405
- De Fabo E C, Noonan F P, Fears T, Merlino G 2004 Ultraviolet B but not ultraviolet A radiation initiates Melanoma. *Journal of Cancer Research* 64: 6372-6
- Elwood J M 1989 Epidemiology of melanoma: Its relationship to ultraviolet radiation and ozone depletion. In Jones R R and Wigley T (eds) *Ozone Depletion: Health and Environmental Consequences*. London, John Wiley & Sons
- Elwood J M, Koh H K 1994 Etiology, epidemiology, risk factors, and public health issues of melanoma. *Current Opinion in Oncology* 6: 179-87
- Elwood J M, Jopson J 1997 Melanoma and sun exposure: An overview of published studies. *International Journal of Cancer* 73: 198-203
- Fears T R, Scotto J, Schneiderman M A 1976 Skin cancer, melanoma and sunlight. *American Journal of Public Health* 66: 461-64

- Fears T R, Bird C C, Guerry D, Sagebiel R W, Gail M H, Elder D E, Halpern A, Holly E A, Hartge P, Tucker M A 2002 Average UVB flux and time outdoors predict melanoma risk. *Cancer Research* 62: 3992-96
- Gold C M 1991 Problems with handling spatial data: the Voronoi approach. *Canadian Institute of Surveying and Mapping Journal* 45: 65-80
- Heywood I, Cornelius S, Carver S 1998 *An Introduction to Geographical Information Systems*. New Jersey, Prentice Hall
- Holman C D J, James I R, Gattey P H, Armstrong B K 1980 An analysis of trends in mortality from malignant melanoma of the skin in Australia. *International Journal of Cancer* 26: 703-9
- Hutchinson M F 1991a *ANUDEM and ANUSPLIN User Guides*. Canberra, Australian National University, Centre for Resource and Environmental Studies Miscellaneous Report 91/1
- Hutchinson M F 1991b The application of thin plate smoothing splines to continent-wide data assimilation. In Jasper J D (ed) *Data Assimilation Systems*. Melbourne, Australian Bureau of Meteorological Research Report 27
- Hutchinson M F 1993 On thin plate splines and kriging. *Computing and Science in Statistics* 25: 203-10
- Hutchinson M F and Gessler P T 1994 Splines-more than just a smooth interpolator. *Geodema* 62: 45-67
- Hutchinson M F 1995 Interpolation of mean rainfall using thin plate smoothing splines. *International Journal of Geographic Information Systems* 9: 385-403
- Hutchinson M F 1998 Interpolation of rainfall data with thin-plate smoothing splines I: Two dimensional smoothing data with short range correlation. *Journal of Geographical Information and Decision Analysis* 2: 152-67
- Hutchinson M F 2003 ANUSPLIN Version 4.3. Canberra, Australian National University, Centre for Resource and Environmental Studies
- Johnston K, Ver Hoef J M, Krivoruchko K, Lucas N 2001. Using ArcGIS Geostatistical Analyst. Redlands, ESRI
- Kricker A, Armstrong B K 1996 International trends in skin cancer. *Cancer Forum* 20: 192-5
- Krivoruchko K Gotway C A 2004 Creating exposure maps using kriging [on-line]
http://www.esri.com/software/arcgis/arcgisextensions/geostatistical/research_paprs.html
- Lam S 1983 Spatial interpolation methods: a review. *American Cartographer* 10: 129-49

Lancaster H O, Nelson J 1957 Sunlight as a cause of melanoma: A clinical survey. *Medical Journal of Australia* 6: 452-56

Laslett M 1994 Kriging and splines: and empirical comparison of their predictive performance in some applications. *Journal of the American Statistical Association* 89: 391-409

Lee J A 1989 The relationship between malignant melanoma of skin and exposure to sunlight. *Journal of Photochemistry and Photobiology* 50: 493-96

Longstreth J D (ed) 1987 *Ultraviolet radiation and melanoma-with a special focus on assessing the risks of stratospheric ozone depletion*. U.S. Environmental Protection Agency, Washington D.C.

Meyers D E 1994 Spatial interpolation: an overview. *Geoderma* 62: 17-28

Scotto J, Fears, T R, Fraumeni J F Jr. 1982 Solar Radiation. In Shottenfeld D, Fraumeni J F Jr. (eds) *Cancer Epidemiology and Prevention*. Philadelphia, W.B. Saunders Company

Siska P P, Hung I K 2005 Assessment of kriging accuracy in the GIS environment [on-line] <http://gis.esri.com/library/userconf/proc01/professional/papers/pap280/p280.htm>

Wahba G 1980 Spline bases, regularization and generalized cross validation for solving approximation problems with large quantities of noisy data. University of Wisconsin, Madison: *Statistics Department Technical Report* 597: 1-8

Wilson J P and Gallant C 2000 Secondary topographic attributes. In Wilson J P and Gallant C (eds) *Terrain Analysis*. New York, John Wiley and Sons: 87- 131

Endnotes:

ⁱ ESRI Environmental Systems Research Institute, Inc., 380 New York Street, Redlands, California 92373-8100

ⁱⁱ NSRAD User's Manual is prepared by National Renewable Energy Laboratory, 1617 Cole Blvd., Golden, CO, 80401. Distributed by: National Climatic Data Center Federal Building, Asheville, North Carolina, 28801

ⁱⁱⁱ redc.nrel.gov/solar/old_data/nsrdb/dsf

^{iv} <ftp://ftp.srrb.noaa.gov/pub/data/surfrad/>

^v www.ngdc.noaa.gov