Proceedings of Spatial Accuracy 2018

Editors:

Yong Ge, Xin Li & Gerard Heuvelink

Organized by:
Reducing the Uncertainty in Precipitation Forecasting among Statistical and Dynamic Models
Lei Xu, Nengcheng Chen ................................................................. 1
Accounting for Access Costs in Optimal Stratification for Validation of Soil Maps
Dick Brus, Lin Yang, A-Xing Zhu ......................................................... 5
A Statistical Framework of Functional Data Analysis for Spatiotemporal Analysis and Modeling
Guoqiang Cao .......................................................... 6
Cross-Comparison between GF-2 PMS2 and ZY-3 MUX Sensor Data
Wu Xiao-qing, Xu Han-qiu ................................................................. 7
Spatial Dependence Modelling for Object-based Classifications of Remote-Sensing Images
Yunwei Tang, Jingxiong Zhang, Linhai Jing, Han Gao, Jiong You, Ying Zhang, Ru Xu ............. 8
Reduction of Systematic Length Error Involved by Polygonal Approximation of Curves using
Polynomial Interpolation Methods
Jean-François Girres ................................................................. 9
Uncertainty in the Relationship between Spatial Autocorrelation and the Modifiable Areal Unit
Problem: A Simulation Study
Sang-Il Lee, Monghyeon Lee, Yongwan Chun, Daniel Griffith ................................. 16
Small Area Estimation of Cancer Rates: A Case Study of Lung Cancer in Florida, 2000-2010
Lan Hu, Yongwan Chun, Daniel Griffith ........................................ 17
The Impact of Locational Uncertainty on Local Geographic Cluster Detection: An Application to
Cancer Data from Florida, 2006-2010
Monghyeon Lee, Daniel A. Griffith, Yongwan Chun ........................................ 18
Object-oriented Classification Based on Deep features for High Spatial Resolution Remote Sensing Images
Han Gao, Linhai Jing, Yunwei Tang, Hui Li, Haifeng Ding ....................................... 25
Comparing the Impact of the Mapping Error on Aggregation Methods
Peijun Sun, Russell G Congalton, Yaozhong Pan ........................................ 26
An Uncertainty-based Approach to Quantify the Spatial Representativeness of Local Health Datasets
Rachel Whitsed, Ana Horta, Amilcar Soares, Herbert Jelinek ................................... 35
Using Multiple Classifiers with AdaBoost in Google Earth Engine for Land Use/Cover
Classification
Yuhao Wang ................................................................. 37
Prediction of Urban PM2.5 Concentrations using a Bayesian Spatial-temporal Modeling Approach
Qiang Pu, EunHye Enki Yoo ................................................................. 39
Dealing with Uncertainty in Real-time Sampling using Portable X-ray Fluorescence
Ana Horta, João Neves, Leonardo Azevedo, Amilcar Soares, Rachel Whitsed .................. 41
Assessing Spatial Uncertainty in 3D Stochastic Mapping of Soil Properties
Leonardo Azevedo, Ana Horta, Amilcar Soares, Tiago Ramos, Rachel Whitsed .................. 43
Geostatistical Prediction of Water Lead Levels in Flint, Michigan: a Multivariate Approach
Pierre Goovaerts ................................................................. 45
Uncertainty Characterization for Cropland Products at the Country Level from Different Sources
Jiong You, Zhiyuan Pei, Haiqi Liu ........................................................... 46
Class-stratified Random Sampling with Sub-stratifications by Edge and Interior Pixels for
Reference Data Collection to Increase Accuracy in Area Estimation
Jingxiong Zhang, Yue Wan, Wangle Zhang .............................................. 47
High Accuracy Surface Modelling
Tianxiang Yue, Zemeng Fan, Zhengping Du, Na Zhao ........................................ 54

Proceedings of Spatial Accuracy 2018
Uncertainties in using Kriging Method to Upscale in situ Soil Moisture to Multiscale Pixel Estimations
Wenjun Zuo, Xiaohu Zhang ................................................................. 55

Study on Topographic Correction Algorithm of Leaf Area Index Products Derived from Remote Sensing Data
Yueting Hu, Xianfeng Feng................................................................. 56

The Monitoring Method of Economic Forest in Southern Jiangsu Based on the High-Resolution Remote Sensing Images
Jia Jin, Xianfeng Feng, Zhou Lu, Liang Pei, Ming Luo .......................................... 60

Cross-Calibration of Passive Microwave Satellite Brightness Temperatures Observed by F13 SSM/I and F17 SSMIS for the Retrieval of Thickness and Area on Antarctic Thin Ice
Yue Liu, Xi Zhao, Xiaoping Pang ................................................................... 66

Comparison between Three Split Window Algorithms on Ice Surface Temperature Retrieval in the Arctic
Pei Fan, Xi Zhao, Xiaoping Pang, Qing Ji ......................................................... 67

Exploring Spatial Heterogeneities in Real Estate Market via Geographically Weighted Summary Statistics: a Case Study in Wuhan
Binbin Lu, Jiayi Xie, Wenzhong Shi, Paul Harris .............................................. 68

Uncertainty of endmember on Super-resolution mapping
Xinyan Li, Xiaodong Li, Yihang Zhang, Yun Du, Feng Ling ................................ 70

Robust Variogram Estimation Combined with Isometric Log-ratio Transformation for Improved Accuracy of Soil Particle-sizefraction Mapping
Wenjiao Shi, Zong Wang ............................................................................. 72

Mapping Soil Particle-Size Fractions: A Comparison Of Compositional Kriging and Log-ratio Kriging
Zong Wang, Wenjiao Shi .............................................................................. 73

A Modified Change Vector Approach for Land Use Change Mapping in the Loess Plateau, China, Supporting Critical Zone Science
Alexis Comber, Yihe Lu, Ying Luo, Yanjiao Ren, Paul Harris .......................... 74

Daily Sea Ice Lead Retrieval from MODIS Images and the Uncertainty Handling in Cloud Mask
Meng Qu, Xi Zhao, Xiaoping Pang .................................................................. 77

Adopting Semantic Signatures to Quantify the Uncertainty of Volunteered Geographic Information
Rui Zhu and Krzysztof Janowicz ................................................................... 78

Remote Sensing Accuracy and Resolution for Precision Agriculture in Small Farms, Malawi
Shengpan Lin, Brad Peter, Joseph Messina ......................................................... 79

Improving the Prediction Accuracy of Forest Aboveground Biomass Benchmark Map by Integrating Machine Learning and Spatial Statistics
Shaoqing Dai, Xiaoman Zheng, Shudi Zuo, Yin Ren ........................................ 80

Maximum Consistency Block based Image Segmentation Methods
Hexiang Bai, Feng Cao .................................................................................. 84

High-resolution Mapping of Direct CO2 Emissions and Uncertainties at the Urban Scale
Shaoqing Dai, Shudi Zuo, Yin Ren .................................................................. 88

Comparison between the Accuracy of SMOS Sea Ice Thickness Products and the Thin Ice Thickness Retrieved from AMSR2 Based on Ship Observation Data by Sikuliaq
Xueyi Ma, Xi Zhao, Meng Qu, Zian Cheng ......................................................... 91
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation of Soil Moisture with Land Surface Temperature Retrieving from Landsat-8 Thermal Infrared Data Using Different Algorithms</td>
<td>93</td>
</tr>
<tr>
<td>An Integrated Runoff Risk Support Tool for Predicting Agrochemical Runoff Risk at Field and Landscape Scales</td>
<td>94</td>
</tr>
<tr>
<td>In-city Vegetation Quantification from Online Satellites Tiles Providers using Tungstene Technology</td>
<td>97</td>
</tr>
<tr>
<td>Investigation of the multivariate spatial structure of a soils micronutrients data set at the farm-scale</td>
<td>99</td>
</tr>
<tr>
<td>Alternative interpretations of the particle size distribution of soils (0 -500cm) in the Loess Plateau, China</td>
<td>101</td>
</tr>
<tr>
<td>Spatial-temporal Reconstructing of Forest Percent Cover by Fusing Multiresolution Remotely Sensed Images</td>
<td>107</td>
</tr>
<tr>
<td>A Platform for Sharing and Analysis of Low-cost Environmental Sensor Data</td>
<td>108</td>
</tr>
<tr>
<td>Study on Spatial Interpolation of Daily Precipitation Considering Spatiotemporal Autocorrelation and Non-stationarity: A Case Study over the Upper Reaches of the Bengbu Sluice in the Huai River Basin</td>
<td>110</td>
</tr>
<tr>
<td>A Comparative Study of Class-aggregated Versus Class-specific Logistic Modelling Approaches for Estimating Local Accuracies in Remotely Sensed Land Cover Information</td>
<td>111</td>
</tr>
<tr>
<td>Mapping Change Detection of Lulc on the Cameroonian Shores of Lake Chad and Its Hinterland through an Inter- Seasonal and Multisensor Approach</td>
<td>115</td>
</tr>
<tr>
<td>Efficient Nonstationary Autocorrelated Field Simulation using Convolution Filtering</td>
<td>117</td>
</tr>
<tr>
<td>A Comparison of GWR Software on Various Platforms</td>
<td>123</td>
</tr>
<tr>
<td>Analysis and Application of Drought Characteristics based on Run Theory and Copula Function</td>
<td>124</td>
</tr>
<tr>
<td>Including Spatial Correlation in Structural Equation Modelling of Soil Properties</td>
<td>125</td>
</tr>
<tr>
<td>Evaluation of Heavy Metal Cu Pollution in Soil Based on Cokriging</td>
<td>126</td>
</tr>
<tr>
<td>Accuracy Assessment of Lidar-Derived Dem for Flood Risk Mapping</td>
<td>127</td>
</tr>
<tr>
<td>Evaluating of Terrain Shading Algorithms Using Artificial Surfaces for Solar Radiation Modeling Over Mountainous Terrain</td>
<td>128</td>
</tr>
<tr>
<td>An Interval Uncertainty Propagation Method based on Kernel Density Estimation</td>
<td>129</td>
</tr>
</tbody>
</table>
A Hybrid Approach for Uncertainty Propagation with Random and Interval Variables
Xiaojing Song, Jianghong Ma, Yee Leung ................................................................. 130

Gslpe: a Spatial Upscaling Method for Multipoint P-normal Measurements
Feng Liu, Xin Li .................................................................................................. 132

A Quantitative Assessment of Spectral Unmixing Strategy on the following Subpixel Mapping for Hyperspectral Imagery
Xiong Xu, Xiaohua Tong, Huan Xie ............................................................................. 134

Constrained Total Least Squares Method with Applications in Map Generalization
Yanmin Jin, Xiaohua Tong .......................................................................................... 135

Super Resolution Mapping of Farmland based on Convolutional Neural Network
Yuanxin Jia, Yong Ge, Yuehong Chen, Yan Jin ............................................................ 136

Spatial Sampling Optimization Design for Monitoring Poverty
Mengxiao Liu, Yong Ge, Jianghao Wang, Shan Hu ..................................................... 137

Registration of Point Cloud Data Based on Principle of Least-Squares
Yuefeng Lu, Yong Ge, Kun Chen, Yangyang Fu .......................................................... 138

Validation of High Resolution Global Land Cover Dataset: Methodology and Practice
Huan Xie, Wen Meng, Fang Wang, Ang Zhao, Yali Gong, Xiaohua Tong ....................... 140
Reducing the Uncertainty in Precipitation Forecasting among Statistical and Dynamic Models

Lei Xu¹, Nengcheng Chen*¹

¹ State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China

*Corresponding author: cnc@whu.edu.cn

ABSTRACT

Precipitation forecasts ahead of months are valuable for early drought warnings and agricultural preparations. The existing precipitation forecasting methods include statistical approaches, climate models and hybrid ones. Here, the uncertainties in precipitation forecasts of 313 rain gauge observations in China for a total of 13 statistical and dynamic models were modeled by Bayesian model averaging (BMA) to obtain a hybrid forecast. As the results indicated, a combination of these models did exhibit an advantage over respective models, with the highest Pearson’s correlation coefficient (PCC) and the least mean absolute error (MAE) in the hybrid model, demonstrating potential superiority in ensemble forecasts by incorporating heterogeneous models.

KEYWORDS: Precipitation forecast; Statistical; Dynamic; BMA

I. INTRODUCTION

The present precipitation forecasting methods can be divided into statistical, dynamic and hybrid models (Hao et al. 2018). Statistical methods model the relationship between rainfall observations and relevant variables from historical data, trying to find the general rule for prediction. Dynamic models simulate the real atmosphere-ocean-land interactions based on physical mechanism. The performances of rainfall forecasts in statistical and dynamic models show considerable differences. It is encouraged to combine the statistical and dynamic models to see if they can be combined to obtain better accuracy.

II. EXPERIMENTAL SETUP

In this study, statistical approaches consist of a series of machine learning methods, including support vector machine (SVM), artificial neural network (ANN) and long short-term memory network (LSTM). The detailed model settings were described in Table 1. For SVM and LSTM, precipitation observations in previous six months were used to forecast the precipitation in the target month. In model SVM_C, only some climate indices (Nino 1+2, Nino 3, Nino 3.4, Nino 4 and Multivariate ENSO Index) were adopted to find the relationship between precipitation and teleconnections using SVM. In model ANN, climate indices, precipitation in previous two months, minimum temperature, maximum temperature and average temperature were combined to forecast precipitation in the current month using ANN. For WSVM and WANN models, wavelet transformation was incorporated to preprocess the predictors. Specifically, the Daubechies wavelet of order 5 was used to decompose the precipitation series into one approximation series and three detail series. The decomposed series were then regarded as the predictors together with other
variables to forecast precipitation in the target month. The data used in predictors and predictands spanned from 1960 to 2016. Parameter settings of these methods were conducted by trial and error.

For the dynamic part, six climate models including 64 members were obtained from the North American multi-model ensemble (NMME) to forecast precipitation at station scale. The NMME (Kirtman et al. 2014) is a global seasonal forecasting system by incorporating multiple climate models. The ensemble size of each model varies from 10 to 28, and the lead time ranges between 0.5 and 11.5. Hindcast periods of the NMME start after 1980 and end at 2010, and forecasting periods start at 2011 and extends to present. The 1°×1° resolution data over China were obtained from International Research Institute/Lamont Doherty Earth Observatory (IRI/LDEO) Climate Data Library (http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/). The NMME forecasts were bias corrected to the nearest stations using quantile mapping. Besides, the ensemble mean (EM) of those six climate models was also considered as a model to see the performance of ensemble forecasts. The available time period in dynamic models ranges between 1982 and 2016.

A hybrid model was constructed by weighting the statistical and dynamic models based on Bayesian model averaging (BMA). Since statistical and dynamic models have own strengths and shortages, it seems promising to combine them in order to maximally reduce the uncertainty in precipitation forecasts. BMA (Raftery et al. 1997) is a statistical method to account for model uncertainty and avoid overfitting based on Bayesian model. Suppose \( \mathcal{M} = \{M_1, M_2, \ldots, M_n\} \) is the set of all the alternative models and \( \theta \) is the vector of parameters to be estimated. Then the posterior distribution of \( \theta \) given data \( D \) is

\[
P(\theta|D) = \sum_{k=1}^{n} P(\theta|M_k, D)P(M_k|D)
\]

It is a weighted average of each model according to their posterior probabilities. The posterior probability of model \( M_k \) is obtained by

\[
P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{i=1}^{n} P(D|M_i)P(M_i)}
\]

where

\[
P(D|M_k) = \int P(D|\theta_i, M_k)P(\theta_i|M_i)d\theta_i
\]

is the marginal likelihood, \( P(M_k) \) is the prior probability of \( M_k \), \( P(\theta_i|M_i) \) is the prior probability of \( \theta \) in model \( M_i \). The implementation of the BMA here was based on the Markov chain Monte Carlo Model Composition (MC^3) model, sampling from the model space using Metropolis algorithm. A total of 100,000 burn-ins and 200,000 iterations were conducted to approximate the posterior distribution through a Markov chain.

Table 1. Experimental setup of this study. Note: P: precipitation; CI: climate index; IC: initial condition; DP: decomposed precipitation; Tmin: minimum temperature; Tmax: maximum temperature; Tave: average temperature; LSTM: Long Short-Term Memory network; BMA: Bayesian model averaging.

<table>
<thead>
<tr>
<th>Number</th>
<th>Model</th>
<th>Inputs</th>
<th>Outputs</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SVM</td>
<td>P(t-6), P(t-5), P(t-4), P(t-3), P(t-2), P(t-1)</td>
<td>P(t)</td>
<td>SVM</td>
</tr>
<tr>
<td>2</td>
<td>LSTM</td>
<td>P(t-6), P(t-5), P(t-4), P(t-3), P(t-2), P(t-1)</td>
<td>P(t)</td>
<td>LSTM</td>
</tr>
<tr>
<td>3</td>
<td>SVM_C</td>
<td>CI</td>
<td>P</td>
<td>SVM</td>
</tr>
<tr>
<td>4</td>
<td>ANN</td>
<td>CI, P(t-2), P(t-1), Tmin, Tmax, Tave</td>
<td>P(t)</td>
<td>ANN</td>
</tr>
<tr>
<td>5</td>
<td>WSVM</td>
<td>DP, CI, P(t-2), P(t-1), Tmin, Tmax, Tave</td>
<td>P(t)</td>
<td>Wavelet SVM</td>
</tr>
<tr>
<td>6</td>
<td>WANN</td>
<td>DP, CI, P(t-2), P(t-1), Tmin, Tmax, Tave</td>
<td>P(t)</td>
<td>Wavelet ANN</td>
</tr>
<tr>
<td>7</td>
<td>CCSM4</td>
<td>IC</td>
<td>P</td>
<td>Climate model</td>
</tr>
<tr>
<td>8</td>
<td>GFDL-A06</td>
<td>IC</td>
<td>P</td>
<td>Climate model</td>
</tr>
<tr>
<td>9</td>
<td>GFDL-AER04</td>
<td>IC</td>
<td>P</td>
<td>Climate model</td>
</tr>
<tr>
<td>10</td>
<td>GFDL-B01</td>
<td>IC</td>
<td>P</td>
<td>Climate model</td>
</tr>
<tr>
<td>11</td>
<td>CMC1</td>
<td>IC</td>
<td>P</td>
<td>Climate model</td>
</tr>
<tr>
<td>12</td>
<td>CMC2</td>
<td>IC</td>
<td>P</td>
<td>Climate model</td>
</tr>
<tr>
<td>13</td>
<td>NMME-EM</td>
<td>Model 7-12</td>
<td>P</td>
<td>Ensemble mean</td>
</tr>
<tr>
<td>14</td>
<td>Hybrid</td>
<td>Model 1-12</td>
<td>P</td>
<td>BMA</td>
</tr>
</tbody>
</table>
III. RESULT

The accuracy of precipitation forecasts of 313 rain gauge observations in China in statistical and climate models was quantified based on the Pearson’s correlation coefficient (PCC) and mean absolute error (MAE). The out-of-sample PCCs and MAEs were obtained from precipitation forecasts during 2006-2016 using a series of statistical, dynamic and hybrid models. It can be seen that the hybrid model achieves the best correlation (Table 2) and the least error (Table 3) among all the models. The underlying uncertainties in precipitation forecasts in statistical and dynamic models were reduced by BMA. In spite of the different mechanisms in various models, they can be combined to reduce the forecasting uncertainty and improve the accuracy.

Table 2. The out-of-sample PCCs of different models in 1 to 8 month lead precipitation forecasts.

<table>
<thead>
<tr>
<th>Model</th>
<th>Lead 1</th>
<th>Lead 2</th>
<th>Lead 3</th>
<th>Lead 4</th>
<th>Lead 5</th>
<th>Lead 6</th>
<th>Lead 7</th>
<th>Lead 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.70</td>
<td>0.69</td>
<td>0.69</td>
<td>0.68</td>
<td>0.68</td>
<td>0.69</td>
<td>0.70</td>
<td>0.69</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.54</td>
<td>0.54</td>
<td>0.48</td>
<td>0.27</td>
<td>-0.18</td>
<td>-0.44</td>
<td>-0.50</td>
<td>-0.54</td>
</tr>
<tr>
<td>SVM_C</td>
<td>0.78</td>
<td>0.79</td>
<td>0.79</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>ANN</td>
<td>0.67</td>
<td>0.42</td>
<td>0.67</td>
<td>0.68</td>
<td>0.70</td>
<td>0.68</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>WSVM</td>
<td>0.82</td>
<td>0.78</td>
<td>0.78</td>
<td>0.75</td>
<td>0.76</td>
<td>0.75</td>
<td>0.73</td>
<td>0.72</td>
</tr>
<tr>
<td>WANN</td>
<td>0.88</td>
<td>0.74</td>
<td>0.67</td>
<td>0.65</td>
<td>0.70</td>
<td>0.65</td>
<td>0.67</td>
<td>0.65</td>
</tr>
<tr>
<td>CCSM4</td>
<td>0.57</td>
<td>0.56</td>
<td>0.56</td>
<td>0.55</td>
<td>0.58</td>
<td>0.57</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td>GFDDL-A06</td>
<td>0.54</td>
<td>0.54</td>
<td>0.55</td>
<td>0.54</td>
<td>0.55</td>
<td>0.53</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>GFDDL-AER04</td>
<td>0.55</td>
<td>0.55</td>
<td>0.58</td>
<td>0.56</td>
<td>0.58</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>GFDDL-B01</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
<td>0.53</td>
<td>0.53</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td>CMC1</td>
<td>0.22</td>
<td>0.28</td>
<td>0.25</td>
<td>0.28</td>
<td>0.30</td>
<td>0.25</td>
<td>0.18</td>
<td>0.19</td>
</tr>
<tr>
<td>CMC2</td>
<td>0.22</td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
<td>0.26</td>
<td>0.30</td>
<td>0.22</td>
<td>0.20</td>
</tr>
<tr>
<td>NMME-EM</td>
<td>0.60</td>
<td>0.60</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>Hybrid</td>
<td>0.89</td>
<td>0.81</td>
<td>0.81</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Table 3. The out-of-sample MAEs (mm) of different models in 1 to 8 month lead precipitation forecasts.

<table>
<thead>
<tr>
<th>Model</th>
<th>Lead 1</th>
<th>Lead 2</th>
<th>Lead 3</th>
<th>Lead 4</th>
<th>Lead 5</th>
<th>Lead 6</th>
<th>Lead 7</th>
<th>Lead 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>66</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>66</td>
<td>66</td>
</tr>
<tr>
<td>LSTM</td>
<td>78</td>
<td>82</td>
<td>87</td>
<td>93</td>
<td>97</td>
<td>99</td>
<td>99</td>
<td>98</td>
</tr>
<tr>
<td>SVM_C</td>
<td>57</td>
<td>57</td>
<td>57</td>
<td>57</td>
<td>57</td>
<td>57</td>
<td>57</td>
<td>57</td>
</tr>
<tr>
<td>ANN</td>
<td>70</td>
<td>115</td>
<td>68</td>
<td>68</td>
<td>66</td>
<td>67</td>
<td>70</td>
<td>69</td>
</tr>
<tr>
<td>WSVM</td>
<td>52</td>
<td>58</td>
<td>58</td>
<td>60</td>
<td>60</td>
<td>61</td>
<td>63</td>
<td>65</td>
</tr>
<tr>
<td>WANN</td>
<td>45</td>
<td>64</td>
<td>72</td>
<td>75</td>
<td>66</td>
<td>72</td>
<td>69</td>
<td>70</td>
</tr>
<tr>
<td>CCSM4</td>
<td>88</td>
<td>85</td>
<td>86</td>
<td>86</td>
<td>84</td>
<td>84</td>
<td>83</td>
<td>87</td>
</tr>
<tr>
<td>GFDDL-A06</td>
<td>91</td>
<td>92</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>98</td>
<td>90</td>
<td>89</td>
</tr>
<tr>
<td>GFDDL-AER04</td>
<td>89</td>
<td>92</td>
<td>84</td>
<td>89</td>
<td>85</td>
<td>89</td>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td>GFDDL-B01</td>
<td>93</td>
<td>92</td>
<td>87</td>
<td>88</td>
<td>91</td>
<td>89</td>
<td>91</td>
<td>89</td>
</tr>
<tr>
<td>CMC1</td>
<td>123</td>
<td>111</td>
<td>115</td>
<td>113</td>
<td>114</td>
<td>112</td>
<td>122</td>
<td>117</td>
</tr>
<tr>
<td>CMC2</td>
<td>111</td>
<td>113</td>
<td>113</td>
<td>111</td>
<td>112</td>
<td>113</td>
<td>113</td>
<td>114</td>
</tr>
<tr>
<td>NMME-EM</td>
<td>81</td>
<td>80</td>
<td>78</td>
<td>79</td>
<td>78</td>
<td>79</td>
<td>79</td>
<td>79</td>
</tr>
<tr>
<td>Hybrid</td>
<td>42</td>
<td>53</td>
<td>54</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>57</td>
</tr>
</tbody>
</table>

ACKNOWLEDGEMENT: This work was supported by grants from Union Foundation of Ministry of Education of the People’s Republic of China (6141A02022318), the National Key Research and Development Program of China (2017YFB0503803), Creative Research Groups of Natural Science Foundation of Hubei Province of China (2016CFA003), the Fundamental Research Funds for the Central Universities (2042017GF0057), the China Meteorological Administration Drought Research Fund (IAM201704), and the China Postdoctoral Science Foundation (No.
REFERENCES
Accounting for Access Costs in Optimal Stratification for Validation of Soil Maps

Dick Brus *1, Lin Yang *2, A-Xing Zhu *1

1 Nanjing Normal University, China
2 Nanjing University, China

*Corresponding author: dick.brus@wur.nl; yanglin@lreis.ac.cn; axing@lreis.ac.cn

ABSTRACT

Soil maps are best validated by additional probability sampling and design-based estimation of map quality indices. In areas with marked differences in accessibility, cost efficiency can be increased by accounting for these differences in the selection of the validation locations. This can be achieved by stratified random sampling. The question then is how to construct the strata. Existing optimal stratification methods such as the cum√f and geometric stratification assume a constant cost among the sampling units, and therefore can be suboptimal when this assumption is violated. A simulated annealing algorithm is proposed for simultaneous optimization of the stratum breaks and the sample size under optimal allocation of the sample size, given a chosen maximum for the total access costs. In the case study optimal stratum breaks are computed for estimating the population mean of the squared prediction error (MSE). The kriging variance is used as a stratification variable. The optimal stratum breaks differed marginally from the cum√f breaks which can be explained by the weak correlation of the access costs and kriging variance. Larger differences in stratum breaks are expected when the correlation is stronger. The variance of the estimated mean of the stratification variable using the optimal stratification was about 5% smaller than with the cum√f stratification. This modest gain in precision is because with cum√f stratification the differences in mean costs among strata are small so that Neyman allocation is nearly optimal, and because the differences are accounted for in the allocation. The proposed algorithm can also be used in the absence of a stratification variable, accounting for differences in costs only. The proposed algorithm can be of use for estimating the population mean of any soil variable.

KEYWORDS: Digital Soil Mapping; Mean Squared Error; Kriging Variance; Simulated Annealing; Design-Based Estimation; Probability Sampling; Pedometrics
A Statistical Framework of Functional Data Analysis for Spatiotemporal Analysis and Modeling

Guofeng Cao*1

1 Department of Geosciences, Texas Tech University

*Corresponding author: guofeng.cao@ttu.edu

ABSTRACT
In this paper, we describe a statistical framework of functional data analysis for modeling and analysis of spatiotemporal footprints of geographic features, and demonstrate the advantages via real-world examples. Functional data analysis (FDA) represents a new perspective and approach in statistics for analyzing a population of functions (e.g., curves and surfaces). Instead of dealing with observations on single locations individually, FDA represents a set of measurements along a continuum (e.g., movement trajectory) as a single functional object and treats it as a random sample and basic analysis atoms. FDA extends the power of traditional spatial statistics by providing a formal way to characterize the dynamic and static features of processes on a spatial continuum while more realistically accounting for the spatiotemporal effects (e.g., dependency and geometry).

Within the context of FDA, we first represent spatiotemporal footprints of geographic objects (e.g., movement trajectory or boundary of geographic features) in forms of elastic functions, and adapted a recently developed FDA framework to characterize and model the spatiotemporal footprints. The adapted FDA framework, based on formal shape theory, provides an elastic metric of functional objects while accounting for geometry and spatiotemporal characteristics of geographic objects, and therefore allows the development of mathematically rigorous methods for statistical analysis of geographic objects. Based on this framework, we extend the commonly used basic statistical operators for scalar values to statistical analysis of spatiotemporal objects, including pairwise comparisons, averaging, covariance, principal component analysis, random sampling and p-value. These statistical operators allow the development of ‘Gaussian-type’ distribution for a collection of spatiotemporal objects, and provide an effective way to characterize and model the spaces of spatiotemporal objects as well as their interactions. To demonstrate the advantages of the framework, we apply the statistical operators for statistical analysis of bird migration paths and user-generated geographic information, and discuss the potential in statistical analysis and modeling of general geospatial data.

KEYWORDS: Spatial Uncertainty; Spatial Statistics; Functional Data Analysis
Cross-Comparison between GF-2 PMS2 and ZY-3 MUX Sensor Data

Wu Xiao-ping1,2, Xu Han-qiu1,2*

1 College of Environment and Resources, Key Laboratory of Spatial Data Mining & Information Sharing of Ministry of Education, Fuzhou 350116, China
2 Institute of Remote Sensing Information Engineering, Fujian Provincial Key Laboratory of Remote Sensing of Soil Erosion, Fuzhou University, Fuzhou 350116, China

*Corresponding author: hxu@fzu.edu.cn

ABSTRACT
Since the implementation of China’s National High Resolution Earth Observation Program, researches on the China-made new sensors are in progress. Nevertheless, no study has been published with respect to the cross-comparison between GF-2 PMS2 and ZY-3 MUX sensor data. Therefore, this study compares the GF-2 PMS2 and ZY-3 MUX sensor data based on three synchronous image pairs using two methods. A pixel-by-pixel comparison method was first used to investigate quantitative relationship between GF-2 PMS2 and ZY-3 MUX sensor data based on the whole test area. The other method is a comparison based on the region of interest (ROI) in common to avoid the problem due to the difference in spatial resolution between the GF-2 PMS2 (4m) and ZY-3 MUX (6m).

The result shows that the TOA reflectance of GF-2 PMS2 and ZY-3 MUX sensors has a high degree of agreement, with R2 values greater than 0.9 for all bands. However, the higher R2 values in blue and green bands indicate that the TOA reflectance between the two sensors in both bands has a better agreement than that of red and near-infrared bands. Scatter plots show that almost all data points lie under the one-to-one line in the spectral feature space with GF-2 data in x-axis and ZY-3 data in y-axis. This suggests that the GF-2 PMS2 sensor data generally have higher TOA reflectance than ZY-3 MUX. This study also finds that the signal difference between the two sensors is mainly affected by bare soil in the red band and by vegetation in the near infrared band. The more vigorous the vegetation grows, the greater the difference between the two sensors is.

The band-by-band comparison has yielded the conversion equations for each corresponding bands of the two sensors. The validation of the conversion equations shows that the conversion equations have high accuracy, which can significantly reduce the spectral difference between the two sensors. The differences in TOA reflectance between the two sensor data result probably from the differences in their spectral response function and spatial resolution.

KEYWORDS: GF-2 PMS2; ZY3-MUX; Sensors; Cross-Comparison
Spatial Dependence Modelling for Object-based Classifications of Remote-Sensing Images

Yunwei Tang¹, Jingxiong Zhang², Linhai Jing¹, Han Gao¹, Jiong You³, Ying Zhang⁴, Ru Xu⁵

¹ Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, China
² Wuhan University, China
³ Ministry of Agriculture, P. R. China / Chinese Academy of Agricultural Engineering, China
⁴ Jilin University, China
⁵ The Chinese University of Hong Kong, China

*Corresponding author: tangyunwei2008@163.com

ABSTRACT

Geographical object-based image analysis (GEOBIA) is widely used for the processing of fine spatial resolution images, with increased research on contextual modelling and classification related to GEOBIA. Specifically, a previously developed object-based image classification method, known as geostatistically weighted k-NN (gk-NN) method, has shown advantages in increasing classification accuracy. This method incorporates spatial weighting into the k-NN classifier through modelling spatial covariance of underlying area classes. However, change-of-support problem (COSP) due to different geometries of image objects is not considered therein. The COSP has been widely studied for the applications of health and remote sensing, and it was concluded that the scale effect between different data supports needs to be considered when transforming information from data across different scales. The existing contextual methods for GEOBIA, however, rarely consider the spatial dependence and related COSP in image objects specifically. In other words, there is a lack of methods for modelling spatial dependence in image objects and pixels with location- and distance-based spatial weighting method for image objects to individual pixels to support a location- and distance-based weighting in spatial structural modelling.

In this paper, we propose a method based on geostatistical de-regularization and regularization for quantifying spatial dependence in area-class occurrences and accounting for scale discrepancy in image objects and pixels. In this new modelling approach, an area-weighted (AW) distance measure is applied for modelling spatial covariance pertaining to sample image objects. The covariance model fitted with image objects sample data is de-regularized to a point-support one, so the spatial covariance over unsampled image objects can then be computed through regularization of a point-support model (RP). Unlike the previous modelling approach in the object-based gk-NN classification, whereby spatial dependence modelling is based on centroids of image objects, this method accounts for change of support and incorporates the geometry of image objects in modelling. The new modelling method was tested on three remote-sensing image subsets with different environments, using regular and irregular segmentation methods at hierarchical scales. It was found that regardless of spatial resolution of image data, the shapes and average sizes of image objects, COSP can be effectively dealt with using the proposed method, even for the pixel-based gk-NN classifications. Spatially dispersed area classes tend to be affected more by the proposed method (i.e., the RP method) than less dispersed classes. The proposed method leads to largely significant increases in classification accuracies, in comparison with the conventional gk-NN method without considering COSP. Further research can be directed towards applications of the RP modelling approach to other spatially weighted methods in GEOBIA. A self-adaptive method for optimal parameter estimation and development of methods for reduction of computational costs incurred in cross-support geostatistical modelling are also under investigation.

KEYWORDS: GEOBIA; Geostatistics; Spatial Structural Modelling; Change-of-Support; Image Classification
Reduction of Systematic Length Error Involved by Polygonal Approximation of Curves using Polynomial Interpolation Methods

Jean-François Girres*1

*1 Université Paul-Valéry Montpellier 3, IRD, UMR GRED 220, F34000, Montpellier, France

Corresponding author: jean-francois.girres@univ-montp3.fr

ABSTRACT

In a geographical database, the geometry of polylines is usually constructed using a succession of vertices connected by line segments. When the real world entities are curves (e.g. a sinuous road), their representation using polygonal lines induces a geometric simplification, called polygonal approximation of curves. The degree of curve approximation varies according to the level of details of the geographical database and its capture scale. As a consequence, the polygonal approximation of curves impacts the positional accuracy of the segments interpolated between two vertices, and also affects measurements computed from the polygonal line, by generating systematic underestimations of the curves lengths. So, in order to estimate the systematic error of lengths involved by the polygonal approximation of curves, this article proposes methods to re-construct realistic curve geometries from an original polygonal line using polynomial interpolation methods. To facilitate the re-construction of curve geometries using an original polygonal line, different methods of polynomial approximation are compared. Finally, the cubic Hermite splines polynomial interpolation method is selected, and more specifically the cardinal splines method, because of its low parameterization requirements. To preserve the realism of the re-constructed curves, a densification of the original polygonal line is performed previously, according to the level of detail of the geographical database, determined by data exploration, using the median inter-vertices distance. Finally, an experiment is performed on a sample of road network located in a mountainous area and presenting curve shapes. After a detection of sinuous roads, the experiments show that the use of cardinal splines with a tension parameter of 0.7 on a densified polyline is adapted for the re-construction of realistic curves. Results finally demonstrate that the proposed methods allow to reduce the systematic length error involved by the polygonal approximation of curves.

KEYWORDS: Polylines; Curves; Splines; Length; Error

INTRODUCTION

In a geographical database, the geometry of polylines is usually constructed using a succession of vertices connected each-other by line segments. When the real world entities are curves (e.g. a sinuous road), the representation using polygonal lines generates a geometric simplification, called polygonal approximation of curves. The polygonal approximation of curves is defined by Garnesson and Giraudon (1992) as the “operation of transforming a chain of related points in a succession of straight segments”. Geographical objects are by nature affected by polygonal approximation. Indeed, in most GIS software, polylines are directly captured by the operator using a succession of line segments, even if capture tools are implemented to build curves, such as Bezier curves. The degree of curve approximation varies according to the level of details of the geographical database and its capture scale. As a consequence, the polygonal approximation of curves impacts the positional accuracy of the segments interpolated between two vertices, and also affects measurements computed from the polyline, by generating systematic underestimations of the curves lengths (figure 1).
In this context, this article presents methods in order to estimate the systematic error of lengths involved by polygonal approximation when the real world entities are curves. To achieve this objective, this article proposes methods to re-construct realistic curve geometries from an original polygonal line using polynomial interpolation methods (figure 2), as proposed in Girres (2012).

In this section, two interpolation methods are presented: Bezier curves and Hermite cubic splines. Even if other interpolation methods are available, such as clothoids, we only focus here on these methods, which are commonly used and implemented.

I. POLYNOMIAL INTERPOLATION METHODS

To create curve geometries from an original polyline, several interpolation methods can be used. In this section, two interpolation methods are presented: Bezier curves and Hermite cubic splines. Even if other interpolation methods are available, such as clothoids, we only focus here on these methods, which are commonly used and implemented.

Bezier curves (from the french engineer Paul Bézier) are polynomial parametric curves defined using control points. For instance, from four control points P0, P1, P2, and P3, the interpolated curve goes from P0 to P3 using the direction defined by P1 and P2 (figure 3).
The main limitation of this method is that the Bezier curve does not go through the points P1 and P2, which is problematic in order to re-construct realistic curves impacted by polygonal approximation, and assess lengths underestimations.

Conversely, the Hermite cubic spline (from the french mathematician Charles Hermite) is a polynomial interpolation method, which goes through all control points, according to the tangents defined on each intermediate point, as illustrated on figure 4.

To define the behavior of the tangents, Kochanek and Bartels (1984) propose three levels of specialization of Hermite cubic splines: Catmull-Rom splines (when b=0), Cardinal splines (when c=0 and b=0), and Finite Difference (when all parameters are fixed at zero). The configuration of the tangents directly impacts the realism of the interpolated curve, so it seems necessary to control the impacts of each parameter on the interpolated curve. Nevertheless, no assumption related to the configuration of the parameters of bias and continuity can be formulated, so the use of Cardinal splines can be considered as sufficient in this work.

Cardinal splines only supposed to define the tension parameter, which has an impact on the realism of the interpolated curve. As illustrated on figure 5, when the tension parameter increases, the interpolated curve sticks with the original polygonal line. On the other hand, when the tension parameter decreases, it can generate unrealistic curves.
In this article, experiments are conducted in order to define the optimal value of the tension parameter in order to provide realistic curves.

II. PRESERVATION OF CURVES REALISM

To estimate the impact of polygonal approximation of curves on length measurements, the strategy proposed in this article supposes to re-construct realistic curves. Nevertheless, if the tension parameter has an impact, some other factors related to the configuration of the original polyline can also generate unrealistic curves. Indeed, in a polyline, the distance between successive vertices is not constant. This distance generally reduces in sinuous sections and increases in rectilinear sections. This heterogeneity can generate problems during the reconstruction of curves with polynomial interpolation methods, especially when a short segment is followed by a long segment (figure 6).

In this example, the tangent defined between the two segments generates an unrealistic curve. To overcome this problem, a segmentation using data analysis can be previously performed. As exposed in figure 7, the densification of the original polyline avoid the generation of unrealistic curves.
The determination of the appropriate distance used to densify the original polyline remains a difficult task because the distribution of inter-vertices distances can be very heterogeneous in a dataset. In this work, we propose to use the median inter-vertices distance, which has already been considered as a pertinent distance in a precedent work (Girres, 2015).

Thus, to estimate the systematic underestimation involved by the polygonal approximation of curves, we finally propose to (1) explore inter-vertices distances of a dataset, (2) densify polylines using the median inter-vertices distance, (3) reconstruct realistic curves using the Cardinal splines method. The next section presents an experiment of the proposed method.

III. EXPERIMENTS

The experiment illustrates the functioning of the method for the reconstruction of curves on a road network. The goal of the experiment is to determine the most appropriate value of the tension parameter for the reconstruction of realistic curves using the Cardinal splines method, and finally assess the impact of polygonal approximation of curves on length reduction.

The dataset used for the experiment is a sample of the BDCARTO (IGN, 2011) road network, located in a mountainous area of the french department of Pyrénées-Atlantiques (figure 8). The total length of the road network is about 191.1 km. A preliminarily detection of sinuous roads could have been provided, using the methods proposed by Plazanet (1996) and Mustière (2001), but in this experiment, the entire dataset is considered as sinuous, and is used for the reconstruction of curves.

![Figure 8. The experimented road network extracted from the BDCARTO database](image)

To avoid the reconstruction of unrealistic curves, a densification of the original dataset is performed using the median inter-vertices distance. Figure 9 shows the distribution of segment lengths of the road network dataset. The minimal and maximal segment lengths are 9 m. and 367 m. and the median distance used for the densification is of 51 m.

![Figure 9. Distribution of inter-vertices distances of the experimented road network](image)
The reconstruction of curves on the road network is performed using the Cardinal splines method, with a variation of the tension parameter from 0.1 to 0.9 (Table 1).

**Table 1. Sensibility of length underestimation according to the tension parameter of curves**

<table>
<thead>
<tr>
<th>Tension</th>
<th>Curve Length (km.)</th>
<th>Length difference (m.)</th>
<th>Length difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>196.69</td>
<td>5606</td>
<td>2.93</td>
</tr>
<tr>
<td>0.2</td>
<td>195.52</td>
<td>4432</td>
<td>2.32</td>
</tr>
<tr>
<td>0.3</td>
<td>194.54</td>
<td>3451</td>
<td>1.81</td>
</tr>
<tr>
<td>0.4</td>
<td>193.71</td>
<td>2625</td>
<td>1.37</td>
</tr>
<tr>
<td>0.5</td>
<td>193.01</td>
<td>1925</td>
<td>1.01</td>
</tr>
<tr>
<td>0.6</td>
<td>192.42</td>
<td>1335</td>
<td>0.69</td>
</tr>
<tr>
<td>0.7</td>
<td>191.93</td>
<td>843</td>
<td>0.44</td>
</tr>
<tr>
<td>0.8</td>
<td>191.53</td>
<td>442</td>
<td>0.23</td>
</tr>
<tr>
<td>0.9</td>
<td>191.23</td>
<td>142</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Results present the sensibility of the length according to the value of the tension parameter. Length reduction between polylines and curves goes from 3% (for \( t = 0.1 \)) to 0.1% (for \( t = 0.9 \)). To identify the most appropriate value of the tension parameter for the reconstruction of curves, a visual comparison between curves is performed. After comparison, curves reconstructed using a tension parameter \( t = 0.7 \) are considered as the most realistic curves for the configuration of the Cardinal splines method (Figure 10).

**Figure 10. Reconstruction of curves using varying tension parameters (left) and tension \( t = 0.7 \)**

This result indicates that for the road network experimented, using the proposed method for the reconstruction of curves, the impact of polygonal approximation of curves is about 843 m., which signifies an underestimation of 0.44% of the original road network length.

**CONCLUSION**

This article presents a simple method to re-construct realistic curves from an original polygonal line, in order to assess the impact of polygonal approximation of curves on length measurements. In this work, we propose to use the Hermite cubic spline polynomial interpolation method for the reconstruction of curves, based on the Cardinal spline method and a tension parameter \( t = 0.7 \). To avoid the reconstruction of unrealistic curves, a preliminary densification of the original polylines is provided using the median inter-vertices distance determined by data exploration. Results finally demonstrate that the proposed method allows to reduce the systematic length error involved by the polygonal approximation of curves. To fit more with particular use cases in the domain of
transportations, we also propose to combine this method with shape analysis methods, such as the detection of sinuous roads for instance.

**REFERENCES**


Uncertainty in the Relationship between Spatial Autocorrelation and the Modifiable Areal Unit Problem: A Simulation Study

Sang-II Lee *, Monghyeon Lee 2, Yongwan Chun 2, Daniel Griffith 2

1 Seoul National University, South Korea
2 The University of Texas at Dallas, United States

*Corresponding author: si.lee@snu.ac.kr

ABSTRACT

Spatial autocorrelation is recognized as a primary source of the modifiable areal unit problem (MAUP) (Amrhein, 1995), as has the presence of a systematic relationship between spatial resolution and spatial autocorrelation (Chou, 1991; Griffith et al., 2003). For example, with a spatial aggregation process, both the variance and Moran's I of an aggregated variable tend to decrease. Nevertheless, relationships between these two concepts continue to be investigated, and analysis outcomes generally are data dependent. The purpose of this paper is to summarize an investigation about how spatial autocorrelation behaves in conjunction with the MAUP to gain a more comprehensive understanding and generalizable outcomes. This paper considers two factors: a spatial autocorrelation level at the finest spatial resolution, and spatial tessellation types. More specifically, it explores the degree to which sample statistics, such as the mean, variance, correlation coefficient, regression coefficients, and Moran's I, are influenced by the effects of the MAUP.

The research supporting this paper utilizes experimental simulations. Methodologically, it employs a random spatial aggregation (RSA) procedure to construct spatial units from a fine spatial resolution by aggregating them into spatial units at a coarser spatial resolution. The experimental design begins with an initial dataset, which is a regular tessellation of 1,024 hexagons. These hexagons are randomly aggregated into 10 different coarser resolutions. For each aggregation, 1,000 different sets of zonations are generated. Spatially autocorrelated random variables are constructed with Moran eigenvectors, which are fundamental components of the eigenvector spatial filtering methodology (Griffith, 2003). To explore different levels of spatial autocorrelation, nine Moran eigenvectors representing spatial autocorrelation levels are selected from among 1,024 eigenvectors, which are extracted from a modified spatial weights matrix for the 1,024 hexagons. For each aggregated zonation in the simulation, three global statistics (i.e., the mean, variance, and Moran's I) are computed based on aggregated spatial units.

Behavior of the three global statistics is inspected to explore any systematic relationship between the initial level of spatial autocorrelation and the effects of the MAUP. Finally, systematic associations are examined in the context of regression that may help numerically formulate how much a sample statistic is influenced by the effects of the MAUP at a given initial level of spatial autocorrelation.

KEYWORDS: MAUP; Spatial Autocorrelation; Spatial Aggregation; Eigenvector
Small Area Estimation of Cancer Rates: A Case Study of Lung Cancer in Florida, 2000-2010

Lan Hu *1, Yongwan Chun1, Daniel Griffith1

1 The University of Texas at Dallas, United States

*Corresponding author: lan.hu@utdallas.edu

ABSTRACT

Small area estimation (SAE) furnishes a cost-effective method for obtaining, and is of importance due to a growing demand for, reliable small area data. For example, in epidemiology, while tremendous sources of health data exist for national and/or state levels, local health departments almost always lack access to small area data. One major problem of SAE is how to generate reliable estimates of interest for small areas with data only available at a coarser geographic resolution. A collection of methods has been proposed and applied to address this issue, including ones based upon generalized linear models, multilevel mixed models, and Bayesian map analyses. These methods tend to borrow information from both individual-level data and area-level covariates to yield relatively smoothed estimates for a finer geographic resolution with good accuracy.

Lung cancer, the leading cause of death for both men and women worldwide, has received substantial research attention. Geographers are interested in exploring spatial patterns of lung cancer rates across various geographic landscapes. The literature notes that positive spatial autocorrelation frequently is detected in a setting where a similar population group clusters (e.g., a Hispanic community). However, spatial autocorrelation enhances the complexity of a model, and a failure to accommodate it can generate biased parameter estimates and misleading statistical inferences. Thus, how to properly account for spatial autocorrelation in SAE is another challenge faced by spatial scientists. The literature presents spatial models to tackle it, including the Spatial Empirical Best Linear Unbiased Predictor; but this topic still remains underexplored.

The goal of this paper is to estimate lung cancer incidence rates at census tract and census block group resolutions in Florida by incorporating demographic risk factors (e.g., age, sex, and race) of individual cancer cases coupled with socioeconomic risk factors (e.g., income, education, and employment) at the county resolution. The estimation methods include creation of a synthetic population, a Poisson regression specification, and a Poisson mixed effects specification. Also, a Moran eigenvector spatial filtering method is utilized to accommodate a spatial autocorrelation component latent in the data. With the availability of geocoded lung cancer incidence points, the estimation results can be directly compared with the actual observed cancer incidence rates. In addition, this paper assesses whether the performance of these methods is consistent between the two different geographic resolutions. Finally, cancer rate maps are produced to evaluate if each estimation method adequately captures the underlying spatial pattern of cancer rates.

KEYWORDS: Small Area Estimation; Cancer Rates; Spatial Autocorrelation; Poisson Regression; Spatial Eigenvector Filtering
The Impact of Locational Uncertainty on Local Geographic Cluster Detection: An Application to Cancer Data from Florida, 2006-2010

Monghyeon Lee *,1, Daniel A. Griffith 1, Yongwan Chun 1

1 The University of Texas at Dallas, USA

*Corresponding author: monghyeon.lee@utdallas.edu

ABSTRACT

Local spatial cluster detection and its related analyses frequently are used in many disciplines, including epidemiology, criminology, demography, economics, policy science, and geography. Techniques that detect local spatial clusters identify spatial patterns that constitute deviation from a spatial random pattern. Finding and analyzing spatial clusters is an extremely useful undertaking in a spatial data analysis, because it can provide a better understanding of geographic phenomena and an underlying process governing those phenomena (Jacquez, 2008). However, spatial cluster analysis is not free from uncertainties that may contaminate any analysis step. Locational uncertainty is a distinctive source in spatial data analyses, one among a number of other uncertainty sources, such as sampling, measurement, specification, calculation, and stochastic behavior (Lee, Chun, & Griffith, 2017). Locational error occurs when a geotagged observation deviates from its true geographic position. This error can propagate to an analysis’ outcome, causing this error-propagated outcome to become less reliable. However, despite the wide range of thematic studies about spatial data uncertainty (e.g., Brown & Heuvelink, 2007; Lee, Chun, & Griffith, 2017), a propagation of spatial uncertainty in a spatial cluster analysis has not been extensively evaluated. Local indicators of spatial association (LISA; Anselin, 1995) and Getis-Ord *G* i statistics (Ord & Getis, 1995) are well-known indices measuring local spatial autocorrelation and use to identify spatial clusters. This paper investigates how location error affects the output of these local spatial autocorrelation measures through simulation experiments with late-stage colorectal cancer (LSCC) data from Florida in 2006 to 2010.

Simulation experiments were designed to analyse location error propagation. During this simulation exercise, we introduce different magnitudes of location error into the LSCC patient points, calculate local spatial cluster indices for both error-free and error-added LSCC patient points, and compare the results in terms of how the location error propagates; we varied the magnitude of location error, and the studied the point patterns, different MSAs in Florida, and local spatial autocorrelation indices. The simulation experiment results show that location uncertainty tends to propagate to local spatial cluster results in some MSAs with specific conditions. A considerable number of true geographic clusters are not identified as being significant in the simulation experiment. In contrast, some spatial units that are not significant clusters with the original error free data become significant after location errors are embedded. Location error propagation becomes more severe as the level of locational uncertainty increases (i.e., 50% of points containing a 100-meter distance displacement), and block group level results are impacted more than census tract level results. This last finding can be explained by the increasing chance that a corrupted observation can be displaced to a nearby areal unit as spatial resolution becomes finer. In addition, our findings indicate that locational uncertainty impacts LISA results more than *G* i results.

KEYWORDS: Location uncertainty; Cluster detection; Cancer
Spatial cluster analysis has a long history dating back to John Snow’s iconic removal of the Broad Street pump handle in London during an 1854 cholera epidemic. Noteworthy research during the 1990s (Anselin, 1995; Getis & Ord, 1992; Kulldorff, 1997; Ord & Getis, 1995; Sokal, Oden & Thompson, 1998) developed spatial cluster detection techniques, which since have been widely used in many disciplines, including epidemiology, criminology, demography, economics, policy science, and geography. These techniques can detect local spatial clusters, which constitute deviation from a random spatial pattern. Finding and analyzing spatial clusters is an extremely useful undertaking in a spatial data analysis, because it can provide a better understanding of geographic phenomena and an underlying process governing those phenomena (Jacquez, 2008).

A spatial cluster analysis, like other analyses, is not free from uncertainties that may contaminate any analysis step. Locational uncertainty is a distinctive source in spatial data analyses, among a number of other uncertainty sources, such as sampling, measurement, specification, and stochastic noise (Lee, Chun, & Griffith, 2017). Locational error occurs when a geotagged observation deviates from its true geographic coordinates. Locational error can propagate in any subsequent spatial data analysis. Although uncertainties may render only a slightly incorrect modelling result, they also can be completely fatal to a georeferenced data analysis, undermining its outcome (Fisher, 1999). However, despite the studies about spatial data uncertainty (e.g., Brown & Heuvelink, 2007; Lee, Chun, & Griffith, 2017), propagation of spatial uncertainty in a spatial cluster analysis has not been extensively analyzed. Local indicators of spatial association (LISA; Anselin, 1995) and Getis-Ord statistics (Ord & Getis, 1995) are well-known indices for measuring local spatial autocorrelation and for identifying spatial clusters. This paper investigates how location error affects output for these local spatial autocorrelation measures through simulation experiments based upon late-stage colorectal cancer (LSCC) data from Florida.

Individual LSCC diagnosed patient points during the 2006–2010 period were obtained from the Florida Cancer Data System (FCDS) for the following six Metropolitan Statistical Areas (MSAs): Pensacola, Tallahassee, Jacksonville, Orlando, Tampa, and Miami. Figure 1 portrays these MSAs with cancer point kernel smoothing maps. Points are densely located in high population areas, and, therefore, significantly high cancer rate clusters are identified in these areas.

The cancer data were cleaned by removing observations that: 1) have no spatial information [i.e., lack (x, y) coordinates], 2) have been georeferenced with partial addresses, such as a zipcode only, are located in census blocks having zero population in both the 2000 and 2010 decennial USCensus reports, or 4) are placed at the same location (i.e., points that have exactly the same geotags) because a single patient had multiple diagnoses. Then, the individual cancer patient points were aggregated for two administrative unit levels, census block groups and census tracks, to calculate spatial cluster indices for LSCC rates.

In the simulation experiments, artificial location errors with a rage of magnitudes were introduced into the LSCC patient point locations (i.e., displacement of the points from their true locations).

The error-embedded LSCC data were aggregated for the two census unit levels, and then the spatial cluster indices for LSCC rates were calculated with the error-embed data. Here these results are compared with results for the original error-free data. The magnitudes of location error were further combined with different proportions of observations subjected to artificial error embedding. Specifically, the proportions of error-embedded observations were set to 10%, 20%, 30%, 40%, or 50%. Distance displacements were set to five different levels (10m, 25m, 50m, 75m, or 100m). These factors result in a total of 600 combinations (i.e., six MSAs, five distance displacement levels, five proportions of patients experiencing error embedding, two levels of administrative units, and two spatial cluster indices). The direction of a displacement was randomly assigned, with the points constrained to stay within the boundaries of their respective MSAs.

LSCC rates are calculated by dividing the number of cancer points within a census unit by its population. A Bonferroni adjustment was utilized because this is a multiple comparisons situation; this adjustment is based upon the geographic effective sample size that accounts for latent spatial autocorrelation (see Chun & Griffith, 2013 and Castro & Singer, 2006). The simulation experiment...
employed 1,000 replicates for each combination (for example, calculate the Local Moran’s I of LSCC rates in the Miami MSA for census block groups for which a random sample of 10% of the LSCC patient points have a 10m distance displacement, repeating this scenario 1,000 times). In summarizing the experimental output, the number of times that each region is detected as a significant cluster is compared with its error-free counterpart results.

The simulation experiment results show that locational uncertainty propagates to local spatial cluster results. A considerable number of true geographic clusters are not identified as being significant in the simulation experiment. In contrast, some spatial units that are not significant clusters with the original error-free data become significant after location errors are embedded.

![Simulation Experiment Results](image1.png)

**Figure 1.** Cancer point kernel smoothing maps for 6 MSAs

For example, Figure 2 portrays the results for the Pensacola MSA. At the block group level, when the 10m error distance (the minimum distance error) is embedded (see Figure 2a), two positive and one negative cluster regions are identified (block groups with red and blue borders). However, one positive cluster region has been detected 861 times, and one non-cluster region 163 times, out of 1,000 replicates. With the maximum level of location error (see Figure 2b), locational uncertainties propagate to a greater extent. Three out of four original cluster regions detected with the error-free data appear as non-significant cluster regions in several error-added simulation replicates, and occasionally four non-original cluster regions are detected as existing cluster regions during the simulation. Furthermore, no Getis-Ord $G^*_i$ cluster region exists in the original data, but two block groups are detected with the error-embedded data (see Figure 2c and 2d). At the census tract level (see Figure 2e–2h), a similar pattern can be observed to that for the census block group level.

Other MSA results are similar (other MSA map results are omitted due to page length restrictions). Table 1 summarizes local spatial cluster detection analysis results that have been
Figure 2. Local Moran’s I and Getis-Ord $G^*_c$ distance error-embedded simulation results for the Pensacola MSA
affected by the distance errors for both the block group and the census tract levels for the six MSAs. This table reports only the extreme levels of location errors, with results for intermediate error levels being between these extreme error level results. This table is formatted as a change detection table. If a unit is continuously identified as having its original status during the 1,000 simulation experiments, then error does not propagate (i.e., cluster → cluster, or not cluster → not cluster; the diagonal elements in the table). In contrast, if a unit is identified at least once as changing its original status during the simulation, then error propagates (i.e., cluster → not cluster, or not cluster → cluster; the off-diagonal elements in the table). According to these results, although errors propagate differently with different conditions, local cluster detection analysis results can be distorted by locational uncertainties.

Overall, location error propagation becomes more severe as the level of locational uncertainty increases (i.e., 50% of points with a 100m distance displacement), and the block group level results are more severely influenced than the census tract level results. This finding can be explained as follows: an error embedded observation can have a higher chance of being displaced to a nearby areal unit at a finer spatial resolution level. In addition, our findings indicate that locational uncertainty impacts LISA results more than $G^*_i$ results.

<table>
<thead>
<tr>
<th></th>
<th>Original data</th>
<th>Distance error added data (1,000)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Minimum error</td>
<td>Maximum error</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cluster (10% 10m)</td>
<td>Not cluster</td>
<td>Cluster (50% 100m)</td>
</tr>
<tr>
<td>Block group</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peninsular MSA</td>
<td>$I_i$</td>
<td>Cluster 4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$G^*_i$</td>
<td>Not cluster 263</td>
<td>4</td>
<td>259</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cluster 0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not cluster 267</td>
<td>2</td>
<td>265</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Census tract</td>
<td>$I_i$</td>
<td>Cluster 1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$G^*_i$</td>
<td>Not cluster 95</td>
<td>2</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cluster 1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not cluster 95</td>
<td>0</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tallahassee MSA</td>
<td>$I_i$</td>
<td>Cluster 1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$G^*_i$</td>
<td>Not cluster 232</td>
<td>2</td>
<td>230</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cluster 0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not cluster 233</td>
<td>0</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Census tract</td>
<td>$I_i$</td>
<td>Cluster 1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$G^*_i$</td>
<td>Not cluster 83</td>
<td>0</td>
<td>83</td>
</tr>
</tbody>
</table>

22 / 140
<table>
<thead>
<tr>
<th></th>
<th>$G_i$</th>
<th>Cluster</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>Not cluster</th>
<th>84</th>
<th>0</th>
<th>84</th>
<th>1</th>
<th>83</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block group</td>
<td>$I_i$</td>
<td>Cluster</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>Not cluster</td>
<td>691</td>
<td>5</td>
<td>686</td>
<td>37</td>
<td>654</td>
</tr>
<tr>
<td>Census tract</td>
<td>$G_i$</td>
<td>Cluster</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>Not cluster</td>
<td>696</td>
<td>4</td>
<td>692</td>
<td>9</td>
<td>687</td>
</tr>
<tr>
<td>Block group</td>
<td>$I_i$</td>
<td>Cluster</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>Not cluster</td>
<td>248</td>
<td>3</td>
<td>245</td>
<td>8</td>
<td>240</td>
</tr>
<tr>
<td>Census tract</td>
<td>$G_i$</td>
<td>Cluster</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>Not cluster</td>
<td>251</td>
<td>2</td>
<td>249</td>
<td>3</td>
<td>248</td>
</tr>
<tr>
<td>Block group</td>
<td>$I_i$</td>
<td>Cluster</td>
<td>14</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>12</td>
<td>Not cluster</td>
<td>819</td>
<td>9</td>
<td>810</td>
<td>30</td>
<td>789</td>
</tr>
<tr>
<td>Census tract</td>
<td>$G_i$</td>
<td>Cluster</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>Not cluster</td>
<td>825</td>
<td>7</td>
<td>818</td>
<td>21</td>
<td>804</td>
</tr>
<tr>
<td>Block group</td>
<td>$I_i$</td>
<td>Cluster</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>Not cluster</td>
<td>385</td>
<td>9</td>
<td>376</td>
<td>13</td>
<td>372</td>
</tr>
<tr>
<td>Census tract</td>
<td>$G_i$</td>
<td>Cluster</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>Not cluster</td>
<td>384</td>
<td>3</td>
<td>381</td>
<td>4</td>
<td>380</td>
</tr>
<tr>
<td>Block group</td>
<td>$I_i$</td>
<td>Cluster</td>
<td>25</td>
<td>15</td>
<td>10</td>
<td>8</td>
<td>17</td>
<td>Not cluster</td>
<td>1981</td>
<td>10</td>
<td>1971</td>
<td>78</td>
<td>1903</td>
</tr>
<tr>
<td>Block group</td>
<td>$I_i$</td>
<td>Cluster</td>
<td>58</td>
<td>18</td>
<td>40</td>
<td>2</td>
<td>56</td>
<td>Not cluster</td>
<td>3319</td>
<td>29</td>
<td>3290</td>
<td>103</td>
<td>3216</td>
</tr>
<tr>
<td>Census tract</td>
<td>$G_i$</td>
<td>Cluster</td>
<td>34</td>
<td>16</td>
<td>18</td>
<td>4</td>
<td>30</td>
<td>Not cluster</td>
<td>3343</td>
<td>20</td>
<td>3323</td>
<td>73</td>
<td>3270</td>
</tr>
<tr>
<td>Block group</td>
<td>$I_i$</td>
<td>Cluster</td>
<td>19</td>
<td>5</td>
<td>14</td>
<td>1</td>
<td>18</td>
<td>Not cluster</td>
<td>1180</td>
<td>3</td>
<td>1177</td>
<td>14</td>
<td>1166</td>
</tr>
<tr>
<td>Census tract</td>
<td>$G_i$</td>
<td>Cluster</td>
<td>17</td>
<td>11</td>
<td>6</td>
<td>0</td>
<td>17</td>
<td>Not cluster</td>
<td>1182</td>
<td>7</td>
<td>1175</td>
<td>9</td>
<td>1173</td>
</tr>
</tbody>
</table>
REFERENCES
Object-oriented Classification Based on Deep features for High Spatial Resolution Remote Sensing Images

Han Gao *, Linhai Jing *, Yunwei Tang *, Hui Li *, Haifeng Ding *

1 Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, China

*Corresponding author: gaohan@radi.ac.cn; jinglh@radi.ac.cn; tangyw@radi.ac.cn; lihui@radi.ac.cn; dinghf@radi.ac.cn

ABSTRACT

In recent years, very high spatial resolution (VHR) earth observation images obtained from satellite and airborne sensors have become increasingly available and have provided more detailed spatial structures and textural features of ground objects. Geographic object-based image analysis (GEOBIA) is a widely used and particularly effective method for the analysis of VHR images that overcomes the limitations of pixel-based image analysis. However, this method requires the inefficient process of manually designed features, and these perceptual level features have limited ability in obtaining discriminative representation of the ground objects in VHR images. Deep learning offers an efficient strategy to fill the gap between complex image patterns and their semantic labels via automatic feature learning. In this paper, we propose an end-to-end framework for the VHR remote sensing images using multiscale convolutional neural networks (MCNN). This framework is built on a combination of deep feature learning strategy and an object-oriented classification without intensive human involvement. Specifically, image patches of certain scales are obtained by using simple linear iterative clustering (SLIC) as input to the MCNN and the high-level feature representations extracted through the MCNN have been systematically investigated over five different layer configurations. Furthermore, the boundary information derived from multiscale segmentation compensate the limitations of the MCNN due to the adoption of convolutional filters such as uncertainty in object boundary partition at pixel level and loss of useful fine spatial resolution detail. The effectiveness of the ensemble framework was tested in both urban and rural areas using aerial photography and the proposed method achieved promising performance. This research paves the way to effectively address the complicated problem of geospatial big data.

KEYWORDS: Deep learning; High-resolution image; Object-oriented classification
Comparing the Impact of the Mapping Error on Aggregation Methods

Peijun Sun ¹², Russell G Congalton ³, Yaozhong Pan ¹²

¹ State Key Laboratory of Remote Sensing Science, Jointly Sponsored by Beijing Normal University and Institute of Remote Sensing and Digital Earth of Chinese Academy of Sciences, Beijing 100875, China.
² Institute of Remote Sensing Science and Engineering, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China.
³ Department of Natural Resources & the Environment, University of New Hampshire, Durham, New Hampshire 03824, U.S.

*Corresponding author: pyz@bnu.edu.cn

ABSTRACT

Rescaling techniques can be used to generate land cover maps at different resolutions to serve Earth science modelling needs. Categorical aggregation methods have been validated as efficient tools for producing such upscaled maps. Various aggregation methods perform differently in generating these upscaled maps. However, the comparisons of the influence of mapping error on the accuracies of different aggregation methods have not been investigated. The purpose of this paper is to compare the performance of categorical aggregation methods on reducing the negative influence of the mapping error. To address the issue, the Fusing class Membership probability and Confidence level probability (FMC) and the Majority Rule Based (MRB) methods, were used. The results illustrated that FMC can perform better compared to MRB. Also, the results suggest that three factors: the mapping error level of the base map, the tolerable error of the upscaled maps for the project, and the budget of the project, should be carefully considered before deciding on the selection of the spatial resolution.

KEYWORDS: Categorical aggregation; upscaling accuracy; class membership probability; confidence level probability

I. INTRODUCTION

Land cover maps generated by remote sensing technologies have been examined as efficient tools serving Earth science modellings (Congalton et al., 2014). In the past decades, many land cover products were produced based on the continuous development of remote sensing technology. As Grekoskis et al. (2015) reported, 23 global and 41 regional land cover mapping products have been produced (e.g., Global Land Cover Characteristics (GLCC) 2.0 (Loveland et al., 2000), Global Land Cover 250m China (GLC250m_CN) (Wang et al., 2015), Finer Resolution Observation and Monitoring-Global Land cover (FROM-GLC) (Gong et al., 2013; Yu et al., 2014)). The resolutions of these products/maps mainly cover 1km, 500m, 300m, 250m, and 30m pixels.

Unfortunately, these resolutions cannot meet the requirements of all the various Earth science modellings that need different land cover maps at different spatial resolutions in order to reduce the uncertainty/error of these models (Sun et al., 2017). Thus, rescaling techniques were developed, which can be divided into two approaches (Atkinson, 2013): (1) downscaling techniques that increase the spatial resolution of the land cover maps, and (2) upscaling techniques that decrease the spatial resolution of the land cover maps. Of the particular interest in this paper is the upscaling techniques that can be implemented either numerically or categorically to produce upscaled maps (Sun et al., 2018). The numerical approach first predicts the coarser resolution imagery as input to produce upscaled maps based on the function between the finer pixel values and their corresponding
coarser pixel values (e.g., Gardner et al., 2008). Per Raj et al. (2013), the categorical way predicts the class type for the coarser pixel based on its corresponding class types of the finer pixels using different aggregation logic. For example, the Majority Rule Based aggregation (MRB) determines the class types for the output coarse-resolution pixels in the upscaled map by assigning the most frequently occurring class from the finer-resolution map contained within each associated coarse pixel (Benson and MacKenzie, 1995; He et al., 2002; Moody and Woodcock, 1995, 1994, Turner et al., 1989a, 1989b).

Recently, Sun et al. (2018) proposed a new categorical method, the Fusing class Membership probability and Confidence level probability (FMC), to produce upscaled maps. This approach works by fusing two different sources of information and can produce more accurate upscaled maps compared to MRB. In addition, Sun et al. (2017) reported that MRB is negatively and greatly impacted by the mapping error of the base map. Whether or not FMC is greatly influenced by the mapping error of the base map has not been investigated. Therefore, the purpose of this paper is to explore if FMC can mitigate the negative influence of the mapping error of the base map on the accuracy of the upscaled map. To assess the accuracies of the upscaled maps, the similarity matrix approach (Sun and Congalton, 2018) was used. The USDA Cropland Data Layer (CDL) data were used for two different study areas with different heterogeneity to conduct the experiments.

II. DATA AND STUDY SITES

The USA Cropland Data Layer (CDL) data produced by the National Agricultural Statistics Service (NASS) of the US Department of Agriculture (USDA) were selected for use in this study. These data can be accessed from NASS’ online geospatial application-CropScape (https://nassgeodata.gmu.edu/CropScape/). The CDL data have been validated as efficient inputs for various research due to their higher accuracies (Boryan et al., 2011). Therefore, this paper used the CDL data at 30m for 2016 to produce the agriculture maps as base maps with an assumed/designated error level of 0%. In addition, the confidence layer data representing how well the output pixels were determined in the CDL data (Boryan et al., 2011) were used to assist in the error simulation for producing base maps with different mapping error levels. The projection for all the data was Albers Equal Area.

Two Agricultural Statistic Districts (ASDs), ASD1870 and ASD4530, with different landscape patterns were selected (Figure 1). These districts are groupings of counties in each State and are defined by geography, climate and cropping practices (USDA, 2017). Both sites have high crop diversity (e.g., corn, winter wheat, soybeans, sorghum, alfalfa, etc.). Patch-Per-Unit (PPU), a measure of landscape heterogeneity (Frohn, 1997), was used to represent the heterogeneity of the two study sites. The first site, ASD1870, is located in southwestern Indiana, centered around 87° 15’ 16.14”W, 38° 29’ 41.33”N. The study site comprises approximately 1.30×10^4 km^2. The heterogeneity of this study site (PPU of 10.10) is higher than the other site (Table 1). The second study site, ASD4811, is located in the northeastern South Carolina, centered around 79° 33’ 55.95”W, 34° 2’ 48.63”N. The total area of the ASD4530 is 1.68×10^4 km^2. The ASD4530 has a PPU of 4.76, which is more homogenous than the first site.
Table 1. Landscape patterns of two study sites. TA means the total area of the study sites. NP means the number of patches (fields) in the study sites. Patch-Per-Unit (PPU) is a measure of heterogeneity. The Fragstats version 4.2, a spatial analysis software package for calculating landscape metrics, was used to obtain all landscape metrics in this paper. Note that all landscape metrics are calculated based on the agricultural thematic maps generated from CDL data.

<table>
<thead>
<tr>
<th>ASD</th>
<th>TA (ha)</th>
<th>NP</th>
<th>Mean, Median, and standard deviation of parcel size (ha)</th>
<th>AI</th>
<th>Dominance</th>
<th>PPU</th>
<th>SqP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1870</td>
<td>1301162.94</td>
<td>131357</td>
<td>9.91, 0.18, 2030.96</td>
<td>91.15</td>
<td>0.44</td>
<td>10.10</td>
<td>0.99</td>
</tr>
<tr>
<td>4530</td>
<td>1681514.28</td>
<td>79995</td>
<td>21.02, 0.09, 5623.75</td>
<td>97.46</td>
<td>0.86</td>
<td>4.76</td>
<td>0.99</td>
</tr>
</tbody>
</table>

III. METHODOLOGIES

Generation of agricultural maps

The overall principle for generating the agricultural maps in this study was that any agriculture crop type covering larger than 0.02% area of the whole study site was maintained, while all other map classes were combined into a non-crop class. The CDL data were then used to extract the crop maps. These maps were then used as the base maps with an error level of 0% based on the assumption that the CDL data were 100% accurate. This assumption is necessary in this analysis in order to quantitatively produce the necessary base maps with specific mapping errors.

Mapping error simulation

The probability-based Monte Carlo simulation (MC) (Sun et al., 2017), a numerical experimentation and statistical sampling technique (Cruse, 1997), was utilized to produce mapping error from the original base maps at eight levels (5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%). The error level represents the percentage of misclassified crop pixels compared to the initial CDL data. The MC simulation was conducted based on the assumption that the pixels at the boundary of agricultural fields were misclassified with a higher probability because error in agriculture areas tends to occur at the boundary of agricultural fields (Congalton, 1988). Thus, the pixel with the lower confidence level probability will be misclassified by the MC algorithm.
Per Sun et al. (2017), the confidence level probabilities, Ci, for the boundary pixels of each agricultural field were first obtained. The required mapping error level, \( \xi \), and the total number of the pixels in the study site, \( N \), were then obtained. Thus, the required number of misclassified pixels, \( N_\xi \), can be computed by \( N_\xi = N \times \xi \). After obtaining \( N_\xi \), a pair of boundary pixels, P1 and P2, with their confidence level probability, C1 and C2, respectively, were randomly chosen. Two random number, n1 and n2, from a uniform distribution ranging [0,1] were also generated.

The condition for swapping the class types of two pixels, P1 and P2, is that n1>C1, and n2>C2. If the condition cannot be met, P1 and P2 would be marked as unswapped pixels. Then, another two pixels, \( P_i \) and \( P_j \) would be randomly selected for processing the misclassification. After repeating 2N times (Sun et al., 2017), the simulation tests to determine if the required number of swapped pixels, \( N_{sw} \), was obtained. If \( N_{sw} < N_\xi \), the simulation would continue until the required error level was obtained. Note that if the number of the boundary pixels was less than \( N_\xi \), the non-boundary pixels in the agricultural fields will also be involved in the error simulation.

**Upscaling methods**

1. The Majority Rule Based aggregation (MRB) determines the land cover class for the coarse pixels in the upscaled map by selecting the most frequently occurring class from the finer resolution map contained within each coarse pixel (He et al., 2002; Saura, 2004). When there is more than one major class, the dominant class is selected at random (Raj et al., 2013).

2. The Fusing class Membership probability and Confidence level probability (FMC) determines the class type for the coarse pixels by assigning the class type with the highest posterior probability. The posterior probability of each class type for one coarse pixel is calculated by fusing the spatial structure information (the class membership probability, CMP) and the uncertainty information of base map (confidence level probability, CLP). The Tau model (Boucher et al., 2008) was used to predict the fused probability of the \( K_\text{th} \) land cover class at the coarse pixel \( V_m \), denoted as \( P_k^{FMC}(V_m) \), by Equation (1).

\[
P_k^{FMC}(V_m) = \left[ 1 + \left( \frac{1-P_k^{CMP}(V_m)}{P_k^{CLP}(V_m)} \right)^{\tau_{CMP}} \frac{1}{\left( \frac{1-P_k^{CLP}(V_m)}{P_k^{CLP}(V_m)} \right)^{\tau_{CLP}}} \right]^{-1}
\]

where \( P_k^{CMP} \) and \( P_k^{CLP} \) are the CMP and CLP for \( K_\text{th} \) class types, \( \tau_{CMP} \in (0,1) \) and \( \tau_{CLP} \in (0,1) \) are two exponent values with a nonnegative constraint to ensure \( \tau_{CMP} + \tau_{CLP} = 1 \).

Per Sun et al. (2018), FMC uses seven steps to produce upscaled maps. (1) a non-overlapping window, \( W_{CLP} \), was constructed to calculate the \( P_k^{CLP} \), which has same size with the coarser pixel. Within each \( W_{CLP} \), \( N_{CLP}^W \times N_{CLP}^W \) finer pixels were covered. (2) Compute the number of pixels for each class \( K \), denoted as \( N_k^{CMP} \) within \( W_{CLP} \), a predefined window to calculate the \( P_k^{CMP} \). The \( W_{CMP} \) covers \( N_{CMP}^W \) coarse pixels. In this paper, the \( W_{CLP} \) covers three coarser pixels, which is same with Sun et al. (2018). Then, \( P_k^{CMP} = \frac{N_k^{CMP}}{N_{CLP}^W \times N_{CLP}^W} \times 100\% \). (3) The pixel, \( v_i \), in the base map and its CLP, \( P_k(v_i) \), within \( W_{CLP} \) was obtained. (4) the number of each class k from the base map within \( W_{CLP} \) was calculated and denoted as \( N_{CLP}^k \). Then, the CLP for each class k within \( W_{CLP} \) were calculated by \( P_k^{CLP} = \sum_{i=0}^{N_{CLP}^k} P_k^i(v_i)/N_{CLP}^k \). To ensure the sum of CLP for different class types for one coarse pixel is 1, the CLP for each class k was modified by \( P_k^{CLP}(V_m) = R_k \times \left( \sum_{i=0}^{N_{CLP}^k} P_k^i(v_i)/N_{CLP}^k \right) = P_k^{CLP}(V_m)/\sum_{k=0}^{N_k} P_k^{CLP}(V_m) \). (5) Based on Tau model (Equation 1), the probability of each class type occurring at the upscaled maps, denoted as \( P_k^{FMC}(V_m) \) was computed. (6) the class type \( k^* \) with the highest \( P_k^{FMC}(V_m) \) will be assigned as the class type for the upscaled map. (7) 101 pairs of exponent values ranging [0,1] at 0.01 changing steps were utilized to select the optimal upscaled map with the highest accuracy (i.e., overall similarity in this paper). Then the outputs include the upscaled map and its corresponding posterior probability layer.

**Accuracy assessment**

The similarity matrix approach proposed by Sun and Congalton (2018) has been validated as an
efficient way to evaluate the rescaled maps. Three measurements, overall similarity (OS), omission error (OE) and commission error (CE), are calculated in the similarity matrix (Table 2). The total area of class i in the base map that is identified as class j in the upscaled maps is denoted as \( A_{ij} \). Then, \( A_{ik} = \sum_{j=1}^{n} A_{ij} \) and \( A_{kj} = \sum_{i=1}^{n} A_{ij} \). After obtaining \( A_{ik} \) and \( A_{kj} \), the OE, CE and OS are calculated by \( (A_{ik} - A_{ii})/A_{ik} + (A_{kj} - A_{ij})/A_{kj} \) and \( \sum_{i=0}^{n} A_{ii} / \sum_{i=0}^{n} A_{ik} \), respectively.

IV. RESULTS

Generation of error maps

There are a total of nine base maps at 30m spatial resolution with error levels ranging from 0% to 40% for each study site. Table 3 shows the proportion of each class for each base map with error level of 0%. The results show that different class types have different proportions. In addition, for each study site, eight base maps with error levels of 5%, 10%, 15%, 20%, 25%, 30%, 35% and 40%, were, respectively, produced using MC simulation method. Table 4 shows that the non-boundary pixels also involved in some of the error simulations. For example, for ASD 4530, the non-boundary pixels contributed about 11.74% errors to the final base map with error level of 40%.

Table 3. Proportion of each class for the base maps with error level of 0% at each study site.

<table>
<thead>
<tr>
<th>Class type</th>
<th>Study area</th>
<th>Non-crop</th>
<th>corn</th>
<th>cotton</th>
<th>sorghum</th>
<th>soybeans</th>
<th>Winter wheat</th>
<th>Alfalfa</th>
<th>peaches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1870</td>
<td>57.13</td>
<td>21.43</td>
<td>0</td>
<td>0.12</td>
<td>20.98</td>
<td>0.06</td>
<td>0.11</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4530</td>
<td>94.66</td>
<td>3.72</td>
<td>1.50</td>
<td>0.07</td>
<td>0</td>
<td>0.03</td>
<td>0</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 4. Percentage of the error contributed by non-boundary pixels at each error level for each study site.

<table>
<thead>
<tr>
<th>Error level</th>
<th>Study site</th>
<th>0%</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>35%</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1870</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.22</td>
<td>9.38</td>
<td>16.25</td>
<td></td>
</tr>
<tr>
<td>4530</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.74</td>
<td>11.17</td>
<td></td>
</tr>
</tbody>
</table>

Comparisons of accuracies and mapping error influence for FMC and MRB

There are a total of 90 upscaled maps (9 base maps x 10 resolutions) for each study site. For each error level, the upscaled maps at ten different resolutions (60m, 90m, 120m, 240m, 360m, 480m, 600m, 720m, 840m, and 960m) were produced. Figure 2 shows the OS for all upscaled maps derived by FMC and MRB, respectively, for two study sites. Figures 4 and 5 show that the OE and CE of corn in the upscaled maps derived by FMC and MRB, respectively, for two study sites. The results show that larger mapping error resulted in lower OS, higher OE and higher CE. For example,
for FMC, the OS of the upscaled maps at 90m in ASD1870 decreased from 91.48% to 85.27%, and the OE and CE for corn in these maps increased from 11.29% to 27.84% and from 7.42% to 21.35%, respectively. Furthermore, for both methods the results show that higher mapping error level resulted in higher decrease in accuracy for ASD1870 compared to ASD4530. For example, when mapping error increased from 0% to 40%, for the upscaled maps at 60m derived by FMC, the OS for the heterogeneous area (i.e., ASD1870) and the homogeneous area (i.e., ASD4530) were reduced from 95.21% to 86.84% and from 98.94% to 97.74%, respectively.

To achieve one of the primary objective of this paper, the accuracies of the upscaled maps derived by FMC and MRB were compared. Figures 2, 3, and 4 show that FMC can obtain higher accuracy under each mapping error level (i.e., higher OS, generally lower OE and CE) compared to MRB. For example, when using the base map with mapping error level of 5% at ASD1870, the OS of the upscaled maps at 240m derived by FMC are higher by about 4.60% compared to the MRB. In addition, FMC obtained a lower decrease in OS when the mapping error level increased from 0% to 40% for each resolution compared to MRB. Furthermore, comparable results between FMC and MRB show that FMC obtained higher accuracy in the heterogeneous area (e.g., ASD1870) when mapping error increased compared to MRB. For example, the OS of the upscaled maps at 240m in ASD1870 was reduced from 88.30% to 82.77% for FMC, while from 83.71% to 78.86% for MRB.

Figure 2. Overall similarity (OS) of the upscaled maps for two study sites. (a) is the OS for ASD, and (b) is the OS for ASD.

Figure 3. An example of omission error (OE) for corn in the upscaled maps derived by FMC and MRB, respectively, for ASD 1870 and ASD 4530. (a) and (b) are the results for the FMC. (c) and (d) are the results for the MRB.
V. DISCUSSION

Two issues should be noted before using any of the base maps. First, the base maps with error level of 0% were produced based on the assumption that the CDL data were totally correct with an accuracy of 100%. This assumption is the basis of the error simulation to ensure that the base maps with specific mapping error can be produced for further explorations. Without making this assumption, it will be very difficult to quantitatively analyze how the mapping error impacts the accuracy of the upscaled maps. Second, in addition to the boundary pixels, non-boundary pixels were also involved in the error simulation. This situation results in the ‘salt and pepper’ issue in the classification maps when using classification techniques (e.g., Shao and Lunetta, 2012; Sun et al., 2014).

The accuracy analysis for both methods demonstrates that the mapping error greatly impacts the accuracy of the upscaled maps, which is consistent with Sun et al. (2017). This result further confirms that mapping error should be used cautiously for any further explorations using the upscaled maps. It seems that increasing the accuracy of the base map is a potential way to increase the accuracy of the upscaled maps. However, the cost for increasing the accuracy of the base map should be carefully considered. Moreover, if user’s models/projects can tolerate the error resulted from mapping error of the base map, it is not wise to spend the time or effort to increase the accuracy of the base map.

The results of this paper confirm that the FMC can obtain a higher accuracy compared to the MRB, which is consistent with Sun et al. (2018). Also, the FMC can mitigate the negative influence of the mapping error and the influence of heterogeneity on the upscaling accuracy due to its higher accuracy under each mapping error level and in the heterogeneous area. These results highlight that fusing multi-sources of information may be beneficial when producing upscaling maps. Any future work for developing upscaling techniques should consider how to efficiently use more sources of information to accurately predict the class type for the upscaled map.

VI. CONCLUSION

The goal of this study was to explore if FMC can mitigate the error influence on the upscaling accuracy compared to the conventional method (i.e., MRB). The CDL data were selected for two Agricultural Statistic Districts (ASDs) to implement the error simulation using a probability-based Monte Carlo simulation, and conduct FMC and MRB. For each study site, nine base maps with different error levels (0%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, and 40%) were applied to produce the upscaled maps at ten different resolutions (60m, 90m, 120m, 240m, 360m, 480m, 600m, 720m, 900m, and 1200m).
840m, and 960m). Several conclusions can be summarized: (1) the results demonstrate that FMC can mitigate the negative influence of the mapping error on the accuracy of the upscaled maps by obtaining higher OS, lower OE and CE under each mapping error level, (2) the results show that the heterogeneity of the study site has lower negative influence on the FMC compared to the MRB since the FMC obtained higher accuracy in the heterogeneous area than the MRB, and (3) the results have important implications that the map users should carefully consider the mapping error, the tolerance of the model, and the cost for increasing the accuracy of the base map and the upscaled maps before selecting the suitable spatial resolution. Overall, extending the achievements of the FMC does not only strengthen our confidence to recommend this new upscaling technique, but also gives important implications for the Earth science community to use upscaled maps.

ACKNOWLEDGEMENT: Partial funding was provided by the Major Project of High-Resolution Earth Observation, China. Partial funding was provided by the New Hampshire Agricultural Experiment Station. This work was supported by the USDA National Institute of Food and Agriculture McIntire Stennis Project #NH00077-M (Accession #1002519). We thank National Agricultural Statistics Service (NASS) Cropland Data for providing CDL data. The authors also thank China Scholarship Council (CSC) for providing foundation for the first author as a visiting researcher in the Basic and Applied Spatial Analysis Lab (BASAL), Department of Natural Resources and the Environment, University of New Hampshire, under the direction of Dr. Russell G. Congalton.

REFERENCES


An Uncertainty-based Approach to Quantify the Spatial Representativeness of Local Health Datasets

Rachel Whitsed 1, Ana Horta 1, Amilcar Soares 2, Herbert Jelinek 3

1 Institute of Land, Water and Society, Charles Sturt University, Australia
2 CERENA, Instituto Superior Técnico, Universidade de Lisboa, Portugal
3 School of Community Health, Charles Sturt University, Australia

*Corresponding author: rwhitsed@csu.edu.au

ABSTRACT

In regional and rural areas of sparsely-populated countries such as Australia, national health databases cannot provide data at a fine enough spatial scale to describe patterns within a rural community. Currently rural primary health networks rely on anecdotal evidence to identify where areas are with poor health outcomes, which is not enough information to fund local intervention plans. On the other hand, surveys of rural populations usually result in small datasets often criticized due to lack of representativeness of the general population.

To address this issue, we borrowed the concept of disease cluster analysis to analyze the representativeness of a small database compared to the background population, using location-based data sourced from a rural clinical health screening. The purpose is to evaluate whether the locations of the small number of participants attending the clinic are spatially representative of the background population in a rural Australian city, and can therefore enable further analysis of disease prevalence and distribution at the local level.

Spatial representativeness was quantified based on the uncertainty associated with the output of two known disease clustering methods, applied to meshblocks (the smallest geographical unit in the Australian census, equivalent to approximately 30-40 households).

One approach used a regression-based Generalized Additive Model (GAM) (Hastie and Tibshirani, 1991) to predict a continuous surface representing the relative number of participants in each meshblock. The GAM method was modified to incorporate multiple span sizes instead of just the optimal span size commonly used when applying GAM for disease cluster prediction (Whitsed et al., 2017). Following the approach of Aamodt et al. (2006), we were able to identify hotspots and coldspots, i.e., areas of over and under-representation of participants relative to population given by the relative risk value. Uncertainty was measured by mapping the reliability of hotspots and coldspots, obtained when overlapping the relative risk surfaces calculated for multiple span sizes.

Another approach used Poisson kriging (Goovaerts, 2005) to map the spatial distribution of the number of participants in each meshblock, considering its population size. In this case, the importance of meshblock population is incorporated through the addition of an error variance term in the block kriging system. Further, Oliveira et al. (2013) developed the Poisson kriging coupled with a block sequential simulation algorithm (Liu and Journel, 2009). Using this work, we produced simulations to map the location of potential hotspots and coldspots, representing the extreme values for the number of participants in each meshblock. In this case, uncertainty is measured by mapping the simulations variance, which represents the reliability of hotspots and coldspots.

We expect to devise a methodology that can be applied to any small health dataset, in particular in sparsely populated rural and regional areas, to identify undersampled areas (coldspots) and to provide an associated uncertainty measure for further disease mapping.

KEYWORDS: health data; spatial representativeness; uncertainty; GAM; Poisson kriging; Stochastic simulation
REFERENCES
Using Multiple Classifiers with AdaBoost in Google Earth Engine for Land Use/Cover Classification

Yuhao Wang *

1 Department of Geography, Michigan State University, USA

*Corresponding author: wangy118@msu.edu

ABSTRACT
Human activities are the major force of driven land use/land cover (LULC) in the past decades; it has a significant impact on the environment systems on the planet. Thus, analyses the LULC is an essential step before we make any further research. Different classification algorithms have been developed to classify the land; while each of them may have pretty good accurate result, but inevitably has disadvantages that none could produce a perfect classification result to fit all LULC categories (Chen et al., 2017). Therefore, a multiple classifier system is brought to combine multiple land classifiers’ results to make a better land classification.

Google Earth Engine (GEE) is a cloud-based platform that is easy to access high performance computing resource to process geospatial datasets online (Gorelick et al, 2017). This platform can provide researchers a seamless satellite data imagery with free of cost and fast response. It also provides many built-in land classifiers that are ready to go (Shelestov et al, 2017). Thus, in this short abstract, a method is proposed, based on the GEE platform to build a multiple classifier system with Adaptive Boosting (AdaBoost) to improve the land classification accuracy (Chen et al., 2017).

The methodology is still under development; however, a general framework and some preliminary process have been finished. The study area is around Zhangye oasis, northwest China. It is an arid region; an inland river, the Heihe River, flows through it. The primary dataset of this study is Landsat 8 OLI products on GEE in 2016, with atmosphere correction conducted.

The general framework of this project would be divided into two parts: the base classifier and a weighted multiple classifier system with AdaBoost. The base classifiers are Support Vector Machine, Classification and Regression Trees and Random Forests. After these base classifiers are trained with sample points, the base classifier will be tested by validation points, after the accuracy of each classifier is recognized, the weighted multiple classifier will use a weighted voting to finalize a classifier, this process will use adaptive boosting to iterate the weighted voting process several times to achieve the ideal classifier to classify the study area.

KEYWORDS: Land Classification; Google Earth Engine; Multiple classifier; Adaboost
Figure 1. The three base classifiers classification result

REFERENCES
Prediction of Urban PM2.5 Concentrations using a Bayesian Spatial-temporal Modeling Approach

Qiang Pu1, EunHye Enki Yoo*1

1Department of Geography, University at Buffalo, SUNY, USA

*Corresponding author: eunhye@buffalo.edu

ABSTRACT

Adverse health impacts of ambient fine particulate matter (PM2.5) on human are well-studied (Dominici, et al., 2006), although quantifying individuals’ exposure to PM2.5 is still a challenge. Estimation of spatially and temporally varying PM2.5 concentrations at fine scale, which is essential for individuals’ PM2.5 exposure risk assessment, requires a spatio-temporal prediction model. Spatio-temporal exposure models in air quality research tend to describe PM2.5 levels as a spatially correlated Gaussian field changing in time. For example, land use regression (Hoek et al., 2008) and downscaler models (Berrocal et al., 2010) are specifically designed under the consideration of meteorological conditions and spatially and temporally varying geographical features such as local-scale emissions. In terms of model inference, Bayesian approach with Markov Chain Monte Carlo (MCMC) has been widely used for its flexibility and capability to handle complex spatio-temporal correlation structures. However, convergence problems and its heavy computational burden limit feasibility particularly when dealing with large dataset. Integrated Nested Laplace Approximation (INLA) proposed by Rue et al., (2009) has been proved as an alternative to MCMC to provide fast Bayesian inference by using numerical approximation.

In the present paper, we estimated daily variation of PM2.5 concentrations at fine resolution in Beijing, China in 2015 using INLA combined with Stochastic Partial Differential Equations (SPDE) approach (Lindgren et al., 2011 and Cameletti et al., 2013). Our model characterizes the spatial dependence of covariates with Gaussian Markov Random Fields, while a first order autoregressive latent process determines temporal dynamics. Results showed that (for Jan.15th) PM2.5 concentrations have significant and positive relationships with temperature, humidity, major road length, but negative associations with wind speed and DEM. The estimated high autoregressive coefficient (0.89) indicated a presence of a strong correlation in time, and the range coefficient estimate of 305km also suggested the presence of a strong spatial correlation of PM2.5. Posterior mean prediction at 3 km resolution (Figure 1c) within a sparse triangulated domain obtained from SPDE (Figure 1b) revealed a heavy PM2.5 pollution event (average of 253 μg/m3). Areas of major “green” lands (northern and western parts) have relatively lower PM2.5 concentrations, while most populated urban and suburban areas fall into units with much higher PM2.5 concentrations. Lower and upper quantiles (2.5% and 97.5%) of posterior PM2.5 predictions in Figures 1d and 1e indicated that there is higher uncertainty at outer suburbs where the monitoring network was absent. We also conducted a cross validation analysis, which yielded a Root Mean Square Error (RMSE) of 16 μg/m3 and relative prediction error (RPE) of 9.5%. In comparison, RMSE and RPE of universal kriging were 56 μg/m3 and 25%, respectively. Based on our preliminary results, we conclude that the Bayesian inference approach of INLA-SPDE is efficient and effective in modeling the PM2.5 spatio-temporal process and mapping heterogeneity of urban PM2.5 concentrations at fine resolution. We expect the uncertainty of estimation will be useful for potential risk management.
Figure 1: PM2.5 concentration daily prediction at 3km resolution. (a) major land use classification; (b) Delaunay triangulation; (c) Posterior mean of PM2.5 estimates; (d) and (e) are posterior lower and upper 2.5% percentiles, respectively.

**KEYWORDS**: Spatio-temporal modeling; Bayesian inference; Integrated nested Laplace approximations; Particulate matter PM2.5

**References**


Dealing with Uncertainty in Real-time Sampling using Portable X-ray Fluorescence

Ana Horta *, João Neves 2, Leonardo Azevedo 2, Amilcar Soares 2, Rachel Whitsed 1

1 Institute of Land, Water and Society, Charles Sturt University, Australia
2 CERENA, Instituto Superior Técnico, Universidade de Lisboa, Portugal

*Corresponding author: ahorta@csu.edu.au

ABSTRACT

Soil contamination due to atmospheric deposition, agricultural practices, intensive industrial activities, inadequate waste disposal, mining and military activities, is still a challenging worldwide environmental issue. Although legislation may exist to safeguard public health and the environment, lack of effective monitoring but mostly the costs associated with contamination quantification and remediation, explain the existence of contaminated sites in Australia (NSW EPA, 2013), Europe (Panagos et al, 2013), and China (Lu, 2015).

A significant reduction of the costs associated to contamination assessments can be obtained if conventional soil sampling for contaminated-site characterisation is partly replaced by real-time sampling using proximal soil sensing devices such as the portable X-ray fluorescence (pXRF) (Crumbling, 2001; Clements et al., 2009).

Real-time sampling using pXRF, as a cheap and fast sampling method, has the potential to provide more data, to minimise lack of statistical representativeness, to reduce the time needed to evaluate the presence and the spatial extent of contamination (Horta et al. 2015).

The main disadvantage of using pXRF to collect soil data is the degree of uncertainty of these in-situ measurements due to its indirect nature, soil heterogeneity and soil moisture content (Rouillon et al., 2017).

To be able to confidently use pXRF real-time sampling data without compromising the accuracy of soil contamination quantification, it is important to characterize and integrate the uncertainty of pXRF measurements when mapping the spatial contamination extent.

The approach proposed by this work uses local probability distributions of pXRF measurements (herein referred to as local pdf-XRF) to characterize uncertainty at each sampling location and further generate stochastic simulations of the contamination extent, using the method proposed by Soares et al. (2017).

For each selected sampling location, collocated pXRF measurements and analytical lab data are taken to create a joint probability distribution. Conditional distributions are then derived to create the local pdf-XRF at each sampling location. The number of soil samples to be collected and analyzed can be optimized based on the type and extent of the contamination, and the information available for the contaminated site. These pdf-XRF can be continuously updated over time once new lab data is available.

After generating the local pdf-XRF, a stochastic sequential simulation algorithm with point probability distribution function (Soares et al., 2017) is used to produce equiprobable realizations of the contamination extent which integrate uncertainty through the local pdf-XRF. In a first step, all locations associated with pXRF measurements are visited following a random path. At each location a value is simulated using the local pdf-XRF and included in the set of conditioning data for the next location to be visited. After all the pXRF locations are visited these are included as experimental data for the stochastic sequential simulation of the remaining nodes of the simulation grid.

This approach was tested for a case study based on true pXRF measurements of heavy metals in a contaminated site. Local and spatial uncertainty are quantified to address the accuracy of the
contamination extent map and to identify areas which require additional sampling.

**KEYWORDS:** soil contamination; pXRF; joint probability; stochastic simulation

**References**


Assessing Spatial Uncertainty in 3D Stochastic Mapping of Soil Properties

Leonardo Azevedo *, Ana Horta 2, Amilcar Soares 1, Tiago Ramos 3, Rachel Whitsed 2

1CERENA, Instituto Superior Técnico, Universidade de Lisboa, Portugal
2Institute of Land, Water and Society, Charles Sturt University, Australia 3MARETEC, Instituto Superior Técnico, Universidade de Lisboa, Portugal

*Corresponding author: leonardo.azevedo@tecnico.ulisboa.pt

ABSTRACT

One of the most important issues in spatial modelling is how to assess the uncertainty intrinsically associated with the modelling procedure due to: poor quality of input data (for example, due to measurement errors or lack of sampling representativeness); model choice (or limitations in the mathematical models used to capture the complexity of the physical systems); and model implementation (estimation error in the model parameters) (Nelson, 2011).

Although possible to identify the uncertainty attached to each step and evaluate its propagation through the model (Heuvelink, 1998), the most common approach is to quantify the magnitude of the resulting uncertainty in the model predictions.

Depending on the objective sought, two different geostatistical approaches exist for uncertainty evaluation of model predictions, namely, the assessment of local uncertainty (i.e., the uncertainty prevailing at a particular location), or of spatial uncertainty (which measures uncertainty simultaneously at many locations). Measures of local uncertainty include assigning a confidence interval to the prediction, or to build an uncertainty distribution for each prediction location by using indicator or multiGaussian kriging (Goovaerts, 2001).

The assessment of spatial uncertainty implies spatial modelling using stochastic sequential simulation techniques. This is particularly relevant when the aim of the study is to use the resulting simulations as input into a physical model for further prediction. For example, in the context of Digital Soil Mapping, stochastic sequential simulation can be used to generate a set of multiple and equally likely realizations of soil properties which will serve as input data for crop models or flow simulators.

Within this context, the uncertainty attached to soil properties prediction is still performed in a simplistic framework by computing a measurement error (e.g., root-mean square error) between predicted values and sampled values not used as conditioning data for the stochastic modelling procedure. This approach, while allows to quantify the prediction error, lacks the assessment of how the model parameter space is being explored, and how different soil scenarios can be generated. This work builds on this premise and proposes a methodology for quantifying uncertainty in stochastic DSM using multidimensional scaling (MDS) (Scheidt and Caers, 2009) and stochastic sequential simulation (Soares, 2001).

A real DSM case study is presented dealing with the 3D mapping of soil texture properties (Ramos et al., 2017), needed as input data for a numerical flow model. A 3D stochastic sequential simulation technique (Soares, 2001) was used to generate a set of multiple realizations using a randomly selected calibration dataset as experimental data. A random validation dataset was also created to quantify and evaluate uncertainty using measurement errors and MDS to assess how the uncertainty space is being explored if the predictions are biased. Finally, and to assess the robustness of the proposed approach under soil uncertainty this procedure was applied to different sets of experimental and cross-validation datasets.
KEYWORDS: spatial uncertainty; MDS; stochastic simulation; soil

References


Geostatistical Prediction of Water Lead Levels in Flint, Michigan: a Multivariate Approach

Pierre Goovaerts *1

1BioMedware, Inc., USA

*Corresponding author: goovaerts@biomedware.com

ABSTRACT

The drinking water contamination crisis in Flint, Michigan has attracted national attention since extreme levels of lead were recorded in local water supplies and the percentage of children with elevated blood lead levels increased in neighborhoods with the highest water lead level (WLL). The culprit was the switch in water supply that resulted in water with high chloride and no corrosion inhibitor flowing through the aging water distribution system characterized by a high percentage of lead pipes and lead plumbing. Since Flint returned to its original source of drinking water more than 24,000 water samples have been collected and tested for lead in more than 10,000 residences. Concentrations above the EPA action level of 15 μg/L still remain and identifying residences at risk is critical to assess past exposure to lead and target future remediation. Concern over lead in water supply has also spread to other U.S. cities and we need an approach that can integrate sparse WLL tests with readily available secondary information, such as age of the house or presence of lead service lines, to predict the likelihood of elevated lead levels.

Despite several environmental crises, little research has been conducted on citywide geospatial modeling of water lead levels in public distribution systems. This paper presents the first application of multivariate geostatistics to lead in drinking water within a distribution system, specifically in Flint, Michigan. One of the key features of the Flint data is their collection through two different sampling initiatives: (i) voluntary or homeowner-driven sampling whereby concerned citizens decided to acquire a testing kit and conduct sampling on their own (10,717 sites), and (ii) State-administered sampling where data were collected bi-weekly at 809 selected sites after training of residents by technical teams (sentinel sites). These two datasets were first averaged over the 41-week sampling period and each tax parcel to attenuate sampling fluctuations and create a set of 420 tax parcels sampled by both protocols. Both variables displayed a correlation of 0.62 while their direct and cross-semivariograms showed substantial nugget effect and a long range of 9 km. WLLs recorded at sentinel sites and deemed more reliable by city officials were then interpolated using cokriging to account for the more densely sampled voluntary data and information on service line composition (lead, other, or unknown) available for each of 51,045 residential tax parcels. Cross-validation demonstrated the greater prediction accuracy of the multivariate geostatistical approach relative to kriging using only sentinel data. This general procedure is applicable to other cities with aging infrastructure where lead in drinking water is a concern.

KEYWORDS: cokriging; cross-validation; geostatistics
Uncertainty Characterization for Cropland Products at the Country Level from Different Sources

Jiong You*1,2, Zhiyuan Pei 1,2, Haiqi Liu1,2

1 Key Laboratory of Cultivated Land Use, Ministry of Agriculture, China
2 Chinese Academy of Agricultural Engineering, China

*Corresponding author: rsjone@sina.com

ABSTRACT

Monitoring the production of main agricultural crops is important to predict and prepare for disruptions in food supply and fluctuations in global crop market prices. Accurate and reliable information on the spatial distribution of major crops is needed for detecting possible production deficits with the aim of preventing food security crises and anticipating response planning. Some of the most widely used global land cover datasets are used in order to focus on cropland areas. These mainly include three kinds of derived land cover products from various sources. First, datasets based on single satellite sensors, for example, the 1 km Global Land Cover 2000, the 500 m MODIS Land Cover dataset, and the recently developed GlobeLand30 which provides global land coverage based on the exploitation of the Landsat archive at 30 m. Second, products rely on the development of a hybrid or synergic map by reconciling the best characteristics of several existing global datasets and integrating them into a single one or a single-category product such as the global cropland map from the IIAA-A-IFPRI. Other approaches also aggregate regional and national land cover products into a single global map such as FAO’s GLC-SHARE.

Global crop production monitoring systems need reliable input information; in particular, total cropland extent and location are absolutely necessary as they have a direct impact on anomaly warnings triggered by a percentage of cropland. However, although several land cover datasets exist, they were developed by using different mapping standards and classification methodologies, resulting in inevitable discrepancies of the available information, which leads users of land cover maps frequently find difficulties in selecting the best map for their specific application. Moreover, with the growing number of global datasets, it remains challenging to establish in a unique way in which a product is the most suitable for monitoring cropland, especially in areas with high fragmentation of landscape as is the case in China. Therefore, it is necessary to examine the comparative advantages of various global datasets for cropland monitoring and quantify uncertainty during the process of deriving cropland products at the national level from global land cover datasets.

This paper discusses uncertainty characterization for cropland products at the country level from different data sources. Global datasets include FAO’s GLC-SHARE produced by the United Nations’ FAO in 2014 at 1 km spatial resolution, GLC2000 developed by the European Commission’s JRC for the reference year 2000 at 1 km spatial resolution, and MODISLC derived by NASA for 2010 at 500 m spatial resolution. Discrepancies in the extent and spatial distribution of cropland were highlighted. Areas from each datasets were compared with FAO agricultural statistics at the country level to examine their comparative advantages for cropland monitoring. Finally, error modeling for uncertainty characterization was provided for downscaling from global to country level, with freely available reference datasets of GlobeLand30 developed by the National Geomatics Centre of China.

KEYWORDS: uncertainty; cropland; global land cover; country level; downscaling; error modeling;
Class-stratified Random Sampling with Sub-stratifications by Edge and Interior Pixels for Reference Data Collection to Increase Accuracy in Area Estimation

Jingxiong Zhang 1, Yue Wan 1, Wangle Zhang 1

Wuhan University, China;

*Corresponding author: jxzhang@whu.edu.cn

ABSTRACT

Area estimates of land cover and other thematic area classes are important for resources management and environmental modeling. Area estimation can be performed based on confusion matrices constructed from remote-sensing image classifications and reference data. For this, the direct method and the model-assisted post-stratification method are frequently used. Map-class-stratified random sampling is commonly used for collecting reference sample data to derive area estimates in combination with image classifications.

The accuracy of resultant area estimates depends on a variety of factors, such as reference sample sizes, accuracies of the classification maps being used, and performances of the area estimators employed. Given the cost of reference sample data collection, it is necessary to optimize sample configurations with consideration for limitations on sample sizes. Consider the accuracies of image classifications. Past research has found that edge pixels near class border-zones featuring greater spatial heterogeneity tend to be more likely mis-classified than interior pixels. It is sensible to allocate more sample units (thus greater inclusion probability) at edge pixels than at interior pixels to decrease their weights on area estimation so that effects of misclassification errors on accuracy of area estimation are reduced. With these in mind, we propose collecting reference data using class-stratified random sampling with sub-stratifications by edge and interior pixels to increase accuracy in area estimation. This proposed method is labeled SSTRA. An alternative method for area estimation exploiting the fact that classification accuracy is spatially varied is to divide a study area into regular sub-areas and to construct spatially constrained confusion matrices using the sample pixels falling within it. Area estimates for the whole study area are derived by summing up sub-area estimates. This alternative method is labeled SPCON.

An empirical study was carried out to test the proposed method SSTRA in comparison with method SPCON and the baseline method of direct estimation. It was confirmed that:

1) methods SSTRA and SPCON provide more accurate area estimation than the direct method; 2) method SPCON’s performance is inferior to that of method SSTRA, perhaps due to the fact that the number of sample points within some sub-areas may be too small, leading to reduced accuracy in area estimation. Therefore, the proposed method SSTRA is the most accurate for area estimation by combined use of image classifications and reference sample data.

KEYWORDS: area estimation; stratified random sampling; subarea; post-stratification estimation; edge; interior

I. The Introduction

Area estimation based on confusion matrices can utilize category information of sample points in classification map and reference map, with the classification map used as auxiliary data, to estimate area and proportion of classes. Area estimation methods based on confusion matrices include direct estimation, calibration estimation and model-based estimation. This paper chooses direct estimation and model-based post-stratified estimation.

Stratified random sampling is usually used to collect sample points in combination with
classification map to estimate area.

As to the classification map, pixels in the edge have greater probability of misclassification with more mixed pixels which reduces the accuracy of area estimation. Therefore, the stratified random sampling considering edge/interior pixels and category information is introduced in the paper. Greater inclusion probabilities attached to the edge pixels in the sampling technique.

Based on this sampling technique, two methods of area estimation are proposed. One is SSTRA which divides the research area into the edge part and the interior part and estimates area of each part respectively. The other is denoted as SPCON which takes spatial variability of classification accuracy into consideration and divides the study area into equal subareas to estimate area.

This paper uses classification map compiled in comparison with high-resolution images as reference data to evaluate area estimation of the two methods.

II. Research Area and Experimental Data

The area of the longitude and altitude of (114°27′43.06″~114°38′0.8″E, 30°26′15.46″~30°34′41.65″N) in Wuhan city is chosen as the study area. TM image is spatially subset into experimental data, which has 500 × 500 pixels. In comparison with high resolution images, the experimental data is processed into reference map as is shown in figure 1 and the classification map of figure 2. The classification hierarchy is set as {soil, water, architecture, forest, grassland} which is abbreviated to {s, w, b, f, g} in the following.

Figure 1 (a) and 1(b): The reference map and the classification map.

III. Experimental process

3.1 Area Estimation Methods Based on Confusion Matrice.

$p_{ij}$ is the ratio of the count of samples whose map class is $i$ and reference class is $j$ to the total sample count, and $p_k+$ and $p_+k$ are row and column marginal totals. $p_{+k}$ is mainly used in this paper, which is the ratio of sample count whose reference class is $k$ to the total sample count, to estimate area.

$$p_{+k} = \sum_{j=1}^{r} p_{jk}$$

Table 1. Sample proportion confusion matrix.

<table>
<thead>
<tr>
<th></th>
<th>ref</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>map</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>P11</td>
<td>P12</td>
<td>P13</td>
<td>P1+</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>P21</td>
<td>P22</td>
<td>P23</td>
<td>P2+</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>P31</td>
<td>P32</td>
<td>P33</td>
<td>P3+</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>P+1</td>
<td>P+2</td>
<td>P+3</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
3.1.1 Direct Estimation.

Direct estimation uses column marginal totals \((p+k)\), to estimate the ratio of pixels whose reference class is \(k\) to the total population in the research area as eq. (2)

\[
P_{+k} = \frac{\sum_{i=1}^{r} n_{ik}}{n} = \frac{n_{+k}}{n}
\]  

(2)

3.1.2 Model-based Post-Stratified Estimation.

Post-stratified estimation can be expressed as eq. (3), eq. (4) and eq. (5) from the view of model-based estimation according to Stehman (2013).

\[
\bar{Y} = \frac{1}{n} \sum_{i=1}^{r} N_{i+} \bar{Y}_{1}
\]  

(3)

\[
\bar{Y}_{1} = \frac{n_{ik}}{n_{i+}}
\]  

(4)

\[
W_{i} = \frac{N_{i+}}{N}
\]  

(5)

Combining eq.(3)(4)(5), eq. (6)is deduced as follows:

\[
P_{+k} = \bar{Y} = \frac{1}{N} \sum_{i=1}^{r} N_{i+} \bar{Y}_{1} = \sum_{i=1}^{r} W_{i} \frac{n_{ik}}{n_{i+}}
\]  

(6)

3.2 Stratified Sampling Design Considering Edge/Interior Pixels and Category Information.

In the sampling method adopted by this paper, pixels are stratified according to whether they are edge pixels or not and their category (Liu Meng, 2016). To be specifically, an pixel is distinguished between edge and interior by the categories count of its eight neighboring pixels. if the eight neighboring pixels have the same class as the center one, it is defined as an interior pixel; if not, it is defined as an edge pixel. In this way, the research area has 10 stratums which are \{se, si, we, wi, be, bi, fe, fi, ge, gi\}.

After the stratification, the sampling is processed, including two steps: sample size calculation and sample size allocation. Firstly, sample size (Cochran, 1977) is calculated using eq. (7).

\[
n = \left( \frac{Z_{1-\alpha/2}}{d} \right)^2 \left( \sum W_{i} S_{i} \right)^2
\]  

(7)

Where \(Z_{1-\alpha/2}\) is the confidence interval of standard normal distribution whose confidence coefficient is \(\alpha\), where \(\alpha\) is set to be 0.01, then \(Z_{1-\alpha/2} = 1.96\), \(d\) is overall allowed error of accuracy and specified to be 0.02, \(W_{i}\) is the proportion of each stratum and \(S_{i}\) is the standard deviation of the ith stratum. \(S_{i}\) can be achieved by the pre-sampling of the research area, using eq. (8)(Cochran, 1977) where \(U_{i}\) is the user's accuracy of class \(i\).

\[
S_{i} = \sqrt{U_{i}(1 - U_{i})}
\]  

(8)

In the pre-sampling process, 30 samples were randomly selected from each stratum, and the confusion matrice is constructed as table 2.\(S_{i}\) and \(U_{i}\) of each stratum can be computed using this confusion matrice, as shown in table 3.
Table 2. Pre-sampling confusion matrix.

<table>
<thead>
<tr>
<th></th>
<th>ref</th>
<th>Se</th>
<th>Si</th>
<th>We</th>
<th>Wi</th>
<th>Be</th>
<th>Bi</th>
<th>Fe</th>
<th>Fi</th>
<th>Ge</th>
<th>Gi</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>map</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Se</td>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Si</td>
<td>1</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>We</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Wi</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Be</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Bi</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Fe</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Fi</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Ge</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Gi</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>25</td>
<td>30</td>
</tr>
<tr>
<td>overall</td>
<td>28</td>
<td>30</td>
<td>24</td>
<td>31</td>
<td>32</td>
<td>29</td>
<td>40</td>
<td>32</td>
<td>29</td>
<td>25</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

The sample size is preliminarily computed with eq. (8), then samples are allocated to each stratum using Neyman Optimum Allocation where product of proportion and standard deviation is used as the weight of sample allocation as eq. (9). In eq. (9), \( n_i \) denotes the sample size in layer \( i \).

\[
n_i = n \frac{w_i s_i}{\sum_{i=1}^{k} w_i s_i} \tag{9}
\]

Then adjust the sample numbers to 20 for strata with too few samples and sample size in each stratum is shown in table 3. The final total sample count is 1018.

Table 3. Sample size distribution in each stratum

<table>
<thead>
<tr>
<th></th>
<th>( W_i )</th>
<th>( U_i )</th>
<th>( S_i )</th>
<th>Sample counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Se</td>
<td>0.068076</td>
<td>0.900000</td>
<td>0.300000</td>
<td>63</td>
</tr>
<tr>
<td>Si</td>
<td>0.02554</td>
<td>0.933333</td>
<td>0.250000</td>
<td>20</td>
</tr>
<tr>
<td>We</td>
<td>0.026444</td>
<td>0.800000</td>
<td>0.400000</td>
<td>66</td>
</tr>
<tr>
<td>Wi</td>
<td>0.081816</td>
<td>1.000000</td>
<td>0.000000</td>
<td>20</td>
</tr>
<tr>
<td>Be</td>
<td>0.198464</td>
<td>0.833333</td>
<td>0.370000</td>
<td>228</td>
</tr>
<tr>
<td>bi</td>
<td>0.065204</td>
<td>0.900000</td>
<td>0.300000</td>
<td>61</td>
</tr>
<tr>
<td>Fe</td>
<td>0.035504</td>
<td>0.933333</td>
<td>0.250000</td>
<td>28</td>
</tr>
<tr>
<td>Fi</td>
<td>0.012988</td>
<td>1.000000</td>
<td>0.000000</td>
<td>20</td>
</tr>
<tr>
<td>Ge</td>
<td>0.242444</td>
<td>0.800000</td>
<td>0.400000</td>
<td>301</td>
</tr>
<tr>
<td>Gi</td>
<td>0.235536</td>
<td>0.833333</td>
<td>0.370000</td>
<td>271</td>
</tr>
</tbody>
</table>

3.3 SSSRA: Area Estimation of Edge/Interior Subregions.
Firstly, the research area is divided into edge subregion and interior subregion, and area is estimated for each of them. This kind of area estimation is denoted as SSSRA. Pixels in the edge part have map classes of \{se,we,be,fe,ge\} and pixels in the interior part have map classes of \{se,we,be,fe,ge\}. The confusion matrix is constructed according to sample points falling in each subregion. Direct estimation and post-stratified estimation are applied to SSSRA respectively. Firstly, eq. (2) is used to estimate area \( p_{k} \) (\( k = 1, \ldots, 5 \)) in the edge subregion and the interior
subregion. Area estimations in subregions are added up and proportions of each subregion serve as weights. Area estimation based on the direct estimation is achieved and marked as result 1(1). Secondly, eq. (6) is used to compute $W_i$, $n_i^+$ and $n_i k$ in the subregions where $W_i$ is the proportion of pixels whose map class is $i$, $n_i^+$ is the count of samples whose map class is $i$ and $n_i k$ is the count of samples which have map class of $i$ and reference class of $k$ ($i, k = 1, \ldots 5$). Then area estimation $p_{i+k}$ ($k = 1, \ldots 5$) is computed for each subregion. Finally, area estimation of the research area based on the Post-Stratified estimation is achieved by adding up area estimation in subregions where proportions of subregions serve as weights. This result is set as result 1 (2).

3.4 SPCON: Area Estimation of Equal Subregions.

The research area is divided into equal subregions and area is estimated separately. This kind of area estimation is denoted as SPCON. Two experiments were designed where there are four equal subregions and one hundred equal subregions respectively.

3.4.1 Area Estimation of 4 Equal Subregions.

In this section, the research area is divided into 4 equal subregions, each of which has the size of $250 \times 250$, whose center pixels are (125,125), (125,375), (375,125), (375,375). The confusion matrix is constructed using the sample points falling in each subregion respectively. According to eq. (2), $p_{i+k}$ ($k = 1, \ldots 5$) is calculated. Adding up area estimation in the four subregions, and area estimation based on the direct estimation is achieved which is set as result 2(1); According to eq. (6), $p_{i+k}$ ($k = 1, \ldots 5$) is calculated. Adding up area estimation in the four subregions, and area estimation based on the post-stratified estimation is achieved which is set as result 2(2).

Area Estimation of 100 Equal Subregions.

In this section, the research area is divided into 100 equal subregions, each of which has the size of $50 \times 50$, whose center pixels are (25,25), … (475,475). The confusion matrix is constructed respectively using the sample points falling in each subregion. According to eq. (2), $p_{i+k}$ ($k = 1, \ldots 5$) is calculated. And area estimation based on the direct estimation is achieved by adding up area estimation in the one hundred subregions, which is set as result 3(1); According to eq. (6), $p_{i+k}$ ($k = 1, \ldots 5$) is calculated. Area estimation based on the post-stratified estimation is achieved by adding up area estimation in the one hundred subregions, which is set as result 3(2).

3.5 Area Calculation of Reference Map as Validation Data

The total coverage of reference classification map is used as the validation data to compare the performance of SSTRA and SPCON and the area distribution in the reference map is set as result 4.

IV Experimental Results and Analysis.

Experimental Results

According to the experimental scheme in subsection 4.0, area estimation results of SSTRA and SPCON are shown in table 5. When those area estimation results are displayed in a histogram (as figure 1), it can be concluded that area estimation of post-stratified estimation (labeled (2)) is superior to area estimation of direct estimation (labeled (1)), which is closer to the real area distribution. Comparing the results of post-stratified estimations in SSTRA and SPCON using a histogram (as figure 2), the conclusion can be drawn that, area estimation of SSTRA is superior to that of SPCON based on post-stratified estimation, closer to the real area distribution.
Table 4. Area estimation results of each methods

<table>
<thead>
<tr>
<th></th>
<th>s</th>
<th>w</th>
<th>b</th>
<th>f</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>7.3231%</td>
<td>9.8948%</td>
<td>25.4923%</td>
<td>12.1673%</td>
</tr>
<tr>
<td>1</td>
<td>(2)</td>
<td>7.0448%</td>
<td>10.3203%</td>
<td>25.3800%</td>
<td>10.8291%</td>
</tr>
<tr>
<td>2</td>
<td>(1)</td>
<td>7.1787%</td>
<td>8.0310%</td>
<td>26.5087%</td>
<td>11.9040%</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>8.9859%</td>
<td>10.4792%</td>
<td>26.3341%</td>
<td>11.7194%</td>
</tr>
<tr>
<td>3</td>
<td>(1)</td>
<td>7.1993%</td>
<td>7.4510%</td>
<td>25.2893%</td>
<td>11.3204%</td>
</tr>
<tr>
<td>3</td>
<td>(2)</td>
<td>9.0465%</td>
<td>10.8397%</td>
<td>25.9440%</td>
<td>11.4566%</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>9.1636%</td>
<td>11.0022%</td>
<td>25.9254%</td>
<td>10.8647%</td>
</tr>
</tbody>
</table>

Figure 2. Comparisons of area estimations of the direct estimation and post-stratified estimation in the method of SSTR and SCON.

Figure 3. Comparisons of area estimations of post-stratified estimation in the method of SSTR and SCON.
4.2 Analysis

1) As a whole, the accuracy of post-stratified estimation is better than that of the direct estimation; 2) Method SSTRA is superior to method SPCON; 3) For post-stratified estimation, area estimation based on equal subregions in regard of accuracy is no better than area estimation without dividing subregions. What’s more, when there are too many subregions, given the too few samples in some subregions, accuracy of area estimation in those subregions are seriously harmed; 4) The reason why area estimation of SSTRA has better accuracy lies in that the method considers the difference of classification accuracy between edge pixels and interior pixels in a class and fully exploited the sampling scheme to express the difference of classification accuracy between strata. Therefore, in this method the error distribution compared with the reference map is well manifested.

REFERENCES


High Accuracy Surface Modelling

Tianxiang Yue *1, Zemeng Fan 1, Zhengping Du 1, Na Zhao1

1 Institute of Geographical Sciences and Natural Resources Research, China

*Corresponding author: vue@lreis.ac.cn; zhaon@lreis.ac.cn

ABSTRACT

Satellite observations can frequently supply spatially continuous information about the Earth’s surface, which is impossible from ground-based investigations, but remotely sensed data by satellites are not able to directly obtain process parameters. Ground observations are able to obtain highly accurate data with high temporal resolution at observation points, but these observation points are too sparse to satisfy some application requirements. The most effective method for Earth’s surface modelling entails the integration of satellite observations with ground observations. However, the full integration was ignored in most of the methods. For finding a solution for this problem, we suggest an alternative method, high accuracy surface modeling (HASM), which takes global approximate information (e.g., satellite observation data) as its driving field and local accurate information (e.g., ground observation data) as its optimum control constraints. A Fundamental Theorem of Earth Surface Modelling (FTEM) is abstracted from HASM development and applications in recent 30 years.FTEM can be described as “a The Earth’s surface or a component surface of the Earth’s surface environment is uniquely defined by both extrinsic and intrinsic invariants of the surface, which can be simulated with an appropriate method for integrating the extrinsic and intrinsic invariants, such as HASM, when the spatial resolution of the surface is fine enough to capture the attribute(s) of interest”. From FTEM, seven corollaries have been deduced, corresponding to interpolation, upscaling, downscaling, data fusion and data assimilation respectively. HASM and FTEM have been successfully applied to simulating surfaces of elevation, soil properties, changes of ecosystem services, and driving forces of the changes on multi-scales.

KEYWORDS: Satellite Observations; Ground Observations; Fundamental Theorem of Earth Surface Modelling
Uncertainties in using Kriging Method to Upscale in situ Soil Moisture to Multiscale Pixel Estimations

Wenjun Zuo¹, Xiaohu Zhang *¹

¹ Nanjing Agricultural University, China

*Corresponding author: zhangxiaohu@njau.edu.cn

ABSTRACT

Using kriging method to upscale in situ soil moisture observations (ISMO) to multiscale pixel estimations is a primary step in the comprehensive usage of ISMO and remote sensing (RS) soil moisture data. However, scale effects would occur and introduce uncertainties during upscaling processes because of spatial heterogeneity and the kriging method. To illustrate multiscale ISMO upscaling processes, a nested hierarchical scale series was established at the field level, and upscaled estimations at each scale were obtained with kriging method. Those uncertainties were described with the results of comparison analysis against RS data and the statistical analysis, the spatial trend surface analysis on multiscale estimations. Then the uncertainties were explained from the spatial heterogeneity perspective with a semivariogram analysis on ISMO. The results show that uncertainties exist and vary in multiscale upscaling processes with block kriging, and the range of the empirical semivariogram could indicate scale effects. When the target scale is shorter than the range, block kriging estimations, rather than the ordinary kriging, maintain similar scale effects and global trends during upscaling processes. And the direct pixel estimation by BK is relatively close to the average of nested pixel estimations.

KEYWORDS: Uncertainty; Scale Effect; Kriging; Semivariogram; Soil Moisture
Study on Topographic Correction Algorithm of Leaf Area Index Products Derived from Remote Sensing Data

Yuetong Hu1,2, Xianfeng Feng*1,2

1 State Key Laboratory of Resource and Environmental Information System, IGSNRR, CAS, China
2 University of Chinese Academy of Science, China

*Corresponding author: fengxf@lreis.ac.cn

ABSTRACT

Leaf Area Index (LAI) is an important parameter to describe the vegetation canopy structure and a key input of many ecological and hydrological models. Therefore, the accurate estimation of LAI is of great significance in climate, weather, and ecological studies. At present, remote sensing method is the only way to obtain LAI in large area, which can also generate stable and reliable LAI products. However, the precise of leaf area index products derived from remote sensing data need to be improved because of terrain errors, especially in hilly area.

Topography of hills and mountainous in South China changed significantly. And there has full hydrothermal conditions, high vegetation coverage, and high research value in ecology. However, topography has a relatively large impact on the estimation of remote sensing leaf area index in this region. In this study, Qianyanzhou was chosen as a typical study area. Based on the standard LAI products, DEM data, and vegetation type data in the typical region, the paper studied on the terrain correction algorithm for the GLOBMAP LAI product considering the vegetation types. And then a terrain correction model was established and the correction results were analysed and verified.

The main research contents and conclusions are as follows.

(1) Terrain correction algorithm of LAI products based on elevation standard deviation. The three-time fitting function of elevation standard deviation and LAI indicates that elevation standard deviation affects LAI. The difference between actual LAI and LAI product can be expressed by the mathematical expression of elevation standard deviation. Based on this relationship, a statistical model of LAI product terrain correction based on elevation standard deviation can be established, and shrubs had the highest correlation of 0.986.

(2) Terrain correction and results verification of LAI products. Using the LAI product terrain correction model established in this paper to calibrate the GLOBMAP LAI products in the typical areas, the average corrected LAI values of the two typical areas were 1.89 and 1.84, respectively, which were all reduced. The difference between the corrected average LAI and the ground measured data decreased from -0.25 to +0.04 after correction. The change trend of corrected LAI with elevation standard deviation was closer to the trend of ground data, but the applicability of the model remained to be improved.

KEYWORDS: Leaf Area Index (LAI); LAI product; Topographical Correction; Hilly Region; Standard Deviation of Elevation

I. INTRODUCTION

The Leaf Area Index (LAI) is defined as half the total intercepting area per unit ground surface area. LAI is an important parameter which describes the canopy structure of vegetation. It controls the
biophysical processes such as photosynthesis, respiration and transpiration of plants and is important for understanding the biophysical processes of forests and crops. LAI plays an important role in accurate assessment of various parameters such as photosynthetically active radiation, evapotranspiration, and crop yield. And it is an extremely important input parameter in physical and chemical models of ecosystems. Therefore, the accurate estimation of LAI is of great significance.

Using satellite imagery to retrieve LAI is the only way to obtain a large range of LAI, which also has the advantage of long-term, rapid, real-time and dynamic. At present, a variety of global and regional LAI products have been retrieved through remote sensing methods, such as GLOBMAP LAI, GLOBCORBON LAI and MODIS LAI. These model algorithms often ignore the topographic variables and satellite imagery as one of input data are not topographical corrected. Therefore, the precise of leaf area index products derived from remote sensing data need to be improved because of terrain errors, especially in hilly area.

The existing method of estimate leaf area index from satellite imagery incorporating topography information can reduce the terrain error in LAI products, but terrain correction algorithm for the existing LAI products which has higher application value for many product users is still blank.

A topological correction scientific algorithm for LAI products which takes topography information into account proved to reduce the terrain error in the existing LAI product (taking GLOMAP LAI as an example). The difference between remote sensing LAI product ignoring terrain and actual LAI has a good correlation with terrain factors. A set of 30-m spatial resolution LAI in Qianyanzhou based on the exponential relationship between Landsat-based NDVI and measured LAI was used as the standard LAI. Standard deviation of elevation calculated by DEM was chosen to establish a relationship with LAI difference among different types of vegetation. The calibration model was determined based on the fitting coefficients, and the GLOBMAPLAi product near Qianyanzhou was corrected in 500m.

II. METHODS

GlobeLand30 data includes 10 land use types such as arable land, forest, grassland, shrub, water, and wetland. The Vegetation Map of China (1:1000000) contains vegetation information such as coniferous forest, broad-leaved forest, and mixed forest. According to these data, a 30-m biome class map was obtained by reclassification and resampled to 500m. The codes of biomes as follows (see Table 1).

<table>
<thead>
<tr>
<th>Biome</th>
<th>Code</th>
<th>Code in GlobeLand30</th>
<th>Code in Vegetation Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conifer Forest</td>
<td>1</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Deciduous Forest</td>
<td>2</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Mixed Forest</td>
<td>3</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Shrub</td>
<td>4</td>
<td>40</td>
<td>4</td>
</tr>
<tr>
<td>Grass and Crop</td>
<td>5</td>
<td>10,30</td>
<td>7,8,11</td>
</tr>
<tr>
<td>Nonvegetated</td>
<td>0</td>
<td>80</td>
<td>9,0</td>
</tr>
</tbody>
</table>

GLOBMAP LAI product is based on AVHRR and MODIS data and is a globally consistent leaf area index data since 1981, provided by the Institute of Geographical Sciences and Natural Resources, Chinese Academy of Sciences. The 500-m data from the day of 201~208 in 2008 was used in the study of terrain correction algorithms.

The standard LAI data of Qianyanzhou in 30m was provided by Nanjing University. Firstly, six vegetation indexes (SR, NDVI, PVI, SAVI, RSR and MNDVI) of Qianyanzhou were calculated using the three TM images in 2008. Measured LAI of 47 forest plots in Qianyanzhou area from July 24th to 27th, 2008, were fitted with these indexes. Finally, the 30- m LAI data was obtained from the statistical model based on the NDVI with the highest $R^2$ of 0.68.
30-m ASTER GDEM V2 data was acquired from the Geospatial Data Cloud website. Neighbourhood analyst tool was used to generate elevation standard deviation data and the size of neighbourhood was set in 17×17cells in 30m, which is consistent with the cell size of the GLOBMAP LAI. The size of 500-m elevation standard deviation data was 3×3cells. The correlation of these topographical factors with the difference between true LAI and GLOBMAP LAI was calculated and the factor with highest correlation was used for the establishment of the topographical correction model.

The elevation standard deviation is a more effective factor that represents topographic relief. Topographically corrected LAI show obvious trend with increasing standard deviation of elevation (σ). Difference between actual LAI and LAI product values are formed by the contributions from terrain effects. The effect of topography was removed in four steps. First, GLOBMAP LAI data set was resampled to 30m and then LAId (difference between TM LAI and GLOBMAP LAI) was calculated pixel-by-pixel without the non-vegetation type and zero-value data. Second, LAId data sets of all biomes were binned into 5-m interval classes by the value of σ for given pixel, and the mean LAI value was calculated for each biome. Third, five cubic polynomials were fitted using LAId and corresponding medians of different σ intervals for conifer forest, deciduous forest, shrub, grass and crop and all of vegetation (lack of mixed forest). Last, the fitting parameters were calculated and terrain correction models for different biome class were obtained as follows.

\[
LAI_{TC} = LAI_{GM} + (p_1\sigma^3 + p_2\sigma^2 + p_3\sigma + p_4)
\]  

where LAITC describes LAI after topographic correction, LAIGM represents GLOBMAP LAI product, and \(p_1, p_2, p_3, p_4\) respectively, represent the correction parameters of each vegetation type.

### III. RESULT

The different vegetation fitting parameters are shown in Table 2.

<table>
<thead>
<tr>
<th>Biome</th>
<th>(p_1)</th>
<th>(p_2)</th>
<th>(p_3)</th>
<th>(p_4)</th>
<th>(R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conifer Forest</td>
<td>0.00003048</td>
<td>-0.0044997</td>
<td>0.19269302</td>
<td>1.36595162</td>
<td>0.91730000</td>
</tr>
<tr>
<td>Deciduous Forest</td>
<td>0.00002080</td>
<td>-0.00393100</td>
<td>0.19881218</td>
<td>1.34375054</td>
<td>0.95670000</td>
</tr>
<tr>
<td>Shrub</td>
<td>0.00005631</td>
<td>-0.00843708</td>
<td>0.34033509</td>
<td>0.97133438</td>
<td>0.90630000</td>
</tr>
<tr>
<td>Grass and Crop</td>
<td>0.00002421</td>
<td>-0.00405818</td>
<td>0.18739643</td>
<td>1.55071412</td>
<td>0.97160000</td>
</tr>
<tr>
<td>All of Vegetation</td>
<td>0.00003740</td>
<td>-0.005505945</td>
<td>0.230329188</td>
<td>1.32661968</td>
<td>0.95810000</td>
</tr>
</tbody>
</table>

The histogram of elevation standard deviations and differences of all vegetation is shown in Figure 1.

![Figure 1: Histogram of elevation standard deviations and LAI Differences of all vegetation](image-url)
When biome class is mixed forest, it can be corrected with all types of parameters. Since the elevation standard deviation of the model data was only within 90m, the model was only applicable to areas with an elevation standard deviation of less than 90m. The corrected LAI obtained from the model is shown in Figure 2.

![Figure 2: comparison of LAI retrievals before and after terrain correction. (a) After correction, (b) before correction](image)

After terrain correction, LAI in most areas has been reduced. LAI was significantly increased after correction in the black circle in the flat areas with low elevations. Blue circle was partly in the hilly area where LAI was reduced after correction. The average LAI after correction is 1.8893, and the maximum value is 10.2318. Before correction, the average and maximum are 2.8007 and 10.07. The minimum value is still 0.

Comparing with 31 plots of ground measured LAI data with 2.72 in average, the average LAI are changed from 2.47 to 2.76 after correction.

**IV MAIN CONCLUSIONS**

The main research contents and conclusions are as follows:

1. The difference between LAI and LAI can be expressed by the mathematical expression of elevation standard deviation. Based on this relationship, a statistical model of LAI product terrain correction based on elevation standard deviation can be established, and shrubs had the highest correlation of 0.986.

2. After using the LAI product terrain correction model established to calibrate the GLOBMAP LAI products in study area, the average corrected LAI value reduce from 1.89 to 1.84. The difference between the corrected average LAI and the ground measured LAI decreased from -0.25 to +0.04 after correction. The trend of corrected LAI with elevation standard deviation increasing was closer to the trend of ground data, but the applicability of the model remained to be improved.
The Monitoring Method of Economic Forest in Southern Jiangsu Based on the High-Resolution Remote Sensing Images

Jia Jin*1,2, Xianfeng Feng 2, Zhou Lu 2, Liang Pei 1, Ming Luo 2

1 Liaoning Technical University, China
2 IGSNRR, CAS, China

*Corresponding author: fengxf@lreis.ac.cn

ABSTRACT

In order to extract economic forest information in southern Jiangsu accurately and quickly, and to complete the information management of it, the paper uses GF-2 images which have the high resolution and GF-1 images which have the medium resolution as data sources, and makes full use of the complementarity between multi-source remote sensing data. Meanwhile, combined with digital elevation model, forest patches, the survey data and other auxiliary information, the paper uses ENVI and eCognition as the data processing platform of remote sensing and IDL programming language to extract economic forest. Southern Jiangsu is chosen as study area where is spread of crushing field, scattered forest and economic forest types with regional characteristics, in order to probe the best monitoring method in the study area. The result shows that the best method for monitoring economic forest in southern Jiangsu is the object-oriented classification with the combination of multi-level, multi-rules and multi-classifiers and the accuracy is high which can fully satisfy the requirement of practical production.

KEYWORDS: economic forest; southern Jiangsu; high-resolution remote sensing images; remote sensing monitoring; object-oriented

I. INTRODUCTION

In recent years, with the continuous expansion in the field of high-resolution remote sensing application, especially the improvement of spatial resolution and the application of country scale, remote sensing provides a strong support to extract, monitor and manage economic forest. Therefore, on the basis of the existing forestry remote sensing research, it has practical significance for study of automatic monitoring methods for economic forest. Furthermore, how to quickly and accurately extract economic forest species with the high-similarity from the complex topography is a key technology to be solved.

In China, forest is divided into five categories: shelter forest, timber forest, economic forest, firewood forests and forest with special purposes. Dihkan, Guneroglu, et al(2013) extracted spectrum and texture feature from high resolution image based on Gabor filter, used MNDVI to distinguish green vegetation and the other LULC classes and used support vector machine (SVM) to classify and get the land-use map. Saumitra Mukherjee, et al(2013) used supervised classification combined with the spectral characteristics and phenological characteristics to obtain the distribution of tea and estimate the output in Sri Lanka. Luo Wei, et al(2014) used HJ-1 images and made full use of NDVI, band operation, topography and geometry characteristic to extract orchard by decision tree model. It realized the extraction of citrus and navel orange. YueJun, Wang Zhenxi, et al(2015). used maximum likelihood, markov distance, neural network and SVM to distinguish four main fruit trees: walnut, apple, red jujube and pear in southern Xinjiang basin based on the 2 m, 8 m, 16 m GF -1 data which has different resolutions. According to spectrum, texture feature and NDVI threshold to distinguish. The result showed that the SVM was the best method which based on the 2 m high-resolution images of GF-1.
Guang-ming Xu (2016) used high-resolution images and multi-source data fusion as data resources. Based on object-oriented multi-level classification rules, extracted tea by analysis of spectral, texture and topography. In addition, compared with the maximum likelihood method, SVM and random forest. The results showed that the multi-level classification rules was better. Feng Zhenfeng (2016) used data dimension reduction method to select the best band combination of Hyperion image. It used maximum likelihood, neural network and SVM to identify main economic forest tree in southern Xinjiang’s basin. Moreover, it used the object-oriented algorithm which was multi-scale segmentation for QuickBird image to obtain economic forest boundary. Finally, obtained the forest compartment on the basis of information recombination automatically. The whole process diagram is shown in Figure 1.

Figure 1: Technology route

II. MATERIALS AND METHODS

Southern Jiangsu is located in the south of the Jiangsu province which has the features of the crushing field, scattered forest and economic forest types with regional characteristics., it includes five cities: Nanjing, Suzhou, Wuxi, Changzhou and Zhenjiang as shown in Figure 2, covers 25532km² which accounting for 26.67% of the total area in Jiangsu. Main economic forest contains: tea and orchard. The paper mainly adopts GF-2 which has the high spatial resolution and GF-1 which has the medium spatial resolution and high temporal resolution as remote sensing data sources. Meanwhile, combined with digital elevation model, forest patches, the survey data and other auxiliary information, uses ENVI and eCognition as the data processing platform of remote sensing and IDL programming language to extract economic forest.
Establishing the features knowledge set of spectral characteristics, phonological features and spatial characteristics (such as texture features, shape features and terrain) to digging endemic information of economic forest in bulk with IDL. And get the optimal combination quickly and accurately by analyzing that information. Then, constructing multi-classification model based on pixels and image objects, respectively. To be specific, the paper selects unsupervised classification such as K-Means, ISODATA, the traditional supervised classification such as maximum likelihood, minimum distance and decision tree classification based on expert knowledge set as the classification model with pixels. On the other hand, makes use of multi-scale segmentation algorithm which uses ESP plug-in to select the optimal segmentation scale, as shown in Figure 3, and takes advantage of the combination of multi-level, multi-rules and multi-classifiers to realize oriented-object classification. Finally, verifying the classification results and the extraction accuracy of economic forest from qualitative and quantitative perspectives, the result can conclude that the best method for monitoring economic forest in southern Jiangsu is the object-oriented classification model, which is the combination of multi-level, multi-rules and multi-classifiers.

III. RESULTS AND DISCUSSION

The results show that the production accuracy of economic forest in proper order is: Object-oriented
hierarchical classification>decision tree classification>maximum likelihood>K-Means> mahalanobis distance>ISODATA>minimum distance>parallelepiped> neural network method, and the user accuracy in proper order is: Object-oriented hierarchical classification>maximum likelihood>decision tree classification>neural network method>parallelepiped>minimum distance>mahalanobis distance>K-Means> ISODATA. In summary, it can conclude that the best method for monitoring economic forest in southern Jiangsu is the object-oriented classification with the combination of multi-level, multi-rules and multi-classifiers. The classification rule as shown in Table 1, the classification result as shown in Figure 4 and the extraction result as shown in Figure 5.

<table>
<thead>
<tr>
<th>Level</th>
<th>Attribute</th>
<th>Classes</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>Non-Vegetation</td>
<td>DEM&lt;40 and NDVI&lt;-0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vegetation</td>
<td>not Non-Vegetation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Water</td>
<td>NDWI≥0.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Woodland</td>
<td>NDVI_cha≥0.2 and not Water</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>Construction Land</td>
<td>Mean Nir≤190 and not Water and not Non-Woodland</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bare Land</td>
<td>not Water and not Non-Woodland and not Construction Land</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vegetation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Water</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Woodland</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Existence of super objects Level1 Distance=1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Tea</td>
<td>Feature Space Optimization 19/39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other Woodland</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Water, Construction Land, Non Woodland, Bare Land</td>
<td>Existence of super objects Level2 Distance=2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>Orchard, Tea, Other Woodland</td>
<td>Existence of super objects Level3 Distance=1</td>
</tr>
</tbody>
</table>

Figure 4: The classification of typical region in southern Jiangsu

According to the latest statistical data of 2017, the total area of orchard and tea in the typical regions
Dongshan and Jingting town is 44.8km² and 19.33km², respectively. Meanwhile, compare with calculated values by Arcgis and calculate the ratio of the area with the statistical data and result as relative accuracy. The classification accuracy as shown in Table 2. Besides, by the confusion matrix based on validation sample, the overall classification accuracy is up to 88.1622%, Kappa coefficient is 0.8755, economic forest production precision is 0.8978 and the user accuracy is 0.8129. The effect of economic forest information has good consistency and can meet the requirements of practical application. As a consequence, the model can be extended to southern Jiangsu to realize monitoring economic forest based on the high-resolution remote sensing images.

![Image](image_url)

**Figure 5**: The extraction results of economic forest in typical region of southern Jiangsu

<table>
<thead>
<tr>
<th>Area(km²)</th>
<th>Orchard</th>
<th>Tea</th>
<th>Economic Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Data</td>
<td>44.8</td>
<td>19.33</td>
<td>64.13</td>
</tr>
<tr>
<td>Extraction Results</td>
<td>42.67</td>
<td>17</td>
<td>59.67</td>
</tr>
<tr>
<td>Relative Accuracy</td>
<td>95.24%</td>
<td>87.93%</td>
<td>93.04%</td>
</tr>
</tbody>
</table>

**REFERENCES**


Cross-Calibration of Passive Microwave Satellite Brightness Temperatures Observed by F13 SSM/I and F17 SSMIS for the Retrieval of Thickness and Area on Antarctic Thin Ice

Yue Liu\textsuperscript{*1}, Xi Zhao\textsuperscript{1}, Xiaoping Pang\textsuperscript{1}

\textsuperscript{1}Wuhan University, China

*Corresponding author: yue.liu@whu.edu.cn

ABSTRACT

The thickness of thin ice (<0.2m) in the Antarctic is not only a significant geophysical variable, but also an important parameter for the study of sea ice mass and energy balance. Brightness temperature has been used to measure the thin ice thickness due to the good linear relationship between thin ice thickness and polarization ratio values, which is computed using vertically and horizontally polarized brightness temperature values. Few of today’s satellite sensors can continually work for more than ten years. So it becomes much necessary to conduct cross-calibration of sensors with similar physical configuration and over-lapped observation to construct a consistent time series images. F13 special sensor microwave/imager (SSM/I) is a seven-channel, four-frequency, linearly polarized passive microwave radiometer system. Its successor, F17 Special Sensor Microwave Imager / Sounder (SSMIS), is an enhanced eleven-channel, eight-frequency system. Due to the difference in space, time and radiation measurement characteristics of these two sensors, the time series observation data will produce some systematic deviations. In particular, the SSMIS sensor replaced the 85GHz with the 91GHz so that the bright temperatures observed from the high frequency channel are no longer the same. In this study, we cross-calibrated the brightness temperatures from F13-SSM/I and F17-SSMIS, and evaluated the error propagation from calibration to thin ice thickness retrieval and ice area estimation. To cross-calibrate the brightness temperatures for both H and V channels at SSM/I 85GHz and SSMIS 91GHz, we select the overlapped observation during the period of June ~August 2007 and June ~August 2008. F13 and F17 brightness temperature data in the overlapping period are fitted by a least-square linear model according to different days, and then the daily fitting coefficients are averaged to get the monthly scale cross-calibration model. In order to avoid the influence of land, the land part of the study area was masked out. Based on the thin ice thickness retrieval method (Tamura T et al. 2007), the differences of thin ice thickness and thin ice area in Antarctic before and after assimilation were estimated and analyzed. The results show that the correlations between the brightness temperature of the two channels are above 0.85. Moreover, the correlations between the H channels of SSM/I 85GHz and SSMIS 91GHz are superior to that of the V channels (above 0.90). The RMSEs of brightness temperature are significantly different, with the H-channel's RMSE around 7.0K and the V-channel’s RMSE around 3.5K. The daily mean difference of Antarctic thin ice thickness before and after the cross-calibration is 0.680% (less than 1cm), whereas the daily mean difference in thin ice area is 2.492% (more than 360km\textsuperscript{2}) in the Antarctic. Thus, cross-calibration has little effect on thin ice thickness retrieval but has a large impact on thin ice area estimation.

KEYWORDS: Cross-Calibration; Brightness Temperature; Thin Ice Thickness; Thin Ice Area; Antarctic

REFERENCES

Comparison between Three Split Window Algorithms on Ice Surface Temperature Retrieval in the Arctic

Pei Fan*, Xi Zhao, Xiaoping Pang, Qing Ji

*Wuhan University, China

ABSTRACT
Leads are linear or wedge-shaped openings in the sea ice cover. They account for about half of the sensible heat transfer from the Arctic Ocean to the atmosphere in winter, though the sea surface area covered by them is only 1%-2% of the total sea ice area. To extract lead distribution from thermal images, we need to carefully retrieve the ice surface temperature (IST) first. Although IST products derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) have a reasonable accuracy and free to access, their utilization is rather limited due to the coarse spatial resolution (1km) compared to numbers of narrow leads (10-200m). The Thermal Infrared Sensor (TIRS) instrument carried on Landsat 8 satellite provides two adjacent thermal bands between 10μm and 12.5μm with a finer resolution of 100m since 2013, based on which split window (SW) algorithm can be executed to retrieve temperature. There are numerous studies retrieving land surface temperature using TIRS imagery. Given that the special weather conditions such as air temperature inversion were not taken into consideration, these algorithms may not always suitable for IST acquisition in the Arctic Ocean. The study area located in the Beaufort Sea in the Arctic and the TIRS imagery was choose on April 25, 2015. Here we checked the adaptability of following three SW methods in the Arctic for ice lead detection: the algorithm introduced by Rozenstein et al. (2014) which allows researchers to calculate their own coefficients for different atmospheric profiles; the general SW algorithm proposed by Jimenez-Munoz et al. (2014) that needs water vapor content in advance and the algorithm derived by Du et al. (2015) which needs no auxiliary data. In this study, the subarctic winter atmospheric profile provided by MODTRAN code was chosen to derive coefficients of Rozenstein’s method and the temporal closest MOD05_L2 total precipitable water vapor products was adopted as input water vapor content data for the first two algorithms. To evaluate the three IST results, we use MODIS IST products MOD29 as the validation data. Results show that Rozenstein’s method is too sensitive to stripes caused by stray light and the result image quality is comparatively poor. Jimenez-Munoz’s method tends to underestimate the IST value, which couldn’t be clearly explained yet. Conversely, Du’s method will overestimate the IST value caused by inaccurate snow emissive, which is different from the one provided by ASTER spectral library. Further improvement can be done by choosing a corrected emissive data in Du’s method. Based on the accurate IST, ice leads could be well extracted by setting an IST threshold because the temperature variation between leads and adjacent ice pixel can reach about 5K.

KEYWORDS: Leads; Ice Surface Temperature; TIRS; Split Window Algorithm

References
Exploring Spatial Heterogeneities in Real Estate Market via Geographically Weighted Summary Statistics: a Case Study in Wuhan

Binbin Lu 1*, Jiayi Xie 1, Wenzhong Shi 2, Paul Harris 3

1School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
2Department of Land Surveying and GeoInformatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
3Sustainable Soils and Grassland Systems, Rothamsted Research, North Wyke, Okehampton, UK

*Corresponding author: binbinlu@whu.edu.cn

ABSTRACT

In recent years, price fluctuations of the real estate market within China have been always drawn great attentions due to a specific phase of rapid urbanization. It could be affected by a number of hedonic characteristics, including locational attributes, structural attributes, neighbourhood attributes and other features (Chin & Chau, 2003). In this study, we employed geographically weighted summary statistics (GWSSs) (Brunsdon et al., 2002) to explore spatial heterogeneities in data relationships between house price and hedonic variables selected with a real estate data set of Wuhan (Figure 1).

![Figure 1 Real estate market data set of Wuhan](image)

In the first place, geographically weighted mean and variance are used to investigate the spatial variations of house price in Wuhan (Figure 2). This spatially varying pattern is further analysed with a number of hedonic variables, including architectural attributes (house area, floor number and construction time), location attributes (proximity to a bus or subway station, good view of a water body, accessibilities to a middle school or commercial area) and neighbourhood attributes (local population density, county-level GDP and green area per capita). The relationships between house price and each hedonic variable are investigated via geographically weighted correlation coefficient. Results show that most of the variables present spatially varying correlations, e.g. floor number of a house (Figure 3). Furthermore, different kernel functions, bandwidths and versions (basic and robust) are also tried to understand and interpret these spatially varying relationships in multiple contexts and scales.
In this study, we take GWSSs as an exploratory data analysis toolset to investigate a real estate market data set with a number of hedonic attributes. The use of GWSSs not only benefits the interpretations of its outputs, but also provides valuables insights prior to some step-forward analysis, like geographically weighted regression (Harris & Brunsdon, 2010). This attempt could form another important solution in modelling real estate market.

**KEYWORDS:** Hedonic Price Model; Spatial Non-Stationarity; Local Statistics; Spatial Statistics

**REFERENCES**


Uncertainty of endmember on Super-resolution mapping

Xinyan Li 1, Xiaodong Li #1, Yihang Zhang 1, Yun Du 1, Feng Ling 1

1 Key Laboratory of Monitoring and Estimate for Environment and Disaster of Hubei Province, Institute of Geodesy and Geophysics, Chinese Academy of Sciences, Wuhan 430077, China

#Corresponding author: lixiaodong@whigg.ac.cn

ABSTRACT

Super-resolution land cover mapping (SRM) is a process whose purpose is to determine the specific spatial distribution of different land cover classes in the mixed pixel, and to obtain a classification map with higher spatial resolution. In SRM, the problem of spectral uncertainty usually has a certain impact on the identification of SRM, and affects the accuracy of mapping. Therefore, the selection of endmembers is crucial. In current SRM methods, the same set of endmembers is used for the spectral simulation of each pixel, which ignores the spectral uncertainty, including the spectral space heterogeneity and endmember variability. By selecting the optimal endmember combination for each pixel, using which to simulate the spectrum, the problem of the spectral uncertainty can be solved to a great extent. In this research, an optimal-endmember based super-resolution land cover mapping model (OESRM) is proposed. Considering the spectral uncertainty, this research takes the minimum sum of spectral angle and spectral distance (SASD) as the criterion to select the optimal endmember combination for each pixel. By replacing the endmember of the traditional SRM model from a single endmember set to an optimal endmember combination, the effect of spectral uncertainty on SRM is greatly reduced, and the accuracy of SRM is considerably improved compared to the old methods such as LSRM and hard classification. A Sentinel-2A multispectral image taken over at Wuhan, Hubei Province, China was used to analyze the performance of the OESRM. For purposes of comparison, linear super-resolution mapping (LSRM) and hard classification were used in the same image as control experiments. In the experiment,
bands with spatial resolution of 10m and 20m were used, and the resolution of the super-resolution mapping result is 2m. The result shows that the overall accuracy increased from 68.52% for LSRM and 65.79% for hard classification to 84.78% for OESRM. The kappa value increased from 0.5521 for LSRM and 0.5119 for hard classification to 0.7768 for OESRM, indicating that the proposed OESRM can significantly reduce the effect of endmember uncertainty on the precision of super-resolution mapping.

**KEYWORDS:** Super-Resolution Land Cover Mapping; Spectral Uncertainty; Optimal Endmember Combination

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall accuracy (%)</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>OESRM</td>
<td>84.7804</td>
<td>0.7768</td>
</tr>
<tr>
<td>LSRM</td>
<td>68.5173</td>
<td>0.5521</td>
</tr>
<tr>
<td>Hard Classification</td>
<td>65.7874</td>
<td>0.5119</td>
</tr>
</tbody>
</table>

**REFERENCES**


Robust Variogram Estimation Combined with Isometric Log-ratio Transformation for Improved Accuracy of Soil Particle-sizefraction Mapping

Wenjiao Shi 1,3,* , Zong Wang 1,2,3

1Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

2State Key Laboratory of Resources and Environment Information System, Institute of Geographical Science and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

3College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

* Corresponding author at: shiwj@lreis.ac.cn

ABSTRACT

Mapping soil particle-size fractions (psf) plays an important role in regional hydrological, ecological, geological, agricultural and environmental studies. To map soil compositional data like soil psf, interpolators such as compositional kriging and the combination of log-ratio transformations with ordinary kriging or cokriging were developed. In addition, robust estimators were proposed for these interpolators to improve the variogram models. However, few studies have focused on how to choose log-ratio transformation, kriging, cokriging, or robust variogram estimation methods based on data characteristics to achieve optimal performance when mapping soil psf by comprehensive comparative analysis. Here, we selected different compositional kriging, log-ratio kriging, log-ratio cokriging and log-ratio cokriging methods combined with a robust variogram estimator to improve the accuracy of spatial predictions of soil psf when using 262 soil samples from the upper reaches of the Heihe River in China. In this study, a comprehensive comparative analysis of soil psf maps generated by using different interpolators is presented, and appropriate methods for mapping psf based on the characteristics of the available data are explored. The results show that using isometric log-ratio (ILR) transformation with different interpolators can achieve relatively better performance than the other log-ratio transformation methods. In addition, combining the interpolators with robust variogram estimators significantly improve the prediction accuracy compared with using standard estimators, which presented reasonable and smooth transitions when mapping soil psf. Combining ILR cokriging with a robust variogram estimator had the best accuracy, with the lowest root mean squared error (sand, 10.50%; silt, 11.24%; clay, 7.32%), an Aitchison’s distance of 0.76, a standardized residual sum of squares of 0.70 and a relatively higher rate of correctly predicting soil texture types 90.04%. In the future, guideline for using log-ratio transformation methods with linear regression, a generalized linear model or random forest should be developed and combined with ancillary variables to improve the interpolators.

KEYWORDS: Soil Particle-Size Fractions; Spatial Interpolation; Log-Ratio Transformation; Log-Ratio Cokriging; Robust Estimator
Mapping Soil Particle-Size Fractions: A Comparison Of Compositional Kriging and Log-ratio Kriging

Zong Wang\textsuperscript{1,2}, Wenjiao Shi\textsuperscript{1,3,*}

\textsuperscript{1}Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

\textsuperscript{2}School of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China

\textsuperscript{3}State Key Laboratory of Remote Sensing Science, School of Geography, Beijing Normal University, Beijing 100875, China

*Corresponding Authors: shiwj@lreis.ac.cn

ABSTRACT

Soil particle-size fractions (psf) as basic physical variables need to be accurately predicted for regional hydrological, ecological, geological, agricultural and environmental studies frequently. Some methods had been proposed to interpolate the spatial distributions of soil psf, but the performance of compositional kriging and different log-ratio kriging methods is still unclear. Four log-ratio transformations, including additive log-ratio (alr), centered log-ratio (clr), isometric log-ratio (ilr), and symmetry log-ratio (slr), combined with ordinary kriging (log-ratio kriging: alr\_OK, clr\_OK, ilr\_OK and slr\_OK) were selected to be compared with compositional kriging (CK) for the spatial prediction of soil psf in Tianlaochi of Heihe River Basin, China. Root mean squared error (RMSE), Aitchison’s distance (AD), standardized residual sum of squares (STRESS) and right ratio of the predicted soil texture types (RR) were chosen to evaluate the accuracy for different interpolators. The results showed that CK had a better accuracy than the four log-ratio kriging methods. The RMSE (sand, 9.27%; silt, 7.67%; clay, 4.17%), AD (0.45), STRESS (0.60) of CK were the lowest and the RR (58.65%) was the highest in the five interpolators. The clr\_OK achieved relatively better performance than the other log-ratio kriging methods. In addition, CK presented reasonable and smooth transition on mapping soil psf according to the environmental factors. The study gives insights for mapping soil psf accurately by comparing different methods for compositional data interpolation. Further researches of methods combined with ancillary variables are needed to be implemented to improve the interpolation performance.

KEYWORDS: Compositional Data; Spatial Interpolation; Compositional Kriging; Log-Ratio
A Modified Change Vector Approach for Land Use Change Mapping in the Loess Plateau, China, Supporting Critical Zone Science

Alexis Comber 1*, Yihe Lu 2, Ying Luo 2, Yanjiao Ren 2, Paul Harris 3

1School of Geography, University of Leeds, UK
2State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Science, China
3Rothamsted Research, North Wyke, Okehampton, Devon, EX20 2SB, UK

*Corresponding author: a.comber@leeds.ac.uk

ABSTRACT

This paper describes the creation of annual land cover / land use maps for the Loess Plateau region, China. Land use data are needed to support analysis of the coupled relationships between soil-water processes, multi-scale agro-ecosystem services with the effects of vegetation cover and changes in land use. While medium scale LCLU data for China exist for 2000, 2005 and 2010, temporally sensitive data are needed to evaluate the impacts on landscape re-greening, restoration agricultural sustainability initiatives (e.g., ‘Three Norths Shelter Forest System Project’, ‘Natural Forest Conservation Program’, ‘Grain to Green Program’). Historical land use data, and critically data on land use change, for the Loess Plateau region was created using MODIS EVI data, extending the methods described in Tsutsumida et al. (2016).

KEYWORDS: Land Use / Land Cover; LULC Change; Change Vector; Critical Zone Science

I TABLES AND FIGURES

Reference data was collected by randomly sampling the MODIS pixels and the proportions of 6 different classes (Woodland, Grassland, Wetland, Farmland, Artificial surface, Barren) was recorded at each location. A total of 8353 locations were sampled (Figure 1) and the mix of land covers recorded using historical very high resolution aerial imagery in Google Earth (Figure 2). The distributions of the samples for each class are shown in Table 1. These were used as inputs to a random forest classifier.
**Figure 1:** The 8335 data points in the Loess Plateau region with an OpenStreetMap backdrop.

**Figure 2:** An example of the Google Earth historical imagery and the grid used to sample the land use class proportions.

Land use layers for each year were constructed and a land cover / land use change (LCLUC) analysis was undertaken using a modified change vector approach (Xu et al, in submission). This explicitly seeks to overcome the problems with post classification change analysis, (ie methods that compare the annual land use maps to derive change statistics), as articulated by Fuller et al (2003). Here, the reference data characteristics in bi-temporal image space were used to construct change models that predict land use changes directly from bi-temporal image data and measures derived from the change vector inner product, without the need to determine the individual vectors at either Time 1 (2013) or Time 2 (2017) relative to each axis in a multidimensional (image) feature space.

We describe and discuss the results, in the context of the objectives of the supporting critical zone science in the Loess Plateau. These include upscaling measures of a) reduced runoff and sediment loss (the main aim of the wider project of which this analysis is a part); b) coupled field scale land use change scaled up to catchment scale runoff risks; c) identifying the contribution of individual land use changes to water balance and soil losses in different regions of the Loess Plateau; and d) to model different land use configurations to support optimal ecosystem services provision. The change methodology is evaluated in the context of the issues related multitemporal land cover / land use change detection using spatiotemporal image data cubes (White et al, 2017; Zhu, and Woodcock 2014) and the further extension of the methods described here with soft classifications such as fuzzy sets.

**Table 1.** Reference data collected for each class for each year.

<table>
<thead>
<tr>
<th>Year</th>
<th>Forest</th>
<th>Grass</th>
<th>Wetland</th>
<th>Farm</th>
<th>Artificial</th>
<th>Barren</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>41</td>
<td>14</td>
<td>0</td>
<td>64</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2001</td>
<td>10</td>
<td>13</td>
<td>0</td>
<td>6</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>2002</td>
<td>61</td>
<td>72</td>
<td>3</td>
<td>141</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>2003</td>
<td>47</td>
<td>68</td>
<td>4</td>
<td>115</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>2004</td>
<td>23</td>
<td>92</td>
<td>1</td>
<td>74</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>2005</td>
<td>45</td>
<td>77</td>
<td>4</td>
<td>107</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>2006</td>
<td>45</td>
<td>86</td>
<td>4</td>
<td>104</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>2007</td>
<td>35</td>
<td>97</td>
<td>2</td>
<td>44</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>2008</td>
<td>170</td>
<td>134</td>
<td>3</td>
<td>155</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>2009</td>
<td>127</td>
<td>164</td>
<td>8</td>
<td>112</td>
<td>29</td>
<td>17</td>
</tr>
<tr>
<td>2010</td>
<td>199</td>
<td>175</td>
<td>8</td>
<td>208</td>
<td>23</td>
<td>39</td>
</tr>
<tr>
<td>2011</td>
<td>116</td>
<td>119</td>
<td>7</td>
<td>192</td>
<td>33</td>
<td>41</td>
</tr>
<tr>
<td>2012</td>
<td>176</td>
<td>136</td>
<td>15</td>
<td>171</td>
<td>29</td>
<td>37</td>
</tr>
<tr>
<td>2013</td>
<td>391</td>
<td>277</td>
<td>10</td>
<td>319</td>
<td>49</td>
<td>50</td>
</tr>
<tr>
<td>2014</td>
<td>350</td>
<td>175</td>
<td>13</td>
<td>291</td>
<td>45</td>
<td>26</td>
</tr>
<tr>
<td>2015</td>
<td>284</td>
<td>208</td>
<td>11</td>
<td>274</td>
<td>55</td>
<td>22</td>
</tr>
<tr>
<td>2016</td>
<td>243</td>
<td>183</td>
<td>10</td>
<td>287</td>
<td>57</td>
<td>27</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

This research was supported by the National Key Research and Development Program of China (No. 2016YFC0501601) and the China-UK bilateral collaborative research on critical zone science (National Natural Science Foundation of China NO. 41571130083 and the Natural Environment Research Council Newton Fund NE/N007433/1).

References


Daily Sea Ice Lead Retrieval from MODIS Images and the Uncertainty Handling in Cloud Mask

Meng Qu 1*, Xi Zhao 1*, Xiaoping Pang 1

1Chinese Antarctic Center of Surveying and Mapping, Wuhan University, China

*Corresponding author: mango@whu.edu.cn(M.Q.); xi.zhao@whu.edu.cn(X.Z.)

ABSTRACT

Sea ice lead, defined as linear feature of open water or thin ice within pack ice, serves as prime window for heat exchange between ocean and atmosphere. Based on the significant temperature difference between leads and surrounding thick ice, previous studies have explored the use of MODIS ice surface temperature (IST) product (i.e. MOD29) to retrieve lead distribution in the Arctic on daily scale. Since the intense heat exchange in lead area might produce ocean smoke or fog, which tend to be detected as low water cloud by many infrared brightness temperature difference (BTD) algorithms, a great part of lead information is lost in conservative cloud masking, causing underestimation of lead occurrence and distribution.

In this study, we propose a modified cloud mask to reduce the uncertainty in the retrieval of daily leads in the Beaufort Sea from MODIS level 1B images product (i.e. MOD02). First of all, IST was calculated from MOD02 11μm and 12μm thermal bands using a split window algorithm. Since the surface temperature difference between leads and pack ice is expected to be amplified in sunlight, only daytime images were used. Channel Bit18 (BTD11μm-12μm) from MODIS cloud mask product (i.e. MOD35) was modified by suppressing its sensitivity to high temperature water cloud and applied to remove pixels contaminated by thick cloud. A high pass filter and threshold segmentation technique is then applied to the masked IST images to classify sea ice and leads into binary lead map. Available lead maps during the same day were incorporated to a daily lead map that was calculated as average lead frequency. Pixels with lead frequency below 0.5 were labeled as false lead, so-called cloud artifact, which resulted mainly from deficiency of cloud mask procedure.

Validation with corresponding Landsat8 and Radarsat-2 images shows the daily lead map captures more lead structure than published maps using MOD29, while cloud effect is efficiently controlled and uncertainty from cloud mask is reduced. Dominated by a north-south linear pattern, the leads system spreads southeastwards from north of the Barrow point to Amundsen Gulf and covers up to 7% of the Beaufort Sea area in April. Fragments of pack ice scatter among the dense leads network. Persistent leads on the outer edge of fast ice along Canadian coast evolves to coastal polynya as sea ice retreats further away. The new approach provides necessary supplement to previous thermal method, and the daily lead map reveals a realistic lead distribution in the Beaufort Sea. Interannual variation of lead occurrence and evolvement in this area requires further research and validation.

KEYWORDS : Digital Soil Mapping; Mean Squared Error; Kriging Variance; Simulated Annealing; Design-Based Estimation; Probability Sampling; Pedometrics
Adopting Semantic Signatures to Quantify the Uncertainty of Volunteered Geographic Information

Rui Zhu 1* and Krzysztof Janowicz 1

1 STKO Lab, Department of Geography, University California, Santa Barbara, USA

*Corresponding author: ruizhu@geog.ucsb.edu

ABSTRACT

With the increasing popularity of volunteered geographic information (VGI), the quality of geospatial data currently attracts more attentions compared to the traditional issue of coverage in the GIScience community. To understand the uncertainty, or accuracy, of these geospatial data, many aspects have been applied, such as uncertainties in positions, attributes, and topological relations. However, there is few research on assessing the uncertainty, or accuracy, of geospatial information from a semantic perspective, which plays an increasingly vital role in an era where place-based GIS emerges to corporate with space-based GIS. To address such as a gap, this work proposes an approach to take into account the semantic uncertainties of places to understand the quality of VGI. The semantics are specifically extracted from the feature type of geographic information (e.g., place types such as Restaurant, Hotel and Police Office) as it is semantically rich in terms of characterizing the functionality, popularity, and human perception of places. In order to assess the semantic uncertainty of feature types, which are categorical data compared to numeric values applied in most conventional uncertainty analysis, we design a set of statistics, named as semantic signatures, to quantitatively represent feature types. Semantic signatures are grouped into three categories: spatial, temporal and thematic, giving the respective perspectives they are extracted from. For example, spatial signatures are designed to capture the spatial structure of instances belonging to a specific feature type. A series of spatial statistics, including spatial point pattern analysis and spatial autocorrelation analysis are investigated to quantify the spatial structure. Having these statistics, we then evaluate the uncertainty of feature types by using traditional techniques of uncertainty analysis (e.g., variance and entropy). Our goal of this work is to gain the ability of answering the kind of questions like: what is the variability (i.e., uncertainty) of a place’s neighborhood knowing only its place type? For instance, the neighborhood of a restaurant will be more uncertain than the neighborhood of a school. By answering these questions, our work has the potential to improve the accuracy of VGI by (1). recommending place types for users; (2). cleaning up the data; (3). aligning place place across different geospatial data sources. We test our approach using a dataset of POIs collected from Foursquare, Google Place and Factual in the area of Maryland, USA.

KEYWORDS: VGI, Semantic Signatures, Spatial Statistics, Semantic Uncertainty, Social Networks

REFERENCES


Remote Sensing Accuracy and Resolution for Precision Agriculture in Small Farms, Malawi

Shengpan Lin*, Brad Peter 1, Joseph Messina 1

1 Michigan State University, United Stated

*Corresponding author: shengpan@msu.edu

ABSTRACT

Background: Increasing spatial and temporal resolutions of remote sensing (RS) imagery on one hand provides more information for precision agriculture (PA), on the other hand also increases the investment in remote sensing data. Agricultural decision-making systems are being built in Malawi to provide small-farm holders guidelines on agricultural practices such as fertilization. Local and international stakeholders are not clear about the investment return rates of existing and newly available RS data, including private microsatellite imageries and high resolution drone data.

Objective/Method: The goal of this study was to decide what RS products were suitable for the agricultural decision-making systems in Malawi. For example, to address the farmers’ concern on whether and how much they should fertilize regarding current crop growth, we related RS signals from different RS sources (i.e. drone, PlanetScope, Sentinel-2 A/B, and Landsat-8) to ground agricultural practices (i.e. different fertilization levels), investigating how well nutrient stress could be indicated by RS NDVI from different sources of data with different spatial and temporal resolutions. The products that can inform the crop nutrient stress while with the lowest cost was the optimal products.

Results/Discussion: Different levels of fertilization in small farms were successfully discriminated by NDVI derived from drone data. The discrimination capability decreased when spatial resolution of the drone data decreased from 6 cm to 30 m. Imagery from Drone (0.06 m GSD), PlanetScope (3 m GSD), Sentinel-2 (10 m GSD), and Landsat-8 (30 m GSD) had different costs and accuracies (signal-to-noise ratios). They differentiated different levels of fertilization in small farms. Daily PlanetScope imagery potentially provided more timely information than Sentinal-2 (8-day revisit time) and Landsat-8 (16-day revisit time), but the cost was much higher. Drone data, with less atmospheric effects and higher accuracy compared to other data, were ideal for accuracy but too costly for practice. The combination of Landsat-8 and Sentinel-2 A/B may be more practical, considering less than 3 days of combination frequency, free of change, and acceptable accuracy.

KEYWORDS: Precision Agriculture; Remote Sensing; Small Farms; Accuracy; Resolution
Improving the Prediction Accuracy of Forest Aboveground Biomass Benchmark Map by Integrating Machine Learning and Spatial Statistics

Shaoqing Dai 1,2, Xiaoman Zheng 1,2, Shudi Zuo 1,2,3, Yin Ren *1,3

1 Institute of Urban Environment, Chinese Academy of Sciences, China
2 University of Chinese Academy of Sciences, China
3 Ningbo Urban Environment Observation and Research Station-NUEORS, Chinese Academy of Sciences, China

*Corresponding author: yren@iue.ac.cn

ABSTRACT

We present a low-cost method to create high-precision, spatially explicit reference maps of large-scale forest aboveground biomass (AGB) to provide a scientific basis for quantitative assessment of forest management decisions involving, for example, forest-destroying carbon emissions. However, due to the limited availability of the sample lands and other space restrictions, the prediction accuracy of current AGB reference maps is associated with deviations, whose source may be related to the poorly understood effects of the spatial heterogeneity of multiple environmental factors (such as topography, soil, and forest structures) on the spatial distribution of AGB. To address these problems, we propose a method that combines machine learning and spatial statistics to determine the influence of biased samples by finding multiple environmental factors of regional covariates through the model algorithm. Subsequently, we consider an empirical case to quantitatively assess the prediction accuracy of the AGB reference map of an Eucalyptus plantation in Nanjing, China. In this case study, the biased AGB analysis data of 90 parse trees are selected and combined with regional forest resource inventory data to build three main machine-learning methods (support vector machine, random forest, and artificial neural network) and a spatial statistical analysis integration technology (a PBSHDE model). The results show that the evaluation indexes of the prediction accuracy (i.e., RMSE, MAE, and MRE) of the method that combines machine learning (random forest) and the PBSHDE model are significantly better than those for machine learning or the PBSHDE model individually. We conclude that the combination of spatial statistical analysis and machine learning can improve the accuracy AGB mapping in regions with biased samples and thus provide accurate predictions from AGB reference maps. Our research methods and conclusions may provide references for AGB remote-sensing mapping and simulation of ecological processes of different types of forests in various countries and tropical regions of the world.

KEYWORDS: AGB, Biased samples Machine learning, Reference map, Spatial statistical analysis

TABLES AND FIGURES

Table 1. Spatial autocorrelation and heterogeneity test

<table>
<thead>
<tr>
<th>Spatial autocorrelation</th>
<th>Spatial heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items</td>
<td>Values</td>
</tr>
<tr>
<td>Moran I</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>z-score</td>
<td>16.22</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1: Venn diagram showing a typical and conceptual framework to estimate AGB (Mg/ha) in Eucalyptus plantations. Schematic of proposed methodology integrating machine learning and spatial statistical analysis for clarifying the relationship between forest patches and reference series. Field measurements of parse trees can be combined with a localization biomass model, machine learning, and spatial statistical analysis to predict regional AGB benchmark maps. These predictions can be summed over 90 individual parse trees to obtain a region-level AGB. (a), (b), and (c) show ANN-RBF, RF and SVM, respectively, and (d) shows a localization biomass model.
Figure 2: Results for Eucalyptus plantation test site in Nanjing County, Fujian, China. The bootstrapped bean plots illustrate the distribution of the mean MAE(a), MRE(b), and RMSE(c). Boxplots represent MAE(a) and MRE(b) for each prediction method, with the median (black line in the box), inter-quartile range (25%–75% in the box), the range 5%–95% (whiskers) and outliers (asteroids) labeled. The histogram distributions of RMSE for each prediction method are calculated. The color-shaded curves represent the distribution of the values. The mean MAE(a), MRE(b), and RMSE(c) values obtained by the leave-out cross validation for each prediction method are also shown (S1=SVM = support vector machines, S2=ANN-RBF = artificial neural network radial-basis function, S3=RF = random forest, S4=P-BSHDE , S5=SVM&P-BSHDE, S6=ANN-RBF&P-BSHDE, S7=RF&P-BSHDE, ML=machine learning , Sp=Spatial statistic).

II. FORMULA of P-BSHADE methods

a. Objective

The objective is to interpolate the AGB data set based on data acquired from other forest patches. A theoretical description and derivation of the formula are expressed as follows:

\[ \hat{y}_j = \sum_{i=1}^{n} w_{ij} y_i \]  

(1)

where \(w_i\) denotes the weight (contribution) of the \(i\)th observed parse tree datum to the \(j\)th forest patch to be interpolated, \(\hat{y}_j\) is the AGB of the \(j\)th forest patch, and \(y_i\) is the AGB of the \(i\)th parse tree datum. As expected, the two properties of the estimate of Eq.(1) are unbiased,

\[ E(y_i) = E(\hat{y}_j) \]  

(2)

and minimum estimation variance,

\[ \min_w \left[ \sigma^2_{\hat{y}_j} = E(\hat{y}_j - y_j)^2 \right] \]  

(3)

where \(E\) is the statistical expectation.

b. Ratio between forest patch and parse tree data
The ratio between forest patch and parse tree data is one of the most important inputs for estimating forest patch AGB and is an index of heterogeneity in the AGB spatial distribution. The relationship between forest patch and parse tree data can be expressed as

\[ b_{ij} E y_i = E y_i \]  \hspace{1cm} (4)

where \( b_{ij} \) is the ratio between AGB at the parse tree data and forest patch. Considering Eq. (1), Eq.(4) can be written as

\[ \sum_{i=1}^{n} w_{ij} y_i = 1 \]  \hspace{1cm} (5)

This equation is generally valid for nonhomogeneous conditions. Clearly, the determination of \( y_j \) requires the coefficients \( w_i (i=1,\ldots,n) \) to be calculated, which is addressed in the following section.

c. Estimation of weight

The estimation problem is to find the weights \( w_{ij} \) that satisfy the unbiased condition and that minimize the estimation variance:

\[ \sigma_{y_j}^2 = E(\hat{y}_j - y_j)^2 = c(\hat{y}_j, \hat{y}_j) + c(\hat{y}_j, \hat{y}_j) - 2c(\hat{y}_j, \hat{y}_j) \]  \hspace{1cm} (6)

These weights can be calculated by minimizing the estimation variance and taking into account the unbiasedness,

\[
\begin{bmatrix}
    c(y_1y_1) & \cdots & c(y_1y_n) & b_1 \\
    \vdots & \ddots & \vdots & \vdots \\
    c(y_ny_1) & \cdots & c(y_ny_n) & b_n \\
    b_1 & \cdots & b_n & 0 & \mu
\end{bmatrix} \begin{bmatrix}
    w_1 \\
    \vdots \\
    w_n
\end{bmatrix} = \begin{bmatrix}
    c(y_1y_1) \\
    \vdots \\
    c(y_ny_n)
\end{bmatrix}
\]

\[ \mu \] is a Lagrange multiplier. The minimized variance in the estimation error can then be written as

\[ \sigma_{y}^2 = \sigma_{\hat{y}_j}^2 + \sum_{i=1}^{n} \sum_{k=1}^{n} c(y_iy_k) - 2 \sum_{i=1}^{n} w_{ij} c(y_iy_j) + 2\mu(\sum_{i=1}^{n} w_{ij} b_{ij} - 1) \]  \hspace{1cm} (8)

References: Xu et al. used the point estimation model of Biased Sentinel Hospitals-based Area Disease Estimation (P-BSHADE) to interpolate missing data in temperature datasets in Xu et al. (2013).

REFERENCES

Maximum Consistency Block based Image Segmentation Methods

Hexiang Bai\textsuperscript{1}, Feng Cao\textsuperscript{1}

\textsuperscript{1}School of Computer and Information Technology, Shanxi University, Taiyuan, 030006 Shanxi, China

*Corresponding author: baihx@sxu.edu.cn

ABSTRACT
Segmentation is an important step for Object-based Image Analyzing of high resolution remotely sensed images. Currently, there are many different ways for decomposing the image into small patches, such as edge detection, region merging, and region growing. However, these methods all generate segmentation results which are partitions of the remotely sensed imagery. This neglects the possible that two objects may share pixels in some situations. This paper proposed a cover-based segmentation methods using spatial extended maximum consistency block. This method divides the remotely sensed image into intersected small parts and is more appropriate for the situation that different categories mixed with each other.

KEYWORDS: Pedometrics segmentation; partition; cover; maximum consistency block

I INTRODUCTION
Segmentation is an important step for Object-based Image Analysis of high resolution remotely sensed images. Currently, there are many different ways for decomposing the image into small patches, such as edge detection, region merging, and region growing (Zhang et. al., 2014, Lu et. al., 2016, Yang et. al., 2017). These methods all divided the remotely sensed image into disjoint small parts. That is all segmentation results are partitions of the remotely sensed imagery. However, different categories may mixed with each other in real life applications. It is impossible to crisply define the boundary for different categories. Accordingly, it is natural to use a cover rather than a partition of the image to form its segmentation. That is two objects in the segmentation may share some pixels.

Maximum consistency block (Leung and Li, 2003) is an effective tool in constructing a cover of the universe of discourse in terms of the similarity relations. A remotely sensed image can be looked upon as a set of pixels, these pixels constitute the universe of discourse, and the cover-based segmentation is to construct a cover of these pixels according to the similarity of the universe. However, although maximum consistency block is shown effective, it did not take spatial adjacency of pixels into account. To overcome this deficiency, it is necessary to extend maximum consistency block to spatial data sets.

II METHODS
Suppose that two pixels \( p_1 \) and \( p_2 \) are similar to each other according to their features (e.g., spectral or texture values). Denote this relation as \( p_1 \approx p_2 \). If relation \( \approx \) matches that \( p_1 \approx p_1 \) and \( p_1 \approx p_2 \Leftrightarrow p_2 \approx p_1 \), then \( \approx \) is called a tolerance relation. It is easy to construct a tolerance relation for two pixels according to application needs. For example, two pixels are similar to each other if their distance in the spectral space is less than a predefined threshold.
Most region merging used similarity measure to construct a partition segmentation of a remotely sensed image. These methods attempted to find the partition which elements have little within-class difference and large difference between-class. However, the most natural way of using similarity between pixels is to construct a cover instead of partition of a remote sensing image. Generally, the first step of build a cover using a tolerance relation is consistent blocks.

**Definition 1.** Suppose that \( \approx \) is a tolerance relation of \( R \), \( C \) is a set of pixels in \( R \). If any two elements \( a_1, a_2 \in C \) satisfies that \( a_1 \approx a_2 \), \( C \) is a consistent block generated by tolerance relation \( \approx \).

The set of consistent blocks that contain \( R \) constitutes one of its cover. However, there is no guarantee that all pixels in a consistent block are spatially connected with each other. This will generate unreasonable segmentation results. Accordingly, it is important to keep each elements in the constructed cover should also be connected in space in segmenting a remotely sensed image.

**Definition 2.** In a remotely sensed image \( R \), a block \( C \) is a set of pixels. For any \( p_1, p_2 \in C \), if there exists a sequence \( u_1', u_2', \ldots, u_n' \) in \( C \) such that \( p_1 \) is adjacent to \( u_1' \) and \( u_n' \) is adjacent to \( p_2 \), \( C \) is a spatially connected block in \( R \).

**Definition 3.** Suppose that \( \approx \) is a tolerance relation of \( R \), \( C \) is a spatially connected in \( R \). If any two elements \( a_1, a_2 \in C \) satisfies that \( a_1 \approx a_2 \), \( C \) is a spatially connected consistent block generated by tolerance relation \( \approx \).

Using spatially connected consistent blocks could generate a cover segmentation of a remotely sensed image. However, there may be many different covers of a remotely sensed imagery. Another question is which cover is the best. Leung and Li (2003) showed that maximum consistent block could generate a unique cover of the universe and has many advantages to other methods. Combining the concept spatially connected consistent block and maximum consistent block could generate a unique cover which elements are spatially connected.

**Definition 4.** Suppose that \( \approx \) is a tolerance relation of \( R \), \( C \) is a spatially connected consistent block generated by tolerance relation \( \approx \). If for any spatially connected consistent block \( C' \) generated by tolerance relation \( \approx \), \( C' \subseteq C \), then \( C \) is a maximum spatially connected consistent block in \( R \).

Given a remotely sensed image \( R \), each pixel/super-pixel \( p \in R \) can generated a maximum spatially connected consistent block \( C_p \). The set of all maximum spatially connected consistent blocks generated by all pixels \( (C_{\text{m}(R)} = \{C_p \mid p \in R\}) \) formed a cover of \( R \). Furthermore, according to Definition 4, it is easy to prove that any tolerance relation \( \approx \) can uniquely determine a \( C_{\text{m}(R)} = \{C_p \mid p \in R\} \) for \( R \) (Leung and Li, 2013). Based on the concept of maximum spatially connected consistent blocks, a new cover based segmentation method is proposed. The algorithm calculates each pixels’ maximum connected consistency block. Because there may exists some blocks are completely covered by other several blocks, the algorithm also removed all the redundant blocks which did not has a pixel when they minus all neighbouring blocks. The flowchart of the algorithm is shown in Figure 1.
III EXPERIMENTS

An illustrative example is given to show the effectiveness of spatial maximum consistency block based segmentation works. Figure 2 is the Panchromatic band of a Quickbird remotely sensed image. The spatial resolution of the panchromatic band was 0.63m. The size of the image was 100 pixels x 100 pixels. During the segmentation, only the panchromatic band is used. The similarity between pixels are measured using euclidean distance of pixel’s spectral value. Two objects are similar to each other when their similarity is less than a given threshold. The threshold are set to 100, 200, and 300, respectively. The segmentation results corresponds to the these thresholds are shown in Figure 3.
Figure 2: The study area.

Figure 3: The segmentation results corresponds to the threshold 100, 200, and 300, respectively.

From Figure 3, it can be found that the larger the threshold, the larger the region of each segmentation unit. Meanwhile, the segmentation unit intersects with each other. This is hard to be shown in the figure. In fact, there may exists many different possible sub-cover of the remotely sensed imagery. A difficult task is how to select the most appropriate cover of the remotely sensed image as the segmentation result.

IV CONCLUSION

This paper proposed using a cover instead of a partition of the remotely sensed image as the segmentation for further analysis. This method could be used as an alternative to address the issue of inevitable intersection between categories. Now, there exists more problems related to this type of methods, such as selecting best cover and assessing of segmentation results. However, this an interesting and valuable approach to explore.

ACKNOWLEDGEMENTS

This work is supported by the Natural Science Foundation of Shanxi Province, China (No. 201701D121055).

References
High-resolution Mapping of Direct CO2 Emissions and Uncertainties at the Urban Scale

Shaoqing Dai 1,2, Shudi Zuo 1, Yin Ren*1,3

1 Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy
2 University of Chinese Academy of Sciences, China
3 Ningbo Urban Environment Observation and Research Station-NUEORS, Chinese Academy of Sciences, Ningbo, 315800, China

*Corresponding author: yren@iue.ac.cn

ABSTRACT

Mapping direct carbon emissions at high-resolution in urban environments could help in the development of measures to mitigate carbon emissions through optimizing the layout of inner structures. It requires the use of a mapping method combining the bottom-up and top-bottom calculations with uncertainty evaluations. This study developed a method for urban scale analyses of carbon emissions, including a theoretical framework of uncertainty distribution and transmission. Using Jinjiang City, China, as a case study, we applied this method to calculate the amount of carbon emissions in grids distributed across a city. This information was used to analyze emission uncertainties and its sources. The calculated emissions were allocated through the accurate spatial identification of three emission sectors and proxy data. Two different population spatialization methods were constructed in order to create 30 m and 500 m resolution grid maps. We designed four different Monte-Carlo simulation scenarios to analyze the uncertainties of the two maps. The results showed that the method developed here was suitable for delineating carbon emissions at the urban-scale. The 30 m resolution map showed that residential emissions were widely distributed, whereas industrial emissions were more concentrated, with the opposite trend being detected in the 500 m resolution map. Calculations of carbon inventory and spatial proxy had more impacts on the 30 m resolution map than on the 500 m resolution map. During the process of spatial superposition, the uncertainties from different sectors showed a nonlinear relationship, which was represented by smaller total uncertainties compared with the sum of uncertainties from the three emission sectors. In conclusion, this study provides important baseline data that could be used to optimize urban form by promoting low-carbon city construction.

KEYWORDS: Monte-Carlo simulation; spatial proxy; carbon tabular inventory; uncertainty propagation; bottom-up and top-bottom

I. TABLES AND FIGURES
Figure 1: Map showing CO2 emissions at 30 m and 500 m resolution. The upper figures are at 30 m resolution, the lower figures are at 500 m resolution. (a) and (A) are the total direct CO2 emissions maps, (b) and (B) are the residential CO2 emissions maps, (c) and (C) are the industrial CO2 emissions maps, and (d) and (D) are the traffic CO2 emissions maps.

Figure 2: Map showing the uncertainty in activity levels. (a) and (b) represent the uncertainty maps of total CO2 emissions at 30 m and 500 m resolution, respectively.
II. FORMULA

In terms of the IPCC accounting methodology, we calculated direct carbon emissions (Scope 1), from which we created two high-resolution CO2 emission maps of different scales (30 m and 500 m) by combining spatial models with top-bottom and bottom-up approaches. Based on Monte Carlo simulations, four different schemes were designed to study the uncertainty, including the overall uncertainty of carbon emissions from different sectors (scheme 1 and 2) and how uncertainties influence different procedures of grid mapping and the superimposed delivery process (scheme 3 and 4).

\[
\text{Grid}_{i, \text{CO}_2} = f(c_l, \text{weight}_i) \tag{1}
\]

Where Grid\(_{i, \text{CO}_2}\) is the CO2 emission value on the \(i\) th grid (\(i = 1, 2, 3, ..., n\)), \(c_l\) represents the total amount of CO2 emissions from the different CO2 emission sources (\(l = \text{res, ind, trans}\)), Weight\(_{i, l}\) is the weight of the \(l\) type of emissions on grid \(i\).

\[
\text{Uncertainty} = \frac{Cl_{95}}{\overline{CO}_2}, Cl_{95} = \frac{CO_{2.975}-CO_{2.25}}{2} \tag{2}
\]

where CO2,97.5 is the 97.5% quantile of CO2 emissions in the Monte-Carlo simulation, CO2,2.5 is the 2.5% quantile of CO2 emissions in the Monte-Carlo simulation, and \(\overline{CO}_2\) is the mean value of direct CO2 emissions from the Monte Carlo Simulation.

\[
\text{error} = \frac{\sum_{i=1}^{n} \text{Difference}_{i, l}}{c_l} = \frac{\sum_{i=1}^{n} \vert \text{real value}_{i, l} - \text{simulation}_{i, l} \vert}{c_l} \times 100\% \tag{3}
\]

where Difference\(_{i, l}\) is the absolute value of the difference between the real pixel and simulation value emission from \(l\) th sector on the \(i\) th grid (\(l = \text{res, ind, trans}\), which represents the residential, industrial, transport and total emissions). CI is total CO2 emissions from \(l\) th sector.
Comparison between the Accuracy of SMOS Sea Ice Thickness Products and the Thin Ice Thickness Retrieved from AMSR2 Based on Ship Observation Data by Sikuliaq

Xueyi Ma *1, Xi Zhao 1, Meng Qu 1, Zian Cheng 1

1 Wuhan University, China

*Corresponding author: maxueyi9517@whu.edu.cn

ABSTRACT

Thin ice (0-0.2m) plays an important role for heat exchange between the ocean and the atmosphere, which affects the local climate and deep water circulation in the Arctic. The SMOS (Soil Moisture and Ocean Salinity) SIT daily maps are newly released products which provide accurate thin ice thickness (SIT). The brightness temperature data from Advanced Microwave Scanning Radiometer 2 (AMSR2) is another data source to retrieve thin ice thickness.

The objectives of the present work are (i) to validate the accuracy of SIT-SMOS from ESA’s SMOS mission; (ii) to explore relationship between SIT from ship observation and the brightness temperature from AMSR2, and to retrieve SIT-AMSR2 (iii) to compare the difference between the SIT-SMOS and SIT-AMSR2. In this study, we use Sikuliaq cruise ship observation data in Beaufort Sea during 2015 winter as validation references.

The SMOS SIT daily maps were downloaded from the websites of Hamburg University (SSITH for short) and Bremen University (SSITB for short), Germany, respectively. The accuracy of SSITH and SSITB were validated based on sea ice record observed aboard on Sikuliaq (SIT-OBS). Results of our study show that SSITH and SSITB have a strong correlation with a correlation coefficient(R) of 0.899, the RMS error is 0.0373 and the average bias is -0.0018. When compared to SIT-OBS at 270 sample points, the SSITH has R of 0.577 (0.585 for SSITB), RMSE of 0.10657(0.10599 for SSITB) and average bias of -0.03623m (-0.03462m for SSITB). Therefore, we found that the bias of SSITB and SSITH are both less than 0.04m, and the accuracy of SSITB is higher than SSITH.

On the other hand, we used the algorithm developed by Katsushi Iwamoto (2014) to detect thin ice by combining brightness temperature at 36GHz and 89GHz from AMSR2. The sea ice concentration daily maps from University of Bremen were used to exclude the influence of open water (sea ice concentration <30%). The fast ice was masked. And then 93 valid SIT-OBS test points were selected. The comparison between SIT-AMSR2 and SIT-OBS shows an obvious correlation with R of 0.288, RMSE of 0.05856 and average bias of -0.01839. We also compared SSITH and SSITB to the 93 test points and concluded that the accuracy of SSITB is higher than SSITH with R of 0.603(0.536 for SSITH), RMSE of 0.05489(0.05824 for SSITH) and average bias of -0.0051m (-0.01419m for SSITH). We found that the RMSE difference between the thin ice thickness retrieved from AMSR2 and SSITH is 0.0032, and the accuracy of SSITB is higher than SIT-AMSR2. To conclude, SIT-AMSR2 and SIT-SMOS both have reasonable accuracy on thin ice thickness, but the accuracy of SIT-SMOS is higher than SIT-AMSR2. Between the two SIT-SMOS products, the SSITB has a relative higher accuracy than SSITH. This conclusion is drawn based on a limited number of ship observations in Beaufort Sea. More in situ thin ice thickness measurements are preferred to make a general conclusion.
KEYWORDS: ship observation data; SMOS daily SIT products; AMSR2 SIT

REFERENCES
Estimation of Soil Moisture with Land Surface Temperature Retrieving from Landsat-8 Thermal Infrared Data Using Different Algorithms

Yuefeng Yao ¹*, Li Li ¹

¹ Guangxi Key Laboratory of Plant Conservation and Restoration Ecology in Karst Terrain, Guangxi Institute of Botany, Guangxi Zhuang Autonomous Region and Chinese Academy of Sciences, Guilin 541006, China

*Corresponding author: yf.yao@gxib.cn / yf.yao@outlook.com

ABSTRACT

Rapid and accuracy estimation of soil moisture is great significance for precision agriculture, also for water resources sustainable management, especially in typical tourism regions and degradation karst ecosystem. Land surface temperature (LST), which can be retrieved from high to medium spatial resolution remote sensing data, plays an important role in water resources management over agricultural sites and others environmental studies. Therefore, in the present study, the thermal infrared data (TIR-1) of Landsat 8 in December 24th, 2017, was used to retrieve LST using three different algorithms, namely Radiative transfer equation (RTE), Mono-window algorithm (MWA) and Single-channel method (SCM), and to estimate surface soil moisture status with retrieving LST, normalized difference vegetation index (NDVI) and temperature vegetation dryness index (TVDI) in the upper Lijiang River watershed. A comparison of retrieved LST among three different algorithms was conducted, and the retrieved LST and surface soil moisture status were characterized. The spatial distribution of LST was found to be consistent among three algorithms, higher around the outlet of upper Lijiang River surrounding by built-up areas, lower in the headstream areas occupying by forest land and natural reserves. The SCM displays the best agreement with surface measurements temperature from meteorological stations (mean values is 13.49°C) than other two algorithms, with an average values of 13.51°C, suggesting that the SCM is better for retrieving LST in the upper Lijiang River watershed, when calibrated and validated. The TVDI showed the same spatial distribution pattern with LST, lower TVDI values in the headstream and higher values near the outlet areas of the upper Lijiang River watershed, indicating that higher soil moisture distributing in the headstream than downstream areas of the upper Lijiang River watershed. Besides these, the factors like NDVI which influence the accuracy estimation of soil moisture through LST were also considered in this study.

KEYWORDS: Temperature vegetation dryness index (TVDI), Land surface temperature (LST), Radiative transfer equation (RTE), Mono-window algorithm, Single-channel method
An Integrated Runoff Risk Support Tool for Predicting Agrochemical Runoff Risk at Field and Landscape Scales

Comber*1, A.L. Collins 2, T.M. Hess 3, D. Haro Monteagudo 3, A. Turner 1, Y. Zhang 2

1 Leeds Institute for Data Analytics (LIDA) and School of Geography, University of Leeds, Leeds LS29JT, UK
2 Sustainable Agriculture Sciences Department, Rothamsted Research, North Wyke, Okehampton EX202SB, UK
3 Cranfield Water Science Institute, Cranfield University, Cranfield, Bedford MK43 0AL, UK

*corresponding author: a.comber@leeds.ac.uk

ABSTRACT

This Pesticides are one of the major pollutants of water resources. They present a significant problem in situations where receiving waters are used for potable (drinking) water supply - they either have to be removed or alternative sources have to be found. Both options are costly and are a technological challenge. Consequently, there is demand for predictive tools to determine pesticide delivery risks to abstraction points. However, despite many simulation risk models developed to evaluate the fate of pesticides, there are a number of key impediments to their use in an operational context:

- i) they have high data and input parameter requirements;
- ii) they suffer from poor performance when evaluated using monitoring data;
- iii) they assume and require in-depth knowledge about model process parameters;
- iv) they require access to spatial datasets which may not be available to commercial companies (such as water companies in the UK), and
- v) critically, they cannot be easily integrated with real-time data to quantify immediate risks (e.g. weather as precipitation is a key driver of agrochemical runoff).

This paper describes a parsimonious pesticide modelling approach for assessing pesticide pollution risks from agricultural diffuse sources at catchment and field scales for use by both drinking water agencies and by farmers. It uses historical and real-time weather data to estimate runoff using a modified Curve Number (CN) method that accounts for both event rainfall and antecedent soil moisture conditions. The parsimonious philosophy resulted in efforts to minimise model complexity but to ensure applicability in different landscape and field contexts using publicly available data. The analysis was undertaken for the Wissey catchment in England where Metaldehyde (used for slug control on wheat and oilseed rape), has, on occasion, been detected as part of the regulatory testing of surface water samples at abstraction points.

KEYWORDS: Soil water models; Pesticide delivery risk; integrated scales; Parsimonious models; Open data; Modelling in R

The results at the catchment scale account for pathway apportionment and associated pesticide loadings, organic carbon associated absorption and consequent Metaldehyde degradation (Figure 1).
Figure 1. Predicted metaldehyde delivery using the 95th (left) and 99th (right) percentiles of long-term (1900-2010).

The results for the meta-modelling approach at field scale show a good representation of runoff production despite different quality of estimation of antecedent soil water conditions (Figure 2).

Soil water antecedent conditions

Forested Runoff

Figure 2. Comparison between meta-model and model linear regression for antecedent soil water conditions (left) and runoff (right).

Figure 3. The interface to the Integrated Runoff Risk Support Tool.
Both catchment and field scale models were integrated into a web-mapping environment using the R shiny, leaflet and flexdashboard packages to create an interface to link the data and models with users. The models presents the user with different surfaces of field level runoff risk at different timescales, with risk predictions made for 0 to 5 days forward. The current state of the tool is illustrated in the snapshot in Figure 3 and the current live version can be found at https://lexcomber.shinyapps.io/webmockup/. The key feature of this work is that modally framework and model scan be used to determine risk for any agricultural application at field or catchment scales.
In-city Vegetation Quantification from Online Satellites Tiles Providers using Tungstene Technology

Fabien Pfaender *, Roger Cozien 

1 Shanghai University / Sorbonne University, China
2 eXo maKina Company, France

*Corresponding author: fabien.pfaender@me.com

ABSTRACT

Vegetation and nature-based areas are increasingly important for cities analysis. Several studies from a wide range of scientific fields point out that vegetation might provide a unique solution to urban space challenges: reducing the various levels of pollution (air, noise, light); promoting social qualities, (safety, health, wellbeing, vibrancy and creativity). Yet if availability of urban data made decisive progress in the past 10 years, in-city localized vegetation data remains too scarce to properly assess the impact of vegetation, especially for a precise “street-level” case studies. One of the reasons we identified was the gap between remote sensing expert focusing on large space usage obtained from exclusive high-end imagery and urban data experts looking for precise vector vegetation data while not being able to use expert processing.

This paper aims at bridging this gap by using an easy-to-find vegetation data source and converting it into a reliable vector layer. We used an online satellite tiles provider to get satellites pictures of Shanghai and Paris at the same zoom level. Then we introduced Tungstene technology to detect vegetation in the picture, highlighting it with a distinctive colour that can then be converted to a polygonal area. This area is finally detected using pixel colour detection and converted into a vector based format. This format is very popular in urban data science tools and easily incorporated into GIS softwares.

Tungstène itself is a new technology that was initially designed for the French Department of Defense. The first aim of this technology is to be able to very deeply process and assess any digital image. In doing so the first aim of the software was to detect, extract and amplify “weak signals” including manipulations and falsifications.

For this study, as Tungstène includes a large variety of filters, we mainly used a multi-spectral approach. Multi-spectral analysis is commonly used to get information beyond the visible spectrum. It can be used to get physical information about the objects pictured on the image such as light waves, turbulence, textures, materials and even, close infra-red information. Moreover, multi-spectral information can be extracted from regular and even low-end satellite imagery. In this study, the tiles come from mapbox.com online service which is itself using an in-house processing of NASA/USGS Landsat 5 & 7 that only include RGB bands.

This paper shows how this technology with its multi-spectral features was applied to rather common images and was able to successfully extracted reliable vegetation information. The process is very useful for urban monitoring and to have a quick and reliable mapping of the vegetation of a city. We tested the process on two examples: Paris, France, where vegetation coverage is already known, as a benchmark; Shanghai, China where vegetation is not directly accessible for urban data scientists.
Figure 1: top (a, b): Shanghai Vegetation (left), Shanghai satellite tile zoom level 11 (right), bottom (c, d): Paris Vegetation (left), Paris satellite tile zoom level 11 (right)

**KEYWORDS:** Vegetation; Multi-spectral Analysis; Urban Space; GIS

**REFERENCES**


Investigation of the multivariate spatial structure of a soils micronutrients data set at the farm-scale

Paul Harris *, 1, Tegan Darch 1, Binbin Lu 2, Alexis Comber 3, Michael Lee 1,4

1Sustainable Agricultural Sciences, Rothamsted Research, North Wyke, Okehampton, Devon EX20 2SB, UK
2School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
3School of Geography, University of Leeds, UK
4University of Bristol, Langford House, Langford, Somerset, BS40 5DU, UK

*Corresponding author: harvid50@gmail.com

ABSTRACT
Grassland is the largest crop by area in the UK covering just over half of the landmass. The forage it provides underpins most ruminant livestock production and we are at a crucial time for livestock agriculture when international pressures for both maximizing production and minimizing pollution on a global scale has led to a renewed focus on Sustainable Intensification as a means of balancing the opportunities and risks for grazing systems. The feeding requirements of ruminant livestock for macro- (water, energy and protein) and micronutrients (minerals and vitamins) are the bedrock of any sustainable farming enterprise. Rationing to match feeding requirements is often imprecise due to: i) intake and quality of pasture at grazing is often unknown; ii) high variability of the forage component of the diet at housing; iii) high genetic variability of ruminant livestock; iv) high environment variability (housing and at grazing); and v) the foregut microbial fermentation and the symbiotic role rumen microorganisms play in metabolism and nutrient composition post-rumen.

The complexities around these factors lead many ruminant livestock enterprises to optimize rations in order to meet production targets (milk yields and live weight gain), through supplementation of the diet with human-edible concentrate feeds (cereals, grains and pulses) or/and rumen-protected products.

Ultimately, this research aims to develop a new spatial tool to improve the precision of feeding requirements with respect to factors i) and iv) above, where, specifically, the spatial distribution of the soil micronutrients is directly contrasted to the spatial distribution of the forage (plant) micronutrients at the within-pasture scale, via a geographically weighted redundancy analysis (GWRDA). Through such a detailed characterization of spatial heterogeneity, within-pasture livestock grazing could be more efficiently managed for micronutrients (e.g. with mob grazing complemented with field to field movements) so that livestock feeding requirements would be less reliant on diet supplements. An improved understanding of the mineral status of soils and pasture should also improve the health status of the herd or flock. Although mineral content of the soil is of primary importance, the absorption of minerals from the soil to the pasture is complex, where livestock movement patterns, plant species, age of plants present, drainage conditions, lime and fertilizer treatments - all exert influences. Thus, consideration of these factors is also required to deliver key minerals to livestock solely from pasture.

As a precursor to the development of GWRDA, of which this study provides insights, this study will investigate a soils micronutrients data sets only, via a geographically weighted principal components analysis (GWPCA) (Harris et al. 2011). This is useful as GWPCA and geographically weighted regression (Brunsdon et al. 1996) provide the core building blocks for GWRDA development. The study data consists of n = 348 observations for 20 micronutrients (Al, As, Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Mo, Na, Ni, P, Pb, S, Se, Ti and Zn) collected across 21 fields of
an experimental research farm platform, in the south-west of England, UK. Through a GWPCA an investigation of the spatial heterogeneity in the structure of the micronutrients data is possible, where this study will also demonstrate the use of GWPCA: (a) to detect multivariate spatial outliers; (b) to optimally re-design a multivariate sampling scheme; and (c) to provide multivariate spatial input variables to improve spatial classification accuracy.

**KEYWORDS**: Geographically Weighted PCA; Visualization; Outliers; Network Redesign; Agricultural Grasslands Research

**REFERENCE**


Alternative interpretations of the particle size distribution of soils (0 -500cm) in the Loess Plateau, China

Paul Harris *,1, Ming’an Shao 2,3, Yihe Lü 4, Lianhai Wu *,1, Alexis Comber *5

1 Sustainable Agricultural Sciences, Rothamsted Research, North Wyke, Okehampton, Devon EX20 2SB, UK
2 College of Natural Resources and Environment, Northwest A&F University, Yangling 712100, China
3 State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Northwest A&F University, Yangling 712100, China
4 State Key Laboratory of Urban and Regional Ecology, Research Centre for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China
5 School of Geography, University of Leeds, UK

*Corresponding author: harvid50@gmail.com; lianhai.wu@rothamsted.ac.uk; a.comber@leeds.ac.uk

ABSTRACT

This study spatially investigates soil particle size distributions (PSD) using a combination of geostatistical and geographically weighted modelling techniques, using as a case study, 2673 disturbed soil samples collected in 243 soil profiles (0-500 cm) across the Loess Plateau of China. Samples for percentage silt, percentage sand and percentage clay were taken at 11 depths (0-10 cm, 10-20 cm, 20-40 cm, 40-60 cm, 60-80 cm, 80-100 cm, 100-150 cm, 150-200 cm, 200-300 cm, 300-400 cm, 400-500 cm) at each of the 243 sites. Soil PSD is a key soil physical parameter that affects soil conservation, water and nutrient movement and vegetation productivity. It is therefore crucial to any ecological restoration planning that may take place. It is commonly used in soil classification and in estimating soil hydraulic properties and is important for understanding the physical and chemical processes of soil development in terrestrial ecosystems.

The study PSD data has been analysed and presented before in Zhao et al. (2016). However, this study provides an alternative and complementary interpretation of the PSD data, where first compositional cokriging (Lark and Bishop 2007; Boogaart and Tolosana-Delgado 2008) is applied to the PSD data at each of the 11 sampling depths to provide 33 cokriging surfaces of PSD for silt, sand and clay (see Figure). The interrogation and interpretation of the 33 PSD surfaces is then facilitated via an application of a compositional geographically weighted principal components analyses (GWPCA) (Lloyd 2010; Harris et al. 2011) to the gridded cokriging predictions. In this manner, GWPCA is essentially used to spatially post-process 33 spatial distributions of PSD, which is extremely useful given their relatively large number.

GWPCA is a localized version of PCA that provides an exploratory tool for investigating spatial heterogeneity in the structure of multivariate data, where localised ‘percentages of variance explained’, localised loadings and localised scores are all outputted and visualized. The same outputs can also be used to (i) detect multivariate spatial outliers; (ii) optimally re-design the multivariate spatial sampling network; and (iii) provide multivariate spatial input variables to improve spatial classification accuracy, all of which are demonstrated to the predicted soil PSD data of this study. In particular, this study will focus on the use of a GWPCA provide a soil PSD classification map for the Loess region, where GWPCA spatial outputs are fed into a standard classification algorithm (Harris et al. 2015). Issues concerning the compositional nature of the PSD data are also discussed.

KEYWORDS: Soil Particle Size Distribution; Compositional Data; Cokriging; Geographically Weighted PCA
I INTRODUCTION

This study spatially investigates soil particle size distributions (PSD) using a combination of geostatistical and geographically weighted modelling techniques, using as a case study, 2673 disturbed soil samples collected in 243 soil profiles (0-500 cm) across the Loess Plateau of China (Figure 1). Samples for percentage silt, percentage sand and percentage clay were taken at 11 depths (0-10 cm, 10-20 cm, 20-40 cm, 40-60 cm, 60-80 cm, 80-100 cm, 100-150 cm, 150-200 cm, 200-300 cm, 300-400 cm, 400-500 cm) at each of the 243 sites. Soil PSD is a key soil physical parameter that affects soil conservation, water and nutrient movement and vegetation productivity. It is commonly used in soil classification and in estimating soil hydraulic properties and is important for understanding the physical and chemical processes of soil development in terrestrial ecosystems. It is crucial to any ecological restoration planning that may take place. This study aims to provide useful and novel insights into the geographical distribution of this rich and high dimensional PSD data set.

Figure 1. The 243 sample locations in the Loess Plateau, China.
Figure 2. Compositional cokriging prediction results at the top six depths (0 to 100 cm).
II METHODOLOGY

The study PSD data has been analysed and presented before in Zhao et al. (2016). However, this study provides an alternative and complementary interpretation of the PSD data, where firstly compositional cokriging (Lark and Bishop 2007; Boogaart and Tolosana-Delgado 2008) is applied to the PSD data at each of the 11 sampling depths to provide 33 cokriging surfaces of PSD for silt, sand and clay\(^1\). The interrogation and interpretation of the 33 PSD surfaces is then facilitated via an application of a compositional geographically weighted principal components analyses (GWPCA) (Fotheringham et al. 2002; Lloyd 2010; Harris et al. 2011) to the gridded cokriging predictions. In this manner, GWPCA is essentially used to spatially post-process 33 spatial distributions of PSD, which is useful given their relatively large number.

\(^1\)Observe that one cokriging system could have been applied to all 33 variables, but this would likely result in difficulties with ensuring positive-definiteness.
GWPCA is a localized version of PCA that provides an exploratory tool for investigating spatial heterogeneity in the structure of multivariate data, where localised ‘proportion of the total variance’ (PTV), localised loadings and localised scores can all be outputted and visualized. The same outputs can also be used to: (i) detect multivariate spatial outliers (Harris et al. 2014a; 2015); (ii) optimally re-design a multivariate spatial sampling network (Harris et al. 2014b); and (iii) provide multivariate spatial input (or texture) variables to improve spatial classification accuracy (Harris et al. 2015; Comber et al. 2016).

This study will focus on the use of a GWPCA to provide a potentially useful soil PSD classification map for the Loess region, where GWPCA loadings from cokriged PSD input data are fed into a (non-spatial) unsupervised hierarchical clustering algorithm.

![Figure 4](image-url). Top - PTV maps from GWPCA. Bottom - Ward cluster analysis on the GWPCA loadings for the first and second components (specified with five, seven and nine classes).

### III RESULTS

The results of applying 11 compositional cokriging systems (one system per depth) to the PSD data are given in Figures 2 and 3, for the PSD cokriging predictions only. Broadly, clay and silty soils tend to dominate in the south, whilst more sandy soils tend to dominate in the north. The PSD cokriging predictions are compositional, so they are transformed using the isometric-log-ratio (ilr) transform (Egozcue et al. 2003) prior to input in to the GWPCA calibration and implementation. The resultant GWPCA PTV maps are given for the first (PC-1), first and second (PC1-2), and first, second and third components (PC1-3) in Figure 4 (top), where a broad middle band of generally high PTV data suggests relatively strong similarities/relationships in the soil PSDs in relation to elsewhere. The results of the hierarchical cluster analysis for five, seven and nine classes (via Ward’s minimum variance method) on the local loadings from the first and second components of the GWPCA are given in Figure 4 (bottom). These PSD classification maps look interesting but need expert evaluation by soil scientists for their relevance.

### REFERENCES


1 Prediction uncertainty outputs are also available via the cokriging errors and the cokriging z-scores, both of which could be investigated through GWPCA.


Spatial-temporal Reconstructing of Forest Percent Cover by Fusing Multiresolution Remotely Sensed Images

Yihang Zhang 1, Feng Ling *, Xiaodong Li 1

1 Institute of Geodesy and Geophysics, Chinese Academy of Sciences, China

*Corresponding author: lingf@whigg.ac.cn

ABSTRACT

Forest cover is significantly important for the delivery of the global ecosystem services. Accurate mapping of spatial and temporal variation in forest cover by using remote sensing technique provides essential information for steering management practices in world’s forest resources. Given that forest stand structure is always heterogeneous and complex, sub-pixel forest percent cover is superior to pixel classification derived forest/non-forest map in accurately distinguishing forest cover from undesired spectral contribution of background features. Generally, there is typically a trade-off between the spatial and temporal resolution for most satellite remote sensing system, and spatial-temporal fusion method is becoming an attractive approach to provide fused remote sensing dataset used to generate fine spatial and temporal resolution forest percent cover. However, traditional image spatial-temporal fusing methods are focusing on the blending of pixel-based spectral reflectance values, and do not directly provide information on land cover dynamics. In this research, a novel Spatial-temporal Reconstructing of Forest Percent Cover Model (SRFPCM) is proposed to reconstruct fine spatial and temporal resolution forest percent cover by fusing multiscale forest percent covers maps that is the outputs of spectral unmixing of the original multiresolution remotely sensed images. Different with traditional pixel based image spatial-temporal fusing model, SRFPCM is used directly for the forest cover at sub-pixel scale instead of spectral reflectance values at pixel scale, and outputs a series of fine spatial and temporal resolution forest percent cover maps. For the experiments, the proposed SRFPCM against other image spatial-temporal fusion approaches, including the popular Enhanced Spatial and Temporal Adaptive Reflectance Fusion Model (ESTARFM) and the Flexible Spatiotemporal Data Fusion (FSDAF) model. Notably, in terms of both visual and quantitative evaluations, SRFPCM was able to reconstruct more accurate fine spatial and temporal resolution forest percent cover maps than those of ESTARFM and FSDAF, which shows the great efficiency and potential of SRFPCM. Moreover, SRFPCM was also applied to reconstruct time-series fine spatial and temporal resolution within one month, and the results showed SRFPCM is able to provide more accurate information about where, when and how the forest percent cover change through time.

KEYWORDS: Forest Cover; Spatial-Temporal Fusion; Spectral Unmixing; Reconstruction
A Platform for Sharing and Analysis of Low-cost Environmental Sensor Data

N. A. S. Hamm*1, E. A. Vásquez Gómez2, R. L. G. Lemmens3, S. Jirka4

1 School of Geographical Sciences, Faculty of Science and Engineering, University of Nottingham,
2 Department of Geoinformatics, San Carlos University of Guatamala, Quetzaltenango, Guatemala
3 Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands
4 52° North Initiative for Geospatial Open Source Software GmbH, Muenster, Germany

*Corresponding author: nicholas.hamm@nottingham.edu.cn

ABSTRACT

The diversity and volume of environmental data has increased enormously in recent years. An important issue has been the development of low cost sensors which can be organized into dense observation networks. Such sensor networks allow us to take more observations, with a higher temporal and spatial frequency than was possible previously. However, there are concerns about the quality of these data. Further the diversity of data sources as well as the heterogeneity of devices and data models pose technical challenges for the use of the data as well as conceptual challenges for the modelling and analysis that are performed on the data. In this research we developed a platform, defined as a set of web-based components, that could be used to archive, retrieve and process data and to distribute observations via the world-wide-web. We developed this within a citizen science context, provided by the AiREAS civic initiative (www aireas.com) and the associated air quality sensor network. This is a network of 35 “Airboxes” in the Dutch city of Eindhoven. Airboxes contain a range of low-cost sensors for particulate matter, NO2, ozone, ultrafine particles and temperature for which data are acquired at 10-minute intervals.

Our platform made use of the Sensor Web Enablement (SWE) framework of standards from the Open Geospatial Consortium (OGC), specifically the Sensor Observation Service (SOS), Observations and Measurements (O&M) and SensorML. Additionally we used the Web Processing Service (WPS) for calling, processing and analysis via the world-wide-web. The fact that our platform is based on standards is of central importance for ensuring interoperability, both in a technical sense and to make the data available to users for subsequent analysis and processing. We demonstrated this using a WPS for automated geostatistical modelling and interpolation which yielded both the kriged prediction and kriging variance. These outcomes could also be handled within our platform and subsequently made available to users. We present our platform within the context of the core interests of the Accuracy community.

Several points within this research are of relevance to Accuracy 2018:

1) We gave explicit attention to data quality and uncertainty, both in the sensor data and the outcomes from the WPS. These were handled using SensorML and the uncertainty markup language (UncertML). This helps address a core concern in the use of data from low-cost sensor networks, namely the data quality. We show that both machine and human users can query and visualize this data quality information.

2) The WPS for automated geostatistical modelling can be replaced by another type of analysis. Relevant services include automated data quality evaluation, outlier detection and data cleaning.

3) We demonstrate how data from the low-cost sensor network could be integrated with data from an authoritative sensor network. The use of standards is important for facilitating this integration.

Our research addresses technical, conceptual and practical challenges in the handing and processing of sensor data. Future work will make more explicit use of the quality information in subsequent analysis.
KEYWORDS: Air Quality; Sensor Network; Uncertainty; Spatial Data Quality; Sensor Web Enablement (SWE); Web Processing Service (WPS)
Study on Spatial Interpolation of Daily Precipitation Considering Spatiotemporal Autocorrelation and Non-stationarity: A Case Study over the Upper Reaches of the Bengbu Sluice in the Huai River Basin

Lingjie Li¹, Qingfang Hu¹, Yintang Wang¹, Leizhi Wang¹

¹State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing 210029, China

*Corresponding author: lilingjie_water@163.com; hqf_work@163.com

ABSTRACT

Due to the high spatial-temporal heterogeneity of precipitation, an accurate estimation of precipitation spatio-temporal distribution based on rainfall gauges observation has been one of the difficulties in hydrometeorological fundamental research. Considering that the Geographically Weighted Regression (GWR) and the Geographically and Temporally Weighted Regression (GTWR) have prominent advantages in characterizing the complex spatiotemporal autocorrelation and non-stationarity of spatial variables, this paper applies the GWR and GTWR models to precipitation spatial interpolation. It is noted that when temporal autocorrelation lag of spatial variable is zero, time dimension of GTWR model disappears and the model degenerates to GWR. Therefore, the two models are unified within the same computational framework through the time lag, consequently the universal precipitation spatial interpolation program using GWR and GTWR models is designed. Specially, due to the complicated influence of spatiotemporal evolution of cloud cluster, the daily precipitation often presents a discontinuity of dry & wet fields. To address the thorny issue, this study nests the wet & dry areas identification process and the precipitation spatial interpolation program mentioned earlier, then a new interpolation algorithm for daily precipitation considering spatiotemporal autocorrelation and non-stationarity is proposed. The method is applied to daily precipitation estimation from June to July 2014 over the upper reaches of Bengbu Sluice in the Huai River Basin with high-density rainfall network and the performance of the method is evaluated. The experiment is conducted using the coordinates of rainfall gauges as explanatory variables firstly. The simulated dry & wet fields can correctly identify 95% of rainy and rain-free sites, which is much better than other daily precipitation interpolation algorithms. Moreover, precipitation simulation at the site has smaller quantification errors (mean absolute error is 0.93mm) and a higher spatial consistency with observations (mean correlation coefficient reaches 0.85). Additionally, affected by the weakness in global & local correlation (global CC is close to 0 and local values are mainly distributed below 0.4) between precipitation and elevation, average accuracy of interpolation results decreases after the gauges elevation are expanded to another regression variable. Overall, from the perspective of application performance, the proposed algorithm owns generally satisfactory capability for reproducing spatiotemporal distribution characteristics of wet & dry fields and daily precipitation, which can provide favorable support for hydrometeorology research on large scales.

KEYWORDS: Daily Precipitation; Spatial Interpolation; Spatiotemporal Autocorrelation and Non-stationarity; GWR and GTWR Models; Dry & Wet Fields

ACKNOWLEDGMENT: The study is financially supported by the National Key Research and Development Program of China (No. 2016YFC0400902), the National Natural Science Foundation of China (No.51479118, No.51509157, No.51609140) and the National Non-profit Research Program of China (No.201501014).
A Comparative Study of Class-aggregated Versus Class-specific Logistic Modelling Approaches for Estimating Local Accuracies in Remotely Sensed Land Cover Information

Jingxiong Zhang 1, Wenjing Yang *2, Wangle Zhang 2, Yingying Mei 3

1 School of Geodesy and Geomatics, Wuhan University, China
2 School of Remote Sensing and Information Engineering, Wuhan University, China
3 Department of Sociology, Wuhan University, China

*Corresponding author: yangwj@whu.edu.cn

ABSTRACT

Accuracy is an important dimension in land cover information. It is increasingly recognized that accuracy in land cover maps should be characterized locally at the level of individual pixels as classification accuracy is spatially varied. There are also considerable differences between spatial variabilities of classification accuracies for different map classes. Thus, estimating local accuracy by class-specific models may be more adaptive to spatial non-stationarity in classification accuracy than by class-aggregated models. In this paper, class-aggregated and class-specific modelling approaches are compared through an experiment based on real datasets.

The experiment consisted of sampling design, validation sample data (concerning class reference labels) collection for model-building (called training data) and performance testing (reference data), modelling, statistical testing. Map-class-stratified random sampling was adopted to collect sample data. The initial sample size for each class was obtained through Neyman allocation, and then adjusted according to the actual situations of the experimental area, economic cost, efficiency, etc. For per-pixel accuracy modelbuilding, logistic regression analyses between classification correctness and certain pattern indices of map class occurrences were performed based on training data. Class occurrence pattern indices (homogeneity, heterogeneity, dominance, entropy, and contagion) were quantified in size-optimized moving windows. Logistic modelling was then carried out using the optimal explanatory variables selected. The same reference dataset was used by all accuracy-predictive models so comparisons among different methods were not confounded by differences in reference datasets.

The performances of class-aggregated and class-specific logistic models for accuracy predictions were evaluated using the area under the curve (AUC) of the receiver operating characteristic. It was confirmed that class-specific modelling approach is significantly more accurate and should be promoted for use in predictive mapping of local classification accuracies.

KEYWORDS: Land Cover; Stratified Random Modelling; Validation Sample Data; Class Occurrence Pattern Indices

1 INTRODUCTION

Land cover information is important for landscape characterization and monitoring, environmental monitoring and modelling, and global change. A variety of land cover information products are generated from different sensor datasets at regional and global scales (Chen et al., 2014; Wickham et al., 2017). However, no maps created from remote sensing can be completely accurate, and their errors will be propagated into resultant spatial analyses and models, affecting the validity of end results in spatial problem-solving. An understanding of the magnitudes and spatial distributions of errors in land cover information is crucial for their proper use.
Conventionally, land cover map accuracy is assessed via error matrices, with measures, such as overall accuracy, user’s accuracy, and producer’s accuracy, reported. However, classification accuracy is spatially varied (Foody, 2002; Smith et al., 2003). Local accuracy characterization is important for better understanding of misclassification’s locational and contextual fingerprints and for land cover data fusion and refinement.

To estimate local accuracies in land cover information, producers and users may employ various methods, such as calibration, empirical modeling, inverse distance weighting, kernel functions, and kriging (Zhang and Mei, 2016), using datasets available to them. In the paper, logistic regression was used to estimate the local accuracy. Logistic regression has been used to analyse the relationship between classification accuracy and land pattern indices (Smith et al., 2003; Van Oort et al., 2004).

In this paper, class-aggregated and class-specific modeling approaches to local accuracy mapping are compared based on real land cover datasets. The aim is to confirm whether estimating local accuracy by class-specific models is more adaptive to spatial nonstationarity in classification accuracy than by class-aggregated models.

II METHODOLOGY

a DATA

As a global fine-resolution land cover information product, GlobeLand30 (for 2000 and 2010), which was produced by the National Geomatics Center of China (NGCC) in 2014, has ten land-cover classes (http://www.globallandcover.com). GlobeLand30 data in 2010 for Wuhan city were used for the experiment. The dominate class is arable land, occupying about 60 percent of the study area, followed by water, forest, and residential areas, accounting for 15 percent, 12 percent, and 7 percent of the study area, respectively. Grass, wetlands, and bare land together accounted for about 6 percent of the study area.

b SAMPLE DESIGN

In this paper, stratified random sampling is adopted to obtain training sample data for model-building. Neyman optimal allocation method was used for sample allocation among different classes (Stehman, 2012). By Neyman optimal allocation, the sample size for a stratum is calculated as:

\[ n_h = n \frac{W_h s_h}{\sum_h W_h s_h} \]

(1)

where \( n_h \) is the number of sample units in stratum \( h \), \( n \) is total sample size, \( W_h \) indicates the stratum’s area proportion, and \( s_h \) represents the stratum’s standard deviation. The variances of individual strata were estimated by 1000 initial sample pixels in advance. The reference land cover class labels for sample pixels were obtained using interpretation of high spatial resolution images (i.e., Google earth images). A total of 3000 training sample pixels were collected for logistic modeling, while another sample set of 1020 testing pixels were collected for performance evaluation.

c LOGISTIC REGRESSION MODELLING

A logistic regression model describes the relationship between a binary response variable \( y_i \) and one or more explanatory variables \( x_k \) (Smith et al., 2003). Results of the regression is the probability of samples being correctly classified. The regression model is

\[ \log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{k=1}^{K} \beta_k x_k \]

(1)
where \( p \) is the probability of a correct classification, \( \beta = (\beta_0, \beta_1, \beta_2 \ldots, \beta_k) \) represents the parameters to be estimated.

Logistic regression analyses between classification correctness and certain pattern indices of map class occurrences were performed based on the training sample data acquired as in Subsection 2.2. Class occurrence pattern indices (homogeneity, heterogeneity, dominance, entropy, and contagion) were quantified in different-sized moving windows. Table 1 shows the candidate explanatory variables in logistic regression.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abb.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homogeneity</td>
<td>Hom</td>
<td>the number of pixels with the same label as the sample point occurred in the window</td>
</tr>
<tr>
<td>Heterogeneity</td>
<td>Het</td>
<td>the number of land cover class occurred in the window centered on the sample point</td>
</tr>
<tr>
<td>Dominance</td>
<td>Dom</td>
<td>the degree of the dominant or inferior status of plant species in plant community</td>
</tr>
<tr>
<td>Entropy</td>
<td>Ent</td>
<td>the average uncertainty of the source</td>
</tr>
<tr>
<td>Contagion</td>
<td>Con</td>
<td>the degree of agglomeration or extended trend of different patch types</td>
</tr>
<tr>
<td>Class</td>
<td>Class</td>
<td>the classifications of the sample point</td>
</tr>
</tbody>
</table>

Regression parameters are obtained by minimizing the \(-2 \log \text{likelihood}\) (also known as the deviance) of the model. The difference between the deviances of two models follows a \( \chi^2 \) distribution, where \( l \) is the number of explanatory variables additional to those shared by the two models. A \( \chi^2 \)-test was then used to test if adding these \( l \) variables to the model significantly improves the fit of the model. Logistic modelling was carried out using the optimal explanatory variables (quantified in size-optimized moving windows, as shown in Table 2) selected.

**d PERFORMANCE COMPARISON**

The area under the receiver operating characteristic curve (AUC) was used to evaluate accuracy predictions by class-aggregated and class-specific logistic models. AUC is a commonly used measure for assessing the performance of models constructed to predict binary outcomes. AUC values can theoretically range from 0 to 1 with larger value indicating greater accuracy in predictions. T-test was performed on AUCs to examine if there exist significant differences between the two modelling approaches.

**III RESULTS**

Table 2 shows the optimal explanatory variables selected in individual models and the statistical test result regarding the comparative performances of class-specific and class-aggregated modelling approaches. In the table, the subscripts of the variables indicate the sizes of moving window for quantifying patterns of class occurrences. The paired t-test comparing AUCs of the two approaches shows that class-specific modelling is significantly more accurate (p-value<0.005).
Table 2. Optimal explanatory variables selected and performances of the two modelling approaches.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Strata</th>
<th>Models</th>
<th>AUCs</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>class-specific</td>
<td>arable</td>
<td>Het7&amp;Ent3&amp;Hom39</td>
<td>0.7582</td>
<td>0.004708</td>
</tr>
<tr>
<td>class-specific</td>
<td>forest</td>
<td>Hom3&amp;Het25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>class-specific</td>
<td>grass</td>
<td>Het3&amp;Dom13&amp;Dom13&amp;Het39&amp;Con31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>class-specific</td>
<td>wetland</td>
<td>Dom13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>class-specific</td>
<td>water</td>
<td>Het3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>class-specific</td>
<td>residential</td>
<td>Het3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>class-specific</td>
<td>bare</td>
<td>Con11&amp;Ent11&amp;Het39&amp;Con17&amp;Dom13&amp;Ent15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>class-aggregated</td>
<td></td>
<td>Ent3&amp;Class&amp;Hom33&amp;Dom13</td>
<td>0.7385</td>
<td></td>
</tr>
</tbody>
</table>

IV CONCLUSIONS

In this paper, class-aggregated and class-specific modelling approaches were compared through an experiment based on real datasets. It was confirmed that the latter is significantly more accurate in predicting local accuracies. Further work is needed to investigate the relationships between the two modelling approaches’ performances and sample sizes.

REFERENCES

Mapping Change Detection of LULC on the Cameroonian Shores of Lake Chad and Its Hinterland through an Inter-Seasonal and Multisensor Approach

Gbetkom Paul Gérard1,2,3, Gadal Sébastien1, El Aboudi Ahmed2, Ngandam Homère3

1Aix-Marseille Univ, CNRS ESPACE UMR 7300, 13545 Aix-en-Provence, France;
2Laboratory of Botanic, Mycology and Environment, Faculty of Sciences, University Mohammed V Rabat;
3Laboratory of Image Processing, Stats n’ Maps, 13015 Audelia road, suite 9128, Dallas; USA

*Corresponding author: gerardgbetkom@yahoo.com, paul-gerard.GBETKOM@etu.univ-amu.fr

ABSTRACT
The aim of this study is to evaluate the land use/land cover (LULC) inter-seasonal changes along the Cameroonian shores of Lake Chad and its hinterland using four Landsat sensors images of MSS, TM, ETM+ and OLI. The generation of land use/land cover inter-seasonal changes is based on classification by Support Vector Machines (SVMs) algorithm. Three categories of land use/land cover are identified to analyze the evolution of the Cameroonian shores of Lake Chad: open water and marshland, vegetation, and bare soils. The results show that, among these three categories, bare soils has the most important percent (around 55%). Moreover, land use/land cover change from one season to another or from one decade to another can be closely linked to evolution of climate conditions. Concerning variation of land use/land cover on the Cameroonian shores of Lake Chad, open water and marshland surfaces varies little compared to variations of other categories. Vegetation has the most important variation compared to others categories (until 162 % of inter-annual changes). In addition, the proportions of bare soils that varies between rainy seasons are different between dry season (between 3 and 255% during rainy seasons and less than 10% during dry seasons).

KEYWORDS: Inter-seasonal changes; LULC; Support Vector Machines; Lake Chad

Table 1. Characteristics of scenes used.

<table>
<thead>
<tr>
<th>Sensors</th>
<th>MSS</th>
<th>TM</th>
<th>ETM+</th>
<th>OLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date of Acquisition</td>
<td>March 1973 and October 1975</td>
<td>March 1987 and October 1987</td>
<td>April and August 2000</td>
<td>April and September 2015</td>
</tr>
<tr>
<td>Projection</td>
<td>UTM, WGS84_zone 33N</td>
<td>UTM, WGS84_zone 33N</td>
<td>UTM, WGS84_zone 33N</td>
<td>UTM, WGS84_zone 33N</td>
</tr>
<tr>
<td>Pixel Size</td>
<td>60 m</td>
<td>30 m</td>
<td>30 m</td>
<td>30 m</td>
</tr>
<tr>
<td>Path/Row</td>
<td>198/051, 199/051</td>
<td>184/051, 185/051</td>
<td>184/051, 185/051</td>
<td>184/051, 185/051</td>
</tr>
</tbody>
</table>
REFERENCES


Efficient Nonstationary Autocorrelated Field Simulation using Convolution Filtering

Ashton Shortridge*, Thomas Bilintoh

Michigan State University Department of Geography, Environment, and Spatial Sciences, USA
*Corresponding author: ashton@msu.edu

ABSTRACT
Geostatistical models are used to generate equiprobable realizations of processes for a number of purposes, including error propagation. Existing implementations for applied purposes are difficult or impossible to use in cases where nonstationary spatial models are preferred. In this paper we present an extension to stationary convolution filtering which can simulate processes with varying nugget, sill, and/or range, as well as means. Performance on simulated and actual environmental data are described. Potential for large-scale Monte Carlo-based error propagation modeling is discussed.

KEYWORDS: geostatistics; error propagation; filtering

I INTRODUCTION
Spatial data contain error, and much spatial accuracy research has considered how best to characterize the effects of this error on applications that use the data. Early research such as Peter Fisher's (1992) work on the sensitivity of the viewshed operation to digital elevation model (DEM) error introduced a general Monte Carlo simulation approach for error propagation: 1) develop an error model, perhaps using limited reference data, that captures the distributional and spatial properties of error, 2) uses this error model to construct equiprobable spatial realizations of the process, and 3) deploys each realization in turn on the application of interest.

A major challenge for this general approach is its computational complexity: generating spatially structured realizations using conventional approaches like sequential simulation (X) suffers when the number of points to be simulated, or the neighborhood to be used, becomes large. A computationally efficient error propagation approach for error propagation on DEMs was implemented by Oksanen and Sarjakoski (2005). This approach uses convolution filtering. First, a raster filter is constructed with a shape matching a Gaussian 2D spatial covariance model with specified range and sill. Second, this filter is scaled to produce a weighting kernel. Third, a raster grid for the area of interest is populated with random noise, and then the weighting filter is applied to smooth this, producing a raster surface with a structure matching the specified Gaussian model. This process is graphically portrayed in Figure 1. Hebeler (2006) implemented this model in Matlab.

A limitation of the convolution filtering approach described above is that it uses a spatially stationary model: parameters like the mean, nugget, sill, and range are constant. At the regional scale these stationary models may only roughly correspond to the actual DEM error distribution, which can display substantial variation in these parameters based on land cover, slope, and other variables.

In this paper we extend the convolution filtering approach to operate on a nonstationary context on continuous data such as DEMs. In particular, we implement an R-based so
olution (R Core Team, 2018) which enables users to input these parameters as numerical constants or as rasters with varying values. The implementation is validated using synthetic datasets and demonstrated with a real-world example.

II MODEL

Our implementation begins with a conversion of the Hebeler Matlab code to R. The resulting functions can be used to rapidly generate raster representations of stationary random fields with spatial structure matching a user-specified Gaussian covariance model. This code is modified so that R raster-class objects can be passed as mean, nugget, (partial) sill, and range parameters. The first three of these are straightforward and very efficient: the mean raster is added via map algebra to the output of the convolution filter; the nugget raster is used to rescale values on a cell by cell basis in a second raster of standardized random noise, and the sill ratio is used in similar fashion to rescale each cell in the spatially structured raster. Implementing the fourth -- a nonstationary range -- is more difficult conceptually and computationally. Our present solution is to generate n rasters of spatially structured noise, one for each unique range value in the raster. These surfaces are then masked and mosaiced together like a jigsaw puzzle.

To test this model, we implement it on several synthetic test cases with varying model parameters and test whether each area subset has spatial properties matching its underlying model. Then we use it to simulate error for a larger region in Africa and evaluate its performance.

III RESULT

In the first experiment a raster array is developed with 500 columns and 400 rows and 10 meter cell resolution. The left hand side of the raster is modeled as a Gaussian spatial process with nugget 8, partial sill 12, range 80, and mean 9. The right hand side of the raster is modeled as a raster process with nugget 16, partial sill 30, range 160, and mean 13. Maps of one realization of the raster (split into left and right portions) with modeled and empirical variograms are presented in Figure 1.

![Figure 1: Maps and variograms for one realization of a nonstationary process.](image)

In Figure 2, 20 realizations of this nonstationary process are represented. As expected, the empirical variograms vary about the process variogram model for each region. The fit is particularly strong out to about 50 meters (five cells) for the left-hand region and 120 meters (12 cells) for the right-hand region. Past that distance realizations generally have
lower spatial variation than the process model for the left-hand region and higher variation for the right-hand region.

![Variograms for 20 realizations of a nonstationary process.](image)

**Figure 2:** Variograms for 20 realizations of a nonstationary process.

For the second experiment using environmental data, we used an NDVI image of central Ghana derived from a Landsat 8 image (obtained from the USGS website). The NDVI image was reclassified to obtain three classes: 1, 2 and 3 corresponding to very low vegetation, low vegetation and high vegetation respectively Table 1.0 summarizes the its characteristics.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columns</td>
<td>1,578</td>
</tr>
<tr>
<td>Rows</td>
<td>1,313</td>
</tr>
<tr>
<td>Cell resolution</td>
<td>30 x 30 m</td>
</tr>
<tr>
<td>Low vegetation cells (1)</td>
<td>584,986</td>
</tr>
<tr>
<td>Med vegetation cells (2)</td>
<td>871,330</td>
</tr>
<tr>
<td>High vegetation cells (3)</td>
<td>611,062</td>
</tr>
<tr>
<td>Coordinate System</td>
<td>UTM zone</td>
</tr>
<tr>
<td></td>
<td>30 N</td>
</tr>
</tbody>
</table>

A Gaussian model was used to generate four (4) simulations with: ranges, partial sills and nuggets from the manipulation of the input NDVI data. As expected, the structure of the realizations varies spatially, with longer ranges and higher variances observable in some regions. This means that the algorithm that has been proposed in this paper is generating a nonstationary autocorrelation field from the given inputs.
In order to produce realizations that have parameters depicting nonstationary autocorrelation, spatially varying ranges, partial sills and nuggets were derived from a manipulation of some of the characteristics of the classified NDVI data. Table 2 outlines the parameters used to generate the nonstationary model with realizations displayed in figure 4.

Table 2. Variogram parameters for different vegetation classes.

<table>
<thead>
<tr>
<th>Vegetation Class</th>
<th>Range</th>
<th>Sill</th>
<th>Nugget</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>300</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>Class 2</td>
<td>600</td>
<td>75</td>
<td>3</td>
</tr>
<tr>
<td>Class 3</td>
<td>900</td>
<td>112</td>
<td>10</td>
</tr>
</tbody>
</table>

Computational efficiency is an important goal. Figure 5 is a graph of the number of simulations produced against processing time to generate 1, 5, 10, 15 and 20 realizations with 3 replications each. The process was run on a Dell Latitude 7480 with an Intel I5 7300U with 4 CPUs running at 2.6 Ghz and 8 GB of RAM. While the lowest time was about 25 seconds to run a single simulation, the highest number of simulations (20) took 49 seconds to run. It is clear from these statistics that the algorithm proposed generates realizations at quite a fast rate (given the size of the input data: 2,071,914 cells) and that production time is linear.
IV CONCLUSION
Initial experiments indicate promise for the use of convolution filtering for computationally efficient production of nonstationary autocorrelated random fields. These fields can be used for the
development of unconditional realizations of data error in global DEMs and other products where statistical models of error can be usefully employed for error propagation analysis. We find that simulation of spatially varying means, nuggets, and sills is quick and efficient, while variation in the range parameter is more challenging. Future work on more efficient processing of varying ranges is warranted. As currently implemented, simulation on large arrays with more than a few unique range values is not practical. The implementation of alternative semivariance model forms (e.g., exponential, spherical, matern) would also make the methodology more flexible. Nevertheless, this method is already a tractable approach to implement more sophisticated geostatistical error models.

REFERENCES
A Comparison of GWR Software on Various Platforms

Ziqi Li

Spatial Analysis Research Center, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe

ABSTRACT

GWR (Geographically Weighted Regression) is a widely-used tool for exploring spatial heterogeneity in many fields. There are many packages currently available on various platforms that can fit a GWR model, but there is no existing literature comparing those packages and providing a guide to researchers. In this study, four popular and frequently-maintained GWR packages, namely GWR4, GWR in PySAL, GWModel, spgwr are compared along three aspects: 1) functionality, 2) consistency and 3) computational speed and scalability. Functionality is compared by reviewing the documentation for each package. Consistency is tested using the same empirical datasets with the same model options to examine if every package produces the same results. Scalability is investigated by examining how computation time scales when increasing the number of data points and the number of covariates. Simulation surfaces are generated in order to test both runtime and scalability. While all four packages have essentially the same options for fitting a basic GWR model, GWModel and GWR in PySAL have additional advanced features developed from recent literature. All four packages yield consistent results which is encouraging although GWR in PySAL is the fastest and most scalable.

KEYWORDS: Geographically Weighted Regression; GWR; PySAL; GWModel; Computational speed
Analysis and Application of Drought Characteristics based on Run Theory and Copula Function

Wang Xiaofeng*1,2, Zhang Yuan*1,3, Feng Xiaoming3

1School of Earth Sciences and Resources, Chang’an University, Xi’an 710054, China
2School of Land Engineering, Chang’an University, Xi’an 710054, China
3State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, CAS, Beijing 100085, China

*Corresponding author: wangxf@chd.edu.cn; 2227513984@qq.com

ABSTRACT

Drought is a complex and recurrent climate phenomenon, and understanding the development of the drought event is of great significance in preventing the damage. Current studies have analyzed the drought duration and severity based on run theory and Copula function but few focuses on the drought recurrent interval and its relationship with agricultural drought disaster. In this study, the drought recurrence interval was investigated based on the drought duration and severity using run theory and Copula function and the results were applied to agricultural drought disaster analysis in order to verify the reliability of the method for agricultural application. The study area was Northern Shannxi and the precipitation data from 1960 to 2015 were collected for drought grade classification based on standardized precipitation index (SPI). The drought duration and severity of drought event was calculated based on SPI values using run theory. Each drought characteristic index was evaluated for its distribution by comparing 7 different distributions. Then, the joint distribution of the 2 indices was established based on their correlation. Then the recurrence interval of each index was calculated for determination of their joint recurrence interval. The agricultural disaster area data were obtained from Statistic Yellow Book. The results showed that the drought occurred 94-103 times in 1960-2015 in the Northern Shannxi. The main drought area with high frequency was the area such as Yunlin, Yanan et al and that with low frequency was in the area such as Baoji, Xian, Xianyang et al. The distribution of drought duration and severity belonged to Weibull and Gamma distribution, respectively. The joint function of the 2 indices could be well fitted by the Frank-copula function with smaller squared euclidean distance and akaike information criterions. With the increase of drought duration and severity, the joint probability and recurrence interval gradually increased. The joint recurrence interval variation of different meteorological stations had different changing trend. The actual drought affected area of the crops in the Northen Shannxi was closely related to the drought duration and severity. The drought events joint recurrence interval was coincide with the actual drought disaster situation in the region. In the period of 1995-2005, the actual and theoretical drought duration was consistent mostly and the absolute errors between the actual and theoretical drought recurrence interval were only 0.1-0.3 a. The results indicated that the method proposed here was reliable in analyzing drought characteristics and could be used for agricultural drought disaster assessment.

KEYWORDS: drought; precipitation; disasters; run theory; Copula function; drought recurrent interval; Northern Shannxi
Including Spatial Correlation in Structural Equation Modelling of Soil Properties

Gerard Heuvelink *, Marcos Esteban Angelini 2

1 Wageningen University, Netherlands
2 INTA, Instituto de suelos, Argentina

*Corresponding author: gerard.heuvelink@wur.nl

ABSTRACT

Digital soil mapping techniques usually take an entirely data-driven approach and model soil properties individually and layer by layer, without consideration of interactions. In recent studies we implemented a structural equation modelling (SEM) approach to include pedological knowledge and between-properties and between-layer interactions in the mapping process. However, SEM typically does not consider spatial correlation. Our goal was to extend SEM by accounting for residual spatial correlation using a geostatistical approach. We assumed second-order stationary and estimated the semivariogram parameters, together with the usual SEM parameters, using maximum likelihood estimation. Spatial prediction was done using regression kriging. The methodology was applied to mapping cation exchange capacity, clay content and soil organic carbon for three soil horizons in a 300km by 200km study area in the Great Plains of the United States. The residuals showed substantial spatial correlation, which indicates that including spatial correlation yields more accurate predictions. We also compared the standard SEM and the spatial SEM approaches in terms of SEM model coefficients. Differences were substantial but none of the coefficients changed sign. Presence of residual spatial correlation suggests that some of the causal factors that explain soil variation were not captured by the set of covariates.

KEYWORDS: Digital Soil Mapping; Geostatistics; Structural Equation Modelling; Regression Kriging
Evaluation of Heavy Metal Cu Pollution in Soil Based on Cokriging

Yingjun Sun*1, Yunting Jiang1, Qi Wei1

1School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, China

*Corresponding author: sdjusyj@126.com

ABSTRACT

Soil plays an important role in the ecological environment. With the rapid development of industry, the increase in category and quantity of agrochemicals, together with more and more vehicle exhaust, different heavy metal enter into soil in urban and rural area. The estimation of heavy metal pollution in unsampled points is the prerequisite for management of environment. Usually, one can interpolate the attribute value of unsampled points based on sample points. In this paper, the spatial distribution of heavy metal Cu was analyzed. A 2×2 km grid was established in the study area. The centre of each grid represents the location of each sample point. Soil samples were mixed by five subsamples from different site within 20×20 m area. 248 sample points were collected in the entire area. Seven heavy metals were extracted from the soil samples. In the exploratory analysis, the paper found high similarity between Cu and Zn. Based on the geostatistics software - Gstat in R Foundation for Statistical Computing Platform, the paper got the coefficient between Cu and Zn. The value was 0.68 which reveal strong relationship between them. Thus, the paper took Zn as the secondary variable to estimate the spatial distribution of Cu. 48 sample points were chosen as validation set. The other 200 sample points were interpolated by means of OK (Ordinary Kriging) and Cokriging independently. The result showed that the relative error of OK was from 0.52% to 19.49%; Cokring was from 0.96% to 19.02%. The standard deviation of OK was 5.33%, Cokring was 5.04%. The root-mean-square error of OK was 5.39%, Cokring was 5.10. So, with the factor of standard deviation and root-mean-square error, the paper drew the conclusion that the results of Cokring were more accurate.

KEYWORDS: heavy metal; Cokring; accuracy; relative error
Accuracy Assessment of Lidar-Derived Dem for Flood Risk Mapping

Xiaohang Liu

1San Francisco State University, U.S.A

*Corresponding author: xhliu@sfsu.edu

ABSTRACT

In recent years, LiDAR has become the standard technology to collect high-density, high-accuracy elevation data for flood mitigation (Stoker et al. 2008). However, the overall accuracy of a LiDAR raw point cloud does not always translate to the accuracy used by flood risk assessment; accuracy assessment of LiDAR-derived DEM is thus necessary. This paper presents approximation theory as a new paradigm for accuracy assessment and illustrate its application to flood risk mapping. Approximation theory is widely used in numerical analysis in computational science to assess the quality of approximating a complex function by simple functions and quantify the errors introduced therein (Atkinson and Han, 2004). Its rationale is simple: if the largest error is even acceptable, then the error at any point must also be acceptable, hence the overall approximation must be acceptable. Based on this idea, the total error in a LiDAR-derived DEM is first separated into three components: LiDAR sensor error, ground error, and interpolation error, and derive a mathematical definition for each. Approximation theory and numerical analysis are then employed to show how an error bound can be obtained for each component as well as the total error at any location.

A case study is used to illustrate how approximation theory was applied to assess whether the Golden Gate LiDAR data (Ehines, 2012) met the FEMA (Federal Emergency Management Agency) Guideline for Regions and Mapping Partners concerning the accuracy and processing of high-quality topographic data including LiDAR (FEMA, 2010). In a rolling to hilly area with a high flood risk, as is the case with our study site, FEMA requires that a LiDAR-derived DEM must have an accuracy equivalent to 4-foot contours. Using approximation theory, it is found that, under the best case scenario, only 72% of the study area met such a requirement, thus the LiDAR-derived DEM is not appropriate for mapping areas with high flood risk. However, it would be adequate for areas whose risk is medium.

The case study illustrates that, compared to existing accuracy assessment methods which lumps all errors together and summarizes them by sample statistics, approximation theory enables the identification of main source of error and uncertainties introduced at different stages. Better yet, it points out where the desired accuracy is or is not met. For locations with excessive error, approximation theory further points out how such error can be effectively reduced by adding additional bare-earth points before gridding.

KEYWORDS: LiDAR; Digital Elevation Model (DEM); approximation theory; flood risk mapping

REFERENCES

Evaluating of Terrain Shading Algorithms Using Artificial Surfaces for Solar Radiation Modeling Over Mountainous Terrain

Shuhua Zhang *1, Xingong Li 2, Jiangfeng She 1

1 Nanjing university, China
2 University of Kansas, United States

*Corresponding author: zhangshuhua11@mails.ucas.ac.cn

ABSTRACT

Solar radiation is the ultimate energy source of earth surface systems and it affects many atmospheric, hydrological, and biological processes. In mountainous terrain, however, shading may result in significant modification on solar radiation, either by the inclination of the surface itself or the blocking of adjacent terrain and solar radiation shows a large spatial heterogeneity. Two typical shading algorithms based on GIS, hillshade and skyshed (viewshed called in Solar Analyst) are used to simulate the shading effects. In this research, a data independent method (mathematical/artificial surfaces) was developed to evaluate the accuracy of the two shading algorithms and its influence on solar radiation modeling. The artificial surface of V- and U-shaped were designed. V- or U-shaped valleys are the fundamental geomorphological forms, especially in the areas of rapid uplift, due to fluvial and glacier erosion processes and the shading effects of these surfaces can be calculated theoretically. The solar radiation model of Kumar’s model (KM) and Solar Analyst (SA) were used to evaluate the accuracy of solar radiation in mountainous area since the Kumar’s model uses hillshade and Solar Analyst uses skyshed to simulate the shading effects. For direct radiation, the result shows that hillshade overestimates shadow with a mean relative difference (RD) of 3.55% and larger shadow errors are found in spring (RD of 8.45%) and autumn (RD of 4.73%). Skyshed underestimates by 8.39% and the skyshed error is smaller on south-facing surfaces than north-facing surfaces. In addition, the boundary of skyshed based on SA is not smooth as theoretical boundary because of the interpolation of horizon angle. For diffuse radiation, sky view factor is an important factor and skyshed models sky view factor very well with mean absolute relative difference of only 0.76%. We also conclude that the sky view factor in SA is most sensitive to the number of zenith divisions but not to azimuth division. Finally, we evaluated two typical solar radiation models on V and U surface and KM performs better than SA for the calculation of global radiation because KM performs better in direct radiation using hillshade which is the major component of global radiation.

KEYWORDS: Terrain Shading; Artificial Surface; Mountainous Area
An Interval Uncertainty Propagation Method based on Kernel Density Estimation

Dandan Feng 1, Jianghong Ma *1, Yee Leung 2

1 Chang’an university, China
2 Chinese University of Hong Kong, China

*Corresponding author: jh.ma@163.com

ABSTRACT

In this paper, we propose a new interval uncertainty propagation method for calculating the range and the distribution of outputs, which is based on kernel density estimation, generalized Latin hypercube sampling and cubic spline interpolation. The proposed method is applied to solve the interval uncertainty in GIS and the Sandia challenge problem. Experimental results have demonstrated the accuracy and advantages of the new method by comparing with other methods in the literature.

A non-parametric estimation method is used to fit probability density functions (PDF) for input variables whose value range is an interval or multiple intervals. This probabilistic method can not only be suitable for the situation in which the range of input variables is an interval or multiple intervals, but also can be applied to solving problems involving uncertainty representation and propagation in the presence of point data and interval data so that the representation and propagation of uncertainty are simplified and unified. The interval uncertainty is expressed by the probabilistic method in this paper. Therefore, the well-developed probabilistic uncertainty propagation method can be used when the uncertainty propagation analysis is performed such as Monte Carlo and the first-order reliability analysis. When fitting a PDF of interval data, we adopt the non-parametric method which makes the obtained PDF more confidential to the original data, which thus reflects the inherent uncertainty of the interval data. In addition, the new method can give the output range and estimate the cumulative distribution function of the output, which is very useful in some real applications.

KEYWORDS : Interval Uncertainty Propagation; Monte Carlo; Nonparametric Probability Density Estimation; Generalized Latin Hypercube Sampling; Cubic Spline Interpolation; Challenge Problems
A Hybrid Approach for Uncertainty Propagation with Random and Interval Variables

Xiaojing Song 1, Jianghong Ma *1, Yee Leung 2

1 Chang’an university, China
2 Chinese University of Hong Kong, China

*Corresponding author: jh.ma@163.com

ABSTRACT

In recent years, uncertainty theory has been constantly developing and improving, and various approaches for uncertainty representation and propagation have been proposed. Uncertainty is unavoidable for data. With the operation of data, uncertainty will spread and accumulate, which may have a significant impact on the decision result. In many practical problems, multiple types of uncertainty often coexist. Therefore, it is a significant challenge to study a novel uncertainty propagation approach for combining aleatory uncertainty and epistemic uncertainty. In this paper a new hybrid uncertainty analysis technique for a non-linear function model is put forward, which can deal with the problems with random and interval variables.

Consider a mathematical function model where some of the input variables are interval variables and the rest are random variables whose uncertainty is described by a known probability distributions where the distribution parameters are interval form. Firstly, equal probability transform is carried out for separating the distribution parameters from the distribution functions of the random variables. With regard to the aleatory uncertainty, Latin hypercube sampling is used for uncertainty simulation of random variables. Then Polynomial Chaos Expansion (PCE) based on Legendre orthogonal polynomials is introduced to interval uncertainty propagation in the system model.

PCE is derived from the homogeneous chaos theory proposed by Wiener in 1938. For the past few years, PCE method has been applied widely to uncertainty propagation problems. The construction of multidimensional Legendre polynomials is an important part of PCE approach. And the determination of the expansion coefficients is the most important step in PCE method, which determines the accuracy of output estimation. The non-intrusive method is used to solve the PCE coefficients. Non-intrusive methods include the projection and regression method.

Finally two numerical examples are investigated to demonstrate the effectiveness of the proposed hybrid method. One of them is oscillator problem. It is concluded that the second-order PCE and the third-order PCE method provide almost the same results, but the computational time of the former is greatly reduced. Another example is a risk model for the design of a flood protection dike. We can use genetic algorithm, simple interval calculation and the improved interval calculation to compare with PCE method in the interval uncertainty propagation process. It has been shown in the exemplification that PCE outperforms other interval analysis methods in term of computational precision of the model output estimation. In addition, the computational time of PCE approach is obviously lower than that of the other methods. The PCE method performs particularly well in computational time and calculating accuracy when there are a large number of interval variables or the nonlinear degree of the function is very high. And in general, the third-order PCE could provide very accurate results for any kind of hybrid model.

The combined method proposed in this paper is available for propagation of probabilistic and interval uncertainty. Future research will be devoted to make the hybrid approach more widespread and accessible in more applications than ever before and to develop methods to deal with this problem arising in system model with fuzzy variables.
KEYWORDS: Hybrid Uncertainty; Interval Variables; Equal Probability Transform; Latin Hypercube Sampling; Polynomial Chaos Expansion (PCE)
Gslpe: a Spatial Upscaling Method for Multipoint P-normal Measurements

Feng Liu\textsuperscript{1,}, Xin Li\textsuperscript{1,2}

\textsuperscript{1}Key Laboratory of Remote Sensing of Gansu Province, Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Science, Lanzhou 730000, China
\textsuperscript{2}Chinese Academy of Sciences Center for Excellence in Tibetan Plateau Earth Sciences, Beijing 100101, China

*Corresponding author: liufeng@lzb.ac.cn

ABSTRACT

This study is based on the assumption that a normal distribution may not accurately reflect the sampling probability distributions of geophysical variables. As a generalization of the normal distribution, the p-normal distribution and its corresponding maximum likelihood estimation (the least power estimation, LPE) are introduced in upscaling method for multipoint measurements. Corresponding algorithm (GSLPE, GeoStatistical LPE) is also developed by considering the spatially correlated information with ordinary kriging. GSLPE is compared with the other 6 methods, namely three normal-based methods, i.e., arithmetic average, least square estimation, block kriging, two p-normal-based methods, i.e., LPE and Inverse Distance Weighted LPE (IDWLPE), and multi-scale cokriging with truncated power variogram.

There are mainly two types of experiments are conducted in this study. The first one is a synthetic experiment to evaluate the performances of the upscaling methods in terms of accuracy, stability and robustness. A 9 × 9 points spatial correlation region is established using the spatially correlated normal random field to produce the measurements at each point, except the one at the center of the region. Since the region is mean homogeneous, the average of the normal random field can be regarded as the macro-scale true value. The criterion of this experiment is the mean absolute error (MAE). The accuracy test shows that GSLPE produced the smallest MAE, indicating that this method had the highest precision, the stability test proves that a normal random field could have an adverse effect on the study of spatial correlation information, and LPE-based methods are more robust in the robustness test. The second experiment is a real-world experiment to produce upscaling estimates using soil moisture data obtained from multi-scale observations during an intensive observation period of the Heihe Watershed Allied Telemetry Experiment Research (HiWATER) project in 2012. The point measurements are based point-scale soil moisture data and upscaled to validate the footprint-scale measurements, namely COsmic-ray Soil Moisture Observing System (COSMOS) observations (Figure 1). The MAE values between the upscaling estimates and the COSMOS data indicate that GSLPE and LPE produced the smallest errors, which correspond to the highest precision.

URL of shared computer code: https://github.com/uniliufeng/LPE-based-upscaling-algorithms
We conclude that careful attention must be given to determining whether the geophysical variables of interest are normally distributed, and introducing appropriate statistical parameters (in this study, the extra parameter is \( p \)) into an upscaling strategy can substantially improve the estimation, especially if the raw measurements are disorganized; however, further investigation is required to determine which parameter is the most effective among variance, spatial correlation information and parameter \( p \).

**KEYWORDS:** Generalized Gaussian distribution; Multi-scale; Least power estimation; Geostatistics; Soil moisture; Heihe Watershed Allied Telemetry Experimental Research (HiWATER)

**REFERENCES**


A Quantitative Assessment of Spectral Unmixing Strategy on the following Subpixel Mapping for Hyperspectral Imagery

Xiong Xu *1, Xiaohua Tong *1, Huan Xie 1

1College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China

*Corresponding author: xvxiong@tongji.edu.cn

ABSTRACT

Subpixel mapping techniques have been widely utilized to determine the spatial distribution of the different land-cover classes in mixed pixels at a subpixel scale by converting low-resolution fractional abundance maps (estimated by a linear mixture model) into a finer classification map. Over the past decades, many subpixel mapping algorithms have been proposed to tackle this problem. It has been obvious that the utilized abundance map has a strong impact on the subsequent subpixel mapping procedure. However, limited attention has been given to the impact of the different aspects in the spectral unmixing model on the subpixel mapping performance. In this paper, a detailed quantitative assessment of different aspects in linear spectral mixture analysis, such as the criteria used to determine the types of pixels, the abundance sum-to-one constraint in the unmixing, and the accuracy of the utilized abundance maps, is investigated. This is accomplished by designing an experimental procedure with replaceable components. By investigating these critical issues, we can further improve the performance of subpixel mapping techniques.
Constrained Total Least Squares Method with Applications in Map Generalization

Yanmin Jin *1, Xiaohua Tong 1

1 Tongji University, China

*Corresponding author: jinyanmin@tongji.edu.cn

ABSTRACT

A data quality control approach for line features simplification in map generalization by the use of the constrained total least squares (CTLS) method is presented. The geometric characteristics of line features are formulated as constraint equations which are solved using the CTLS method. More and more research attention has been paid to the data quality in map generalization. The least squares (LS) method has been employed as an optimization operator to solve the multi-constraints problem in map generalization with the aim been to maintain the data quality (Harrie and Sarjakoski 2002; Sester 2005). However, the classical LS method only takes into account the errors in the observation vector. When the design matrix in the observation equations system also contain errors, it is appropriate to use the total least squares (TLS) method. The TLS method is designed for the errors-in-variables (EIV) model where both the observation vector and the design matrix contain observed variables (Golub and van Loan 1980). Furthermore, various CTLS methods have been developed to solve the EIV model with constraints. Tong et al. (2015) presented a polygonal boundary simplification with area constraints by the use of structured and constrained total least squares adjustment in order to maintain the area of the original polygons after the simplification. We further extend the data quality constraints to include more geometric constraints and solve the line simplification problem based on the CTLS method.

KEYWORDS: Total least squares; Constraints; Map generalization

REFERENCES

Super Resolution Mapping of Farmland based on Convolutional Neural Network

Yuanxin Jia\textsuperscript{1,2}, Yong Ge\textsuperscript{*1}, Yuehong Chen\textsuperscript{3}, Yan Jin\textsuperscript{1,2}

\textsuperscript{1} State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
\textsuperscript{2} University of Chinese Academy of Sciences, Beijing 100049
\textsuperscript{3} School of Earth Sciences and Engineering, Hohai University, Nanjing 210098, China

*Corresponding author: gey@lreis.ac.cn

Abstract:
Land-use and land-cover (LULC) change is fundamental information for global change, city planning, sustainable development. Especially, farmland dynamic change has manifold effects on both human society and ecosystems. Remote sensing image is the most important data source to obtain regional farmland map. However, due to the inter-constraint of spatial, temporal and spectral resolution of remote sensing images, high spatio-temporal land cover map of large area is difficult to obtain rapidly. Super-resolution mapping is a land cover mapping method, which it can get the finer land cover map directly from the coarse remote sensing image. The objective of this paper is to obtain a series of multi-temporal 10m spatial resolution Sentinel-like farmland map from Landsat images by super-resolution mapping method with convolutional neural. Specifically, a super-resolution mapping method based on multi-scale convolution neural network is proposed to get the finer land cover map, and the proposed method was deployed for the purpose to rapidly obtain a time series finer land cover map from 2008 - 2017. The proposed method was conducted on the farmland mapping of Zhangbei County in China between 2008 - 2017 and an average overall accuracy of 82.5\% was achieved, which brings a new perspective for farmland dynamic study on finer and accurate scale.

KEYWORDS: Super resolution mapping; farmland; Zhangbei County; Convolutional neural network
Spatial Sampling Optimization Design for Monitoring Poverty

Mengxiao Liu¹, Yong Ge¹*, Jianghao Wang¹, Shan Hu¹

¹ State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, China

*Corresponding author: gey@lreis.ac.cn

ABSTRACT

Poverty reduction is the most urgent task for China in the building of a well-off society in an all-round way. The Outline for Development-oriented Poverty Reduction for China’s Rural Areas (2011–2020) delineates total 832 poor counties in China. In order to evaluate the precision of poverty alleviation and exit assessment, the common practice is to select the poor villages in poor counties to carry out questionnaire surveys to evaluate and adjust the early precision recognition results. In the process of collecting samples to monitor poor villages, it is necessary to divide the different regional types according to the characteristics of the poverty-stricken village, and to design a village sampling based on the types of different regions and the representative and characteristic of the sample. Spatial sampling takes into account the spatial distribution characteristics of the object to be monitored. In this paper, SBK(Stratified block kriging) and MSN(Mean of Surface with Nonhomogeneity) method is adopted to study the spatial sampling optimization design for monitoring poverty in Yunyang county in Hubei Province. Firstly, poverty head count ratio data is collected to denote poverty and auxiliary data including population, GDP, topology, land use map, nightlight data are collected to help stratify poverty in space. Then the two methods SBK and MSN is performed to get the spatial sampling optimization design.

KEYWORDS: Spatial sampling; Poverty; Stratified Block Kriging; MSN
Registration of Point Cloud Data Based on Principle of Least-Squares

Yuefeng Lu*, Yong Ge, Kun Chen, Yangyang Fu

1 School of Architecture and Engineering, Shandong University of Technology, Zibo 255049, China
2 State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101 China
3 College of Urban and Environmental Science, Central China Normal University, Wuhan 430079, China

*Corresponding author: yflu@lreis.ac.cn

ABSTRACT

With the rapid development of information technology, 3D laser scanning technology has developed rapidly as an emerging measurement method. The technology of three dimensional laser measurement has a lot of applications in many fields, including reverse engineering, cultural relic protection, 3D printing and so on. Accordingly, the processing of point cloud data measured by three-dimensional laser measurement is becoming more and more important. In the processing of point cloud data, registration is an important step that cannot be indispensable. The essence of point cloud registration is that by determining the coordinate transformation parameters between common points in the point cloud data containing the repeated regions under different scanning coordinate systems obtained by scanning sampling, all the non-common points are converted into unified coordinates system. This is the same as the theory of coordinate transformation. Therefore, we apply the 3D coordinate transformation to the point cloud registration and use the coordinate transformation method to find the transformation parameters of the coordinates of the common points contained in the two sets of repeating regions, and then complete the non-public Registration of point cloud data. In the model of coordinate transformation, the method of seven-parameter is the most widely used. However, this traditional seven-parameter method ignores a number of mutual multiplicative items when the rotation angle is very small. The elements of the rotation matrix are all composed of a sine function and a cosine function, so when faced with a small rotation angle, the elements in the rotation matrix tend to be similar. The sine similarity is 0, the cosine similarity is 1, and the corresponding scale parameters and the The relationship between the partial products is also ignored. In face of a large angle of rotation, the estimated parameter estimates will deviate from these approximations and may even cause distortion. For the reason that, how to realize the coordinate transformation of arbitrary rotation angle is of great significance. In point cloud data registration, the problem of large rotation angles is often encountered, and the traditional seven-parameter model cannot accurately perform coordinate transformation of large rotation angles. Therefore, how to realize the application of three-dimensional coordinate transformation in point cloud registration, that is, to realize coordinate transformation of arbitrary rotation angle has important research significance. The main work of this article can be summarized as follows:

(1) The theory of Least-Squares Principle is analyzed, including its function model and random model. The Least-Squares (LS) is a mathematical solution to the group of nonlinear equations. The observation vector is taken as the estimated parameter to solve the three-dimensional coordinate transformation parameter. The principle of Least-Squares and its optimization criterion are the most commonly used models of function in the process of solving the optimal estimation of parameters. The optimal adjustment criterion of Least-Squares is introduced. Under this criterion, different adjustment methods can be constructed for different problems. The indirect adjustment method is a way commonly used to solve the undetermined parameters under the requirements of the adjustment criterion. The method of indirect adjustment with restricted conditions is put forward,
and the formula for solving the problem under the optimization criterion which satisfies the principle of Least-Squares is given.

(2) The shortcomings of the traditional method of seven-parameter in coordinate transformation are analyzed. In face of coordinate transformation with a large rotation angle, the way that seven-parameter method of rotating matrix elements not only ignore some of the multiplication terms, but also ignore the relationship between the corresponding scale parameter and the partial product will lead to some deviation of the parameters of the model solution, and even lead to the distortion of the model. Aiming at the shortcomings of the method of seven-parameter in the face of coordinate transformation with large rotation angle, a new research on the transformation model is carried out. The seven parameters include a scaling parameter, three translation parameters of the origin of the coordinates, and three angles of rotation of the three axes. In order to realize the coordinate conversion of arbitrary rotation angle, the method of seeking no rotation angle is adopted, and the nine elements in the rotation matrix are all regarded as an unknown parameter to be solved, and the seven-parameter model is converted into a thirteen-parameter model. This avoids the elements in the rotation matrix being similarly processed during the solution.

(3) The solution of coordinate transformation model is analyzed. For the case that the models of coordinate transformation are all nonlinear, the nonlinear seven-parameter are linearized and the Taylor series expansion method is used to obtain the linearized model. When the corresponding control point is equal to or greater than 3 pairs, the model of coordinate transformation can be constructed. The rotation matrix is an orthogonal matrix, which satisfies the condition of orthogonal matrix, that is, the nine elements of the rotation matrix satisfy the corresponding nonlinear conditions. Aiming at the established model of coordinate transformation, proposing a method which called indirect adjustment that contains restrictive conditions to solve the transformation parameters. The linearized model is expressed in the form of indirect adjustment with constraints. The mathematical model of indirect adjustment with constraints is composed of error equation and constraint equation. The constraint is that the rotation matrix should satisfy the conditions of the orthogonal matrix. The parameters are solved by the Least-Squares Estimation. The specific process of solving the parameter needs to be iteratively calculated, the initial value of the iteration is set by ourself. We will be used to make the first calculation of the rotation matrix is set to the unit matrix, the translation parameters in all three directions are set to 0, and the scale parameter is set to 1, the iterative calculation until the parameters satisfy the convergence requirements. Using the point cloud data collected by a subway tunnel in a certain city for registration experiments, using the thirteen-parameter coordinate transformation model proposed above, according to the coordinates of common targets belonging to different coordinate systems in the point cloud data of two sets of repeated regions, Find their coordinate conversion parameters and then complete the registration of all non-common point cloud data. The experimental results of point cloud coordinate registration are analyzed, and it is verified that the three-dimensional coordinate conversion method is feasible and effective in point cloud registration.

**KEYWORDS:** Point cloud; Least squares; Three-dimensional coordinate transformation
Validation of High Resolution Global Land Cover Dataset: Methodology and Practice

Huan Xie 1, Wen Meng 1, Fang Wang 1, Ang Zhao 1, Yali Gong 1, Xiaohua Tong*1

1 College of Surveying and Geo-Informatics, Tongji University, China

*Corresponding author: xhtong@tongji.edu.cn

ABSTRACT

The purpose of this paper is to present an approach for validation of high-resolution global land cover datasets, which includes global- and local-scale sampling strategy, two-rank sampling plan, spatial allocation of samples, degree of trust in sample judgment, and accuracy assessment. In each scale, sample size is calculated by the use of the optimized sampling model. In the sample allocation of global level, the density of land-cover data and effective land area were considered. In the sample allocation of regional level, Moran’s I was used to reduce the correlation between samples by analysis of local spatial relations. An operational flowchart for global validation is proposed, with the aim to provide quality information of the product. Meanwhile, invalid samples analysis was conducted to improve the control of degree of confidence and reduce judgmental error in future work. And the factors which may affect the accuracy of globe land cover datasets were discussed, such as classification methods, global eco-region distribution, global phenological conditions, and original TM images' quality, reference data and external factors.

As an example, GlobeLand30 was validated by the use of the proposed approach. A total of 80 map sheets and over 15000 pixel samples are selected for final accuracy assessment by confidence level of sample. The result of the validation was reported.

KEYWORDS: Global Land Cover, Validation, GlobeLand30, Accuracy

ACKNOWLEDGEMENTS: This work is supported by National Natural Science Foundation of China (Project No. 41631178), and the International S&T Cooperation Program of China (Project No. 2015FDA11360).