UNIVERSITY OF CALIFORNIA

Santa Barbara

Identifying Geographical Features with Spatial Data: Multi-scale Approaches for Representing Local Extrema

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Geography

by

Boleslo Edward Romero

Committee in charge:

Professor Keith C. Clarke, Chair

Professor Michael F. Goodchild

Professor Phaedon C. Kyriakidis

June 2019

The dissertation of Boleslo Edward Romero is approved.

Professor Michael F. Goodchild

Professor Phaedon C. Kyriakidis

Professor Keith C. Clarke, Chair

June 2019

Identifying Geographical Features with Spatial Data:

Multi-scale Approaches for Representing Local Extrema

ACKNOWLEDGMENTS

I would like to thank all of the people who supported me in completing my dissertation and doctoral program. Professor Keith C. Clarke provided a perfect example of steady support and guidance. He helped me to chart a path across countless scholarly activities and provided guidance through to the completion of my program. Professor Michael F. Goodchild, Professor Phaedon C. Kyriakidis, and Dr. Geoffrey M. Jacquez provided considerable time and energy to review my work and to offer significant and thoughtful advice. As a whole, the University of California, Santa Barbara (UCSB) Department of Geography provided a solid foundation for my success. Together, the faculty, staff, and students created an inspiring environment. In addition, several organizations and sponsors aided my scholarship through both funding and research experience. I am grateful for the support of the University of California, Santa Barbara; the USGS Center of Excellence for Geospatial Information Science; the UCSB Climate Hazards Center; Jack and Laura Dangermond; the California Institute for Research on Hazards; and the Center for Spatio-temporal Thinking and Computing and Applications. I appreciate those who provided mentorship and encouragement, especially Professor Krzystof Janowicz and the community involved in NSF and USGS funded GeoVoCamps, Dr. Chris Funk of the UCSB Climate Hazards Center; and Professor Donald Janelle of the UCSB Center for Spatial Studies. I would also like to express my deepest appreciation for the love and support of my family, my mother, my father, and my sweetheart Tammy.

Curriculum Vitae

Boleslo E. Romero
Ph.D. Candidate
Department of Geography
University of California, Santa Barbara
http://www.geog.ucsb.edu/~bo_romero
bo_romero@geog.ucsb.edu

Education

Present-2010 *Ph.D. in Geography,* University of California, Santa Barbara.
Dissertation: Identifying Geographical Features with Spatial Data: Multi-scale Approaches for Representing Local Extrema

2010-2008 *Master of Science in Geography,* Michigan State University.
Master's Thesis: Toward the Detection of Landscape Features: Clustering 3D Points using Spatial and Thematic Characteristics. ProQuest UMI Number: 1483477.

2008-2005 *Bachelor of Science in Geography, cum laude,* University of New Mexico.
Sr. Honors Thesis: Potential Locations for Wind Energy Generation in Eastern New Mexico.

Academic Honors

Golden Key International Honour Society: Academic Honor Society.
Phi Kappa Phi: Academic Honor Society.
Gamma Theta Upsilon: Honor Society in Geography.
Departmental Honors in Geography, Sr. Honors Thesis, University of New Mexico.
Dean's List: Spring 2008, Fall 2007, Spring 2007, University of New Mexico.

Fellowships and Scholarships

2015-2016 *Jack and Laura Dangermond Graduate Fellowship*, University of California, Santa Barbara.
2014-2010 *Four-year Graduate Student Support*, University of California, Santa Barbara.
2013 *Graduate Student Fellowship*, California Institute for Research on Hazards.
2011 *Predoctoral Fellowship*, alternate, Ford Foundation.
2010 *Graduate Opportunity Fellowship*, University of California, Santa Barbara.
Predoctoral Fellowship, honorable mention, Ford Foundation.

Refereed Journal Articles

2017 **Romero, B. E.**, & Clarke, K. C. Exploring uncertainties in terrain feature extraction across multi-scale, multi-feature, and multi-method approaches for variable terrain. *Cartography and Geographic Information Science*, published online 7/18/17, DOI: 10.1080/15230406.2017.1335235.

2016 Clarke, K. C., & **Romero, B. E.** On the Topology of Topography: A Review. *Cartography and Geographic Information Science*, 44(3):271-282, DOI: 10.1080/15230406.2016.1164625.

2015 Sinha, G., Kolas, D., Mark, D. M., **Romero, B. E.**, Usery, E. L., Berg-Cross, G., & Padmanabhan, A. Surface Network Ontology Design Patterns for Linked Topographic Data. *Under open-review in the open-access Semantic Web Journal*

2014 **Romero, B. E.**, & Bray, T. L. Analytical Applications of Fine-scale Terrestrial Lidar at the Imperial Inca Site of Caranqui, Northern Highland Ecuador. *World Archaeology*, 46(1):25-42, DOI: 10.1080/00438243.2014.890910

Refereed Book Chapters

2014 Sinha, G., Mark, D., Kolas, D., Varanka, D., **Romero, B. E.**, Feng, C., Usery, L. E., Liebermann, J., Sorokine, A. An Ontology Design Pattern for Surface Water Features. In: Duckham, M., Pebesma, E., Stewart, K., Frank, A. U. (Eds.) *Geographic Information Science, 8th International Conference, GIScience 2014*, Lecture Notes in Computer Science, Vol. 8728, Vienna, Austria, 9/23/14 - 9/26/14, Springer International Publishing, Switzerland, 187-203. DOI: 10.1007/978-3-319-11593-1_13.

Refereed Conference Proceedings

2017 **Romero, B. E.**, & Clarke, K. C. Patterns of terrain feature extraction from variable terrain using multiple methods. *28th International Cartographic Conference of the International Cartographic Association (ICC 2017)*, Washington D.C., 7/1/17 - 7/7/17.

2014 **Romero, B. E.** Spatial Outlier Detection of Gaussian Shapes. *11th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Spatial Accuracy 2014)*, East Lansing, MI, 7/8/14 - 7/11/14.

2013 **Romero, B. E.**, & Clarke, K. C. The Impact of Algorithm, Relief, and Resolution on Drainage Network Extraction from Digital Terrain. *Proceedings of the Cartography and Geographic Information Society / American Society of Photogrammetry and Remote Sensing 2013 Specialty Meeting (CaGIS/ASPRS 2013)*, San Antonio, TX, 10/30/13.

2012 **Romero, B. E.** 3D Modeling of an Inca Site with Fine-scale Terrestrial LiDAR. *17th International Symposium on Automated Cartography (AutoCarto 2012)*, Columbus, OH, 9/17/12.

Refereed Posters at Professional Meetings

2017 Sinha G., Arundel S., Stewart K., Mark D.M., Hahmann T., **Romero B.E.**, Sorokine A., Usery L.E., McKenzie G. A Reference Landform Ontology for Automated Delineation of Depression Landforms from DEMs. *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017)*, L'Aquila, Italy, 9/4/17 - 9/8/17.

2012 **Romero, B. E.**, & Shortridge, A. M. Geographical Feature Detection: Accuracy Assessment, Metrics, and Visual Analytics. Presented at the *7th International Conference on Geographic Information Science (GIScience 2012)*, Columbus, OH, 9/20/12.

Professional Meeting Competition and Award

2013 **Romero, B. E.**, Chen, H., Chen, M., & Karaffa, B.; Clarke, K. C., faculty advisor. Time-enabled Web-mapping of Flood Risks: Finding Evacuation Routes for Hurricane Ike. Awarded Second Place in the Student Map Competition, including a map, a paper, and a presentation at the *Cartography and Geographic Information Society / American Society of Photogrammetry and Remote Sensing 2013 Specialty Meeting (CaGIS/ASPRS 2013)*, San Antonio, TX, 10/30/13.

Research Projects

2018-2016 ***USGS Center of Excellence for Geospatial Information Science***. Semantic web terrain prototype.

2015-2012 ***USGS Center of Excellence for Geospatial Information Science***. Topographic ontologies and evaluation of deriving digital elevation model surface features with various algorithms, spatial resolutions, and terrain relief. Ontological design patterns for surface networks.

2015-2008 ***LiDAR Data Collection and Analysis at an Inca Archaeological Site in Caranqui, Ecuador***. Lidar site mapping and total station survey for site documentation, visualization, and water flow analysis.

2013-2012 ***UCSB Climate Hazards Group*** */ FEWS Net / USGS / NOAA / USAID.* Interpolation of 100+ years of rainfall station data and subsequent satellite data adjustment for analysis of global rainfall, agricultural production, and food security risk assessment.

2009 ***University of New Mexico LiDAR Lab***. Volunteered and assisted with lidar projects related to sensitive dune changes in the Mancos, CO, Colorado State Park and also related to endangered species habitat of the Rio Grande silvery minnow.

2007 ***LiDAR Data Collection Pilot Project at Petroglyph National Monument****. (2007)*. National Park Service Research Project (PETR-2007-SCI-0003), Principal Investigator. Lidar mapping of petroglyphs for analysis, accessibility, and archival.

Research Meetings and Workshops

2017 *GeoVoCamp SOCoP 2017: General Session*, at the United States Department of the Interior, Washington D.C., online, 11/28/17 – 11/30/17.

USGS ***Center of Excellence for Geospatial Information Science Research Meeting***. Online conference / presentations, 6/13/17 - 6/14/17.

2016 *GeoVoCamp SOCoP 2016:* ***Gradient/Depressions Ontology Design Pattern Working Group***, at the University of Maryland, College Park, MD, online, 11/30/16 - 12/1/16.

2015 *GeoVoCamp SOCoP 2015:* ***Topographic Eminence Ontology Design Pattern Working Group***, at the United States Geological Survey (USGS), Reston, VA, 12/1/15-12/3/15.

USGS ***Center of Excellence for Geospatial Information Science Research Meeting***. Online conference / presentations, 6/9/15 - 6/11/15.

2014 *USGS* ***Center of Excellence for Geospatial Information Science Research Meeting****,* at the United States Geological Survey (USGS), Rolla, MO, 6/24/14 – 6/26/14.

GeoVoCamp Santa Barbara 2014: ***Map Legend Ontology Design Pattern Working Group****,* at the University of California, Santa Barbara, Santa Barbara, CA, 3/14/14 - 3/16/14.

2013 *GeoVoCamp SOCoP 2013:* ***Surface Water Ontology Design Pattern Working Group****,* at the National Science Foundation, Arlington, VA, 11/18/13 – 11/19/13.

USGS Center of Excellence for Geospatial Information Science Research Meeting. Online conference / presentations, 6/25/13 – 6/26/13.

GeoVoCamp Santa Barbara 2013: ***Map Ontology Design Pattern Working Group***, at the University of California, Santa Barbara, Santa Barbara, CA, 3/21/13 - 3/22/13.

2012 *GeoVoCamp SOCoP 2012:* ***Terrain Ontology Design Pattern Working Group****,* at the United States Geological Survey, Reston, VA, 11/29/12 - 11/30/12.

Professional Meeting Presentations

2015 **Romero, B. E.** An Evaluation of Sampling Methods for Spatial Outlier Detection. Presented at the *Association of American Geographers 2015 Annual Meeting*, Chicago, IL, 4/22/15.

2014 **Romero, B. E.** Spatial Outlier Detection of Gaussian Shapes. Presented at the *11th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Spatial Accuracy 2014)*, East Lansing, MI, 7/8/14 – 7/11/14.

Romero, B. E., Clarke, K. C. Empirical Surface Network Analysis: Effects of Algorithms, Relief, and Resolution. Presented at the *Association of American Geographers 2014 Annual Meeting*, Los Angeles, CA, 4/12/14.

2013 **Romero, B. E.**, Clarke, K. C. Comparison of Stream Channel Estimation. Presented at the *Association of American Geographers 2013 Annual Meeting*, Los Angeles, CA, 4/11/13.

Pedreros, D., Funk, C., Peterson, P., Landsfeld, M., Verdin, A., Husak, G., Michaelsen, J., & **Romero, B. E.** The FEWS NET's Rainfall Enhancement Process. Presented at the *Association of American Geographers 2013 Annual Meeting*, Los Angeles, CA, 4/13/13.

2011 **Romero, B. E.** 3D Point Segmentation Using Spatial and Thematic Data. Presented at the *Association of American Geographers 2011 Annual Meeting*, Seattle, WA, 4/13/11.

2010 **Romero, B. E.** 3D Point Segmentation and Feature Representation. Presented at the *Association of American Geographers 2010 Annual Meeting*, Washington D.C., 3/18/10.

2009 **Romero, B. E.** High Resolution LiDAR Mapping and Water Flow at an Inca Archaeological Site in Ecuador. Presented at the *Association of American Geographers 2009 Annual Meeting*, Las Vegas, NV, 3/27/09.

Romero, B. E. High Resolution LiDAR Mapping and Representation of an Inca Archaeological Site in Ecuador. Presented at the *Michigan Academy of Science, Arts, and Letters 2009 Annual Meeting* at Wayne State University, Detroit, MI, 3/20/09.

ACADEMIC TEACHING and SUPPORT

UC Online, University of California Office of the President

2016, Fall GEOG W12, **Maps and Spatial Reasoning**, Teaching Assistant
2016, Spring GEOG W12, **Maps and Spatial Reasoning**, Teaching Assistant
2015, Fall GEOG W12, **Maps and Spatial Reasoning**, Teaching Assistant
2015, Spring GEOG W12, **Maps and Spatial Reasoning**, Teaching Assistant
2014, Fall GEOG W12, **Maps and Spatial Reasoning**, Teaching Assistant

Department of Geography, University of California, Santa Barbara

2011, Fall GEOG 176A, **Introduction to GIS**, *Lab Instructor, Developer*

Department of Geography, Michigan State University

2010, Spring GEO 324-V, **Remote Sensing of the Environment**, *Virt. Univ. Instructor*

2009, Fall GEO 324, **Remote Sensing of the Environment**, *Lab Instructor (x2)*

2009, Spring GEO 424, **Advanced Remote Sensing**, *Lab Instructor*

2009, Spring GEO 203, **Introduction to Meteorology**, *Teaching Assistant*

2008, Fall GEO 324, **Remote Sensing of the Environment**, *Lab Instructor*

2008, Fall GEO 203, **Introduction to Meteorology**, *Teaching Assistant*

ACADEMIC SERVICE and BROADER IMPACTS

Affiliations

Association of American Geographers.

American Society of Photogrammetry and Remote Sensing.

Cartography and Geographic Information Society

Gamma Theta Upsilon: Honors Society in Geography.

Golden Key International Honour Society: Academic Honor Society.

Phi Kappa Phi: Academic Honor Society.

Official Positions

2014-2013 *UCSB American Society of Photogrammetry and Remote Sensing Student Chapter* Vice-president and President.

2013-2012 *UCSB American Society of Photogrammetry and Remote Sensing Student Chapter* Vice-president and founding member.

Professional Conference Service

2015 Chair, Paper Session 2511, GIS: Data, Methods, and Models, at the *Association of American Geographers 2015 Annual Meeting*, Chicago, IL, 4/22/15.

2014 Chair, Paper Session 5156, Geospatially Integrated Modeling and Mapping Scale 2 Sensitivity Analysis, at the *Association of American Geographers 2014 Annual Meeting*, Los Angeles, CA, 4/12/14.

Academic Committees

2014-2015 *Computing Committee, UCSB Department of Geography.*

2010-2012 *Computing Committee, UCSB Department of Geography.*

Journal Manuscripts Reviewed

2018 Manuscript for *Cartography and Geographic Information Science*, 5/1/18.

2017 Manuscript for *Cartography and Geographic Information Science*, 11/20/17.

Manuscript for *Cartography and Geographic Information Science*, 1/24/17.

2014 Manuscript for *Cartography and Geographic Information Science*, 3/10/14.

2009 Manuscript for *Applied Geography*. Co-reviewed with Shortridge, A. M., 11/18/09.

Conference Manuscript Reviewed

2009 Paper for the *2009 Applied Geography Conference*, 6/30/09.

Public Presentations

2011 **Romero, B. E.** Geographical Uncertainty: Evaluation of Outliers. Presented at the *University of California, Santa Barbara, Geography Colloquium*, Santa Barbara, CA, 10/27/11.

2010 **Romero, B. E.** 3D Modeling with Terrestrial LiDAR. Presented at the *University of California, Santa Barbara, Maps and Spatial Reasoning Guest Lecture*, 10/29/10.

Romero, B. E. New Dimensions of Landscape Mapping. Presented at the *UCSB Geography Awareness Week Community Guest Lecture*, at the San Roque School (now Garden Street Academy), Santa Barbara, CA, 11/17/10.

2009 **Romero, B. E.** 3D Modeling with Terrestrial LiDAR. Presented at the *Michigan State University Geography Graduate Group Spring Colloquium*, 2/6/09.

Romero, B. E. Interactive Demonstration of Terrestrial LiDAR. Presented at the *Wayne State University Anthropology Department*, Detroit, MI, 2/13/09.

2007 **Romero, B. E.**, Cabrera, M., Padilla, R., Valle, J., & Garcia, J. Approaches to Success: Lifelong Learning. Presented at Eldorado High School, Albuquerque, NM, 4/9/07.

EXTRA-ACADEMIC EXPERIENCE

Internships

2007-2008 UNM: GEOG 470, Concepts of Applied Geography
Environmental Considerations of
Local and Regional Transportation Planning and Engineering
UNM: C&J 492, Internship in Communication
Workplace Training
UNM C&J 491, Internship in Communication Education
Small Group Facilitator

Computational Experience

Proficiency:

R Language for Statistical Computing and Graphics
ESRI ArcGIS products
GRASS GIS
Python
Matlab
Inkscape
GNU Gimp
Adobe Photoshop
Adobe Illustrator
AutoCAD Civil3D
Innovmetric Polyworks
Survey equipment/software (i.e. GPS, RTK, total station, terrestrial lidar)
Canvas Learning Management System (LMS)
Blackboard LMS
ANGEL LMS
Moodle Open-source LMS
Emacs

Working knowledge:

Many free and open-source GIS
Geographic databases (e.g. MySQL, PostgreSQL/PostGIS)
Geographic services (e.g. WMS, WFS, etc.)
GDAL library
PROJ.4 library
Meshlab
Blender
Windows Server
Linux Ubuntu
ERDAS Imagine
ENVI / IDL
SPSS
SAS
STATA
HTML/CSS
Adobe Flash
Google Earth
Google SketchUp

Familiarity:

Point Cloud Library (open source lidar tools)
OpenCV (open source computer vision tools)
ffMpeg (audio and video processing tools)
JavaScript
Java
C/C++

ABSTRACT

Identifying Geographical Features with Spatial Data:

Multi-scale Approaches for Representing Local Extrema

by

Boleslo Edward Romero

This dissertation concerns the properties and relationships of discernible geographical features or their parts, particularly local extrema. As distinctive cases of rapid change, their local variation imbues a high degree of uncertainty. Scale is involved with this uncertainty, partially by generalization with samples, but also by cross-scale variation of spatial dependence. This research investigates whether geographical features and their parts are classifiable by attribute values and also by patterns of spatial dependence with respect to scale. The first chapter, on surface network features, evaluated classifications of terrain data as either peak, pit, pass, ridge, or course features, all local extrema. Results were reviewed with regard to spatial resolution, terrain variability, and algorithms. Large differences in algorithm results were nearby potential features. Quantitative measures found smaller differences for the crisp features of high variability terrain compared to ambiguous low variability terrain. Due to multi-scale characteristics, every location had a degree of membership in every feature class. Membership values enabled the extraction of dominant features and a quantification of uncertainty. The second chapter focuses on spatial outliers, similar to peaks as locally extreme values. A controlled study was performed to extract, with three algorithms, spatial outliers simulated as Gaussian forms of various heights and widths. Raster grids of outliers were created with various resolutions and assignment operators. Results varied most by outlier width and spatial resolution. The algorithms missed the top regions of wide outliers, spanning multiple raster cells, likely due to the presence of high local spatial autocorrelation with a mismatch between the scale of analysis and the scale of an outlier. The third chapter investigated whether

non-random patterns of high local spatial autocorrelation exist in wide outliers. Simulations of sets of outliers in variable fields enabled a quantitative comparison with regard to various outlier shapes, fields, and methods. Results of three common random sampling strategies were compared to another method that employed higher probabilities for locations with high local spatial autocorrelation. The latter method resulted in higher rates of both samples on outliers and unique outliers found. Intermediary data revealed patches of high spatial autocorrelation around the outlier tops. The fourth chapter evaluated characteristics of parts of wide spatial outliers. With similar synthetic outliers and fields, patterns of local spatial autocorrelation were compared across classes of the top, side, and base parts of outliers. Small samples and correlated proxy variables were also considered. For samples in each class, a novel multi-scale Local Moran's I "distogram" was computed: a series of values representing the local spatial autocorrelation within each of several non-overlapping spatial bands outward from the point of analysis. The results indicated that the top and side classes have distinctive signatures, while the base co-mingles with the background. In challenging scenarios of small outliers in highly variable fields, differentiation was maintained in bands at about the scale of an outlier. Small samples and proxy variables maintained various degrees of distinction between the signatures. In conclusion, this dissertation investigated the properties and relationships of discernible geographical features and spatial outliers with special regard to representation across scales. Multi-scale information indicates a potential for multiple feature classes at every location. Controlled experiments indicate limitations of spatial outlier detection techniques if the scale of analysis does not match the scale of the feature. Finally, distinctive patterns of local spatial autocorrelation were found for parts of spatial outliers. This research provides empirical evidence that broad-scale local variation involves spatial dependence at a finer scale. As such, this research informs the identification of geographical features or their parts by their variation and spatial dependence across scales.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Context and subject of investigation

Science increases our understanding of the world through observation and analysis. This dissertation research is based in the scientific discipline of Geography, more specifically Geographical Information Science, or GIScience (Goodchild, 1992). This specialty is concerned with processes of understanding the geographical world and examining a variety of issues that arise from the observation and analysis of geographical data.

One of the core concerns of GIScience is the study of geographical uncertainty (J. Zhang & Goodchild, 2002). Although many aspects of geographical uncertainty are already identified, there remains a need for further research. The University Consortium on Geographic Information Science (UCGIS) Research Agenda identified "Research Issues on Uncertainty in Geographic Data and GIS-Based Analysis" (McMaster & Usery, 2005). The U.S. Geological Survey (USGS) Geospatial Information Office also suggested uncertainty-related research with the question "What roles do scale, resolution, and uncertainty of scientific information play in addressing different types of issues?" (National Research Council, 2006). The New Directions for Research, by the National Geospatial-intelligence Agency (NGA), included issues of "validation, data quality, spatial uncertainty" (National Research Council, 2010). The previous examples are evidence that top-level authorities of geographical knowledge have interest in resolving issues of geographical uncertainty.

Falling within the broad scope of geographical uncertainty, investigations of this dissertation involved various aspects of spatial scale. In short, this research focused on the extraction of portions of spatial data that represent geographical features for mapping

(Robinson, Morrison, Muehrcke, Kimerling, & Guptill, 1995). The aim was to locate defined features. The following is an overview of the basic process. To begin, consider that a spatial variable is a measurable property that exists throughout a spatial domain. At any point in the domain, the property has a quantifiable attribute. For scientific inquiry and analysis, the variable is measured at a sample set of locations. Each of the observations includes location information and the combined data set represents the variable across space. Next, consider that features are identifiable patterns of variation. Models of features include defined properties, which might include primary values, derived values, or relationships to other features. Then, if characteristics of particular data approximate those of a model, their locations are labeled, or extracted, as particular data that represent features.

Problems arise in the representation of the real variable as a set of spatial data. No matter how the data are collected, perhaps as a simple random sample or as a regular grid, the sample is incomplete. It will miss patterns of variation and not all of the real features will be represented. The scale of analysis is controlled by the distance between sample locations (Nyquist, 1928). Generalization of attribute values between data samples is one of the main concerns of this investigation.

Another consideration is the property of spatial dependence (Tobler, 1970). The theory that nearby values are more similar than distant values provides an opportunity for estimating the missed information. For example, geostatistical interpolation utilizes models of variation to estimate attribute values between samples (Matheron, 1963). Nearby values are modeled as similar to some degree and subject to some assumed spatial decay function.

However, such opportunities provided by the theory of spatial dependence might not necessarily benefit the extraction of features with definitions based upon variation. Geographical features are recognizable, distinctive things embedded in our geographical surroundings. A featureless variable is continuously smooth, of equal value everywhere, and without variation. Introducing only monotonic variation results in what is commonly called a trend. A variable with only a trend does not necessarily have equal values everywhere but does exhibit flatness. Other than the trend's constant rate of change, there are arguably no other distinctive features. Introducing non-stationary variation, different degrees of change across the domain, results in non-monotonic variation and a surface that transitions between high-point peaks and low-point pits. Relative to any given neighborhood, the peaks and pits are local extrema or anomalies that give texture to the variable of interest. Feature classes, such as peaks or pits, are potentially special features with identifiable characteristics of variation. It is also possible for defined sets of features to be required parts that comprise higher-level features (Couclelis, 2010). Different degrees of variation, in special patterns, are the essence of identifiable features.

1.2 Analytical assumptions

This section briefly describes various analytical assumptions employed in this dissertation. A variable attribute has a defined quality. Observations and measurements of the attribute comprise a data set. Attribute values are defined in some scale of measurement, based upon the level of information captured by the measurements and what types of operations the values enable (Stevens, 1946). Containing a high degree of information, such as a equality of ratios, and the property of additivity, the scale of attribute values within this dissertation involve the ratio scale of measurement.

Spatially referenced data include information about the spatial location of each datum, relative to a reference datum and a domain of spatial scale. Spatial data are indexed per location, usually in a Cartesian coordinate system. Various data structures are available for storing and organizing observations. Based upon the field model of geographical data, the initial experiments of this dissertation employed a regular tessellation of square cells organized into a raster grid structure to represent the attribute values spanning across two-dimensional (2D) space. Beyond a raster structure of grid cells, essentially point values, other irregular point sampling strategies were included in later experiments.

To determine if a datum is similar or different from its surroundings, a neighborhood definition is required. Although neighborhoods can be defined in terms of adjacency, or graph-based network connectivity, common distance-based neighborhoods were employed throughout the analyses. Although the distances may be transformed for evaluating different scales of phenomena, particularly with multivariate analyses, this dissertation involves Euclidean distances.

1.3 Research summary

1.3.1 Introduction

This introductory chapter concerns a description of the context and the subjects of the investigation. General concepts of geographical analysis are presented and basic assumptions are also included. Following this introduction, a progression of four separate experimental investigations is presented in chapters 2 through 5. General conclusions about uncertainties of locating features, with respect to methods of representation and

extraction, are in the sixth and final chapter.

The features of interest in this dissertation are different types of local extrema, both points and lines. With relatively extreme attribute values compared to a local neighborhood, they are considered discernible geographical features. Such straightforward definitions are assumed to be readily identifiable and locatable. Beginning with the extraction of geographical features, such as topographical peaks and ridges, the investigations proceed to more general subjects of local extrema. The features were examined at various scales of analysis, both as representations of topographical data and as pairs of synthetic source and sample data.

1.3.2 Surface networks

Chapter 2, on surface network features, evaluated classifications of locations as either peak, pit, pass, ridge, or course features. All of the features classes are related to local extrema in two-dimensional data, in this case representing terrain elevation. The investigation concerned the impacts on feature extraction of three factors, extraction algorithms, terrain variability, and spatial resolution.

1.3.3 Spatial outlier detection

Chapter 3 focuses on the feature class of peaks, which have characteristics of spatial outliers as locally extreme values. In a controlled study of extracting spatial outliers, individual outliers were simulated as Gaussian forms with various heights and widths. The raster grids representing singular outliers were created with various spatial resolutions and by various assignment operators. Three common spatial outlier detection algorithms

provided a range of extraction results.

1.3.4 Spatial dependence of anomalies

Chapter 4 includes an investigation into whether the characteristic of high local spatial autocorrelation exists more than randomly within the extents of spatial outliers. The Local Moran's I statistic was used as a measure of local spatial autocorrelation to provide an indication of clusters of data with similar values. Whether spatial outliers have high local spatial autocorrelation is relative to a surrounding background field. For this controlled study, multiple simulations of sets of outliers embedded at known locations in a variable field. A quantitative comparison was performed across various values of parameters related to the outlier shapes, the variable fields, the Local Moran's I sampling, and the sample size.

1.3.5 Multi-scale spatial dependence

Chapter 5 concerns an evaluation of characteristics of parts of wide spatial outliers, following the ontological concept of composite objects, or features composed of other features with defined relationships. Similar to the third chapter, multiple simulations of sets of outliers embedded in a variable field were employed. In this case, the simulations enabled a controlled study of outlier parts. The top, side, and base parts of a wide outlier were defined by class boundaries at the first, second, and third standard deviations from the center of the Gaussian shape. For those parts, and extended distances into the background field, patterns of local spatial autocorrelation were compared to ascertain whether each of the classes were distinct. The experiment also addressed two other considerations regarding a limited set of data for the spatial variable of interest: knowledge

of only small samples and the potential of using correlated proxy variables.

A novel quantitative multi-scale signature of spatial dependence characteristic involving local spatial autocorrelation was calculated for each sampled location in every class for the variable of interest and for the proxies, separately. The characteristic, a Local Moran's I distogram, was obtained by assembling a series of Local Moran's I statistics calculated from sets of neighboring samples. Each value in the series was calculated from samples within one of several non-overlapping spatial bands defined by boundary distances from the point of analysis.

1.3.6 Summary

The impacts of scale on the results of feature extraction are considered, with regard to the size, shape, and identity of the features, or their parts, and the scale of analysis. Consideration of uncertainties related to scale inform various approaches to address the issues. Across the analyses, spatial scale is addressed in various ways: by combining results of several scales of analysis and by evaluating neighborhood characteristics at various distances. Deeper investigations examined parts of spatial outliers to ascertain potential patterns of spatial dependence with respect to scale. The goal is to better inform the process of feature extraction and potentially reduce scale-based uncertainties encountered with representation of spatial data in order to address a core concern of geographical uncertainty in the context of GIScience.

2 Surface networks

2.1 Introduction

The land surface of the Earth, or terrain, has abundant variation. People observe spatial patterns and identify parts of the terrain as features (Peuquet, 1988; Couclelis, 1992; I. S. Evans, 2012), such as ridges and peaks. Although subject to cultural perspectives (D. M. Mark & Turk, 2003, 2017), the identification of terrain features can support universal needs of people, such as locomotion and way-finding for navigation (Montello, 2005). Another purpose is to better understand processes and interactions among topographic features and water, soils, flora, and fauna (I. Evans, Hengl, & Gorsevski, 2009).

Surfaces such as terrain are often represented as a digital elevation models (DEMs) for quantitative analysis. One type of DEM is a raster, a set of points or cells attributed with elevation values and spatially organized as a regular grid. Such a model of terrain enables the analysis of topographic characteristics across its spatial extents (Miliaresis, 2008; Minár & Evans, 2008). Derived information and higher-level information is needed for the identification of particular instances of basic or composite terrain features (Drăguţ & Blaschke, 2008; Minár & Evans, 2008; Gerçek, Toprak, & Strobl, 2011; Stepinski & Jasiewicz, 2011).

A surface network (Reech, 1858; Cayley, 1859; Maxwell, 1870; Rana, 2004), which is a part of applied mathematics (Morse, 1925), is a topologically-rich model that offers a compact means of storing information about functional surface features. From a surface network, information can be derived for purposes such as flow analysis (Warntz, 1966; Warntz & Woldenberg, 1967). Network nodes are peaks, pits, passes, and pales. Network

edges are ridge lines and course lines. Saddles (passes and pales combined) connect to peaks by ridges and to pits by courses. A tripartite graph is created by connecting the three types of nodes by edges (Pfaltz, 1976). Graph-based analysis (Reeb, 1946) and semantic reasoning about the terrain (D. Mark & Sinha, 2012; Sinha et al., n.d.) is then possible by means of various defined relationships and numerical rules. Surface network components are suitable for representing identifiable features in the terrain (D. Mark, 1977) and a special terrain-based version that represents salient features and their intrinsic relationships is the topographic surface network (D. Mark & Sinha, 2012; Sinha et al., n.d.; Clarke & Romero, 2017).

Surface network feature extraction from raster DEMs of terrain is an ongoing research problem in digital cartography (Clarke & Romero, 2017). Considering that problem, the research question of this experiment is: can surface networks be consistently located with various scales of analysis (i.e. resolutions), various algorithms, and in various terrain types? Of particular concern is the identification and extraction of terrain features in the following set of classes: {peaks, pits, saddles, ridges, courses}. The objective is to consider various aspects of terrain feature uncertainty and explore quantitative representations with potential to better inform terrain analysis (Usery, 1996; MacMillan, Pettapiece, Nolan, & Goddard, 2000; Schmidt & Hewitt, 2004; Shi, Zhu, & Wang, 2005; Cheng, Fisher, & Li, 2004). Addressed throughout the analysis are the variety of results, representations across scales, and the impacts of terrain variability.

Previous work has suggested that due to scale effects, each location has simultaneous fuzzy memberships for several terrain surface feature classes (Wood, 1996; Fisher, Wood, & Cheng, 2004). For example, a broad valley floor is a pit at a coarse scale of analysis, but a small boulder within that region produces a peak at a finer resolution.

To extend this theory, multi-scale feature class membership values were used for a quantitative multi-feature comparison in order to better inform conclusions about the terrain structure and its uncertainty. Previous work that uses multiple spatial attributes for the prediction of soil constituents (Burrough, van Gaans, & Hootsmans, 1997) was adapted to instead evaluate fuzzy terrain features and was also informative with regard to the confusion index metric as a related estimate of terrain feature uncertainty.

Finally, to evaluate whether terrain feature uncertainty is related to the variability of the terrain, three geographic regions were included in this analysis. They were similar with regard to factors such as climate, precipitation, land cover, and erosion but differed with respect to terrain variability. Although numerous metrics exist to describe terrain variability, the standard deviation of elevation values was sufficient to show that the variability of terrain influences the uncertainty of terrain features.

2.2 Methods

2.2.1 Study location

To evaluate the extraction of features in the landscape, topographic surface data was evaluated. The north-eastern portion of Santa Cruz Island, California was selected (Figure 1). Several benefits of this location are that: it avoids complications of built environments and variable land cover; it is represented by a range of digital elevation models (DEMs), including high-resolution lidar data; and, as a portion of the Channel Islands National Park, it has the potential for future multidisciplinary research.

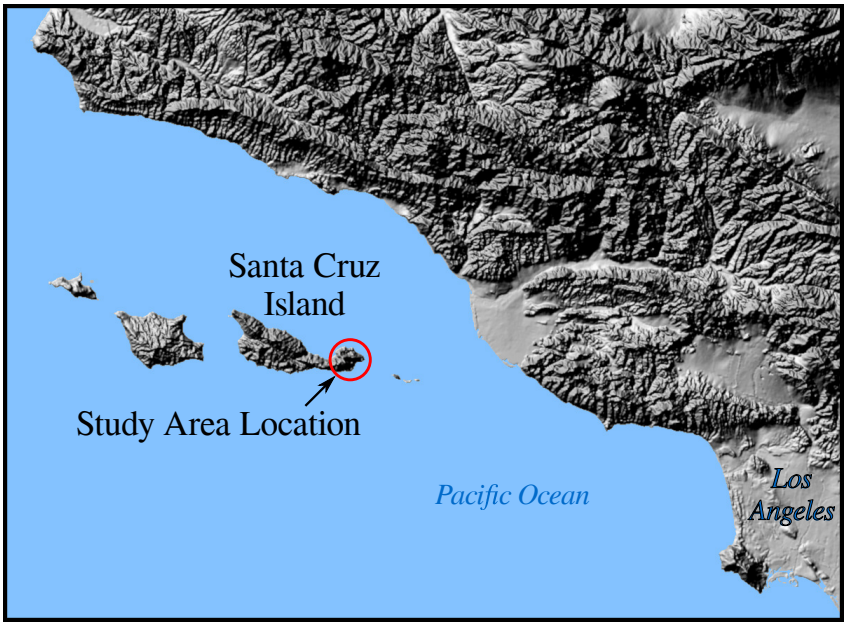


Figure 1: Study area location: the north-eastern portion of Santa Cruz Island, California. (Source: USGS, 2016.)

2.2.2 Sources and resolutions

Four spatial resolutions of raster grid DEMs were collected from the USGS National Map (USGS, The National Map, 2016) and the NSF OpenTopography (OpenTopography, 2010) web portals, of which all DEMs were derived from the same lidar point cloud source data (Arundel et al., 2015). Three resolutions from the National Map were 1-, 1/3-, and 1/9-arc second data, projected to resolutions of 30 meters (m), 10 m, and 3 m, respectively, in the Universal Transverse Mercator (UTM), Zone 11, coordinate system on the 1983 North American Datum (NAD83). The fourth data set was from an OpenTopography high resolution lidar point cloud, pre-processed by the online service. To improve representation of potentially obscured crevices of courses and pits, the minimum elevation values in local areas were used for inverse distance weighting interpolation, creating 1 m raster data in the same coordinate system as the other three DEMs (Figure 2). To prepare data for ridge extraction workflows, all resolutions of the data were inverted

by multiplying the source data by negative one and shifting the resulting elevation values to the original minimum and maximum, so that the original and modified elevation ranges matched.

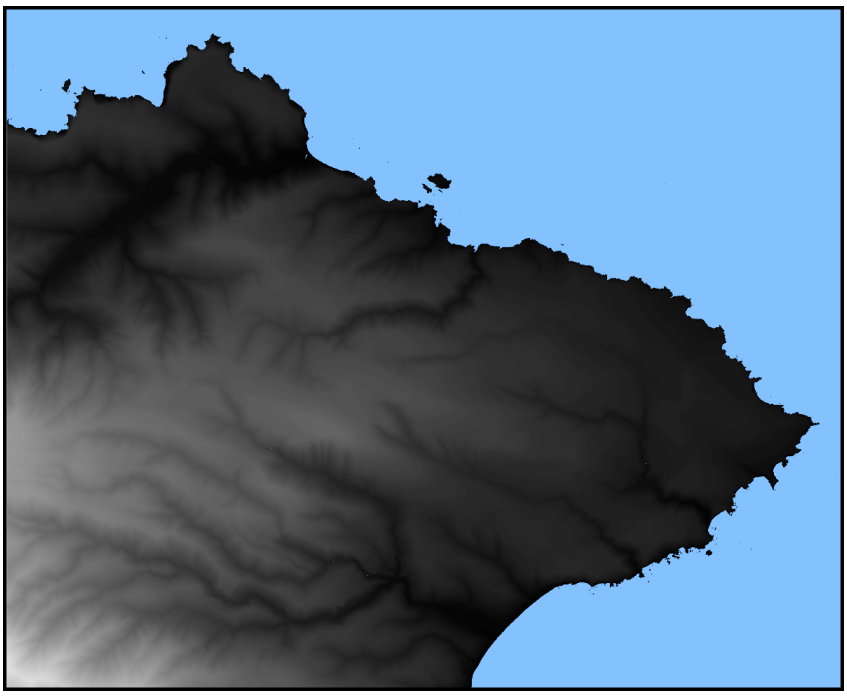


Figure 2: Digital elevation model of the study area location on the northeast portion of Santa Cruz Island, showing high elevations in lighter shades and the Pacific Ocean in blue.

2.2.3 Terrain types

The three regions selected for analysis represented natural terrain of differing degrees of elevation variability, all with similar land cover and subject to similar climate and erosional processes. The size of each region was 900 m x 900 m, allowing for the representation of landscape feature sizes ranging from a few meters to several hundred meters. The 900 m units were also evenly divisible by each of the resolutions (i.e., 1, 3, 10, and 30 m) and the regions were selected to best align with the cell boundaries.

To determine regions that are representative of different terrain types, a 901 m x 901 m moving window analysis was performed over each cell of the 1 m data, excluding locations with edge effects. The standard deviation of the elevation was calculated and each cell was classified as high, medium, or low variability with classification range boundaries, in standard deviation values of vertical meters, specified as: 0, 25, 45, and 88 m (Figure 3). To ensure the inclusion of observed stream features, stream data from the USGS National Hydrography Dataset (nhd.usgs.gov) was overlain to help guide the location selection. Three 900 m x 900 m terrain type regions were subset from the original four elevation data sets of 30 m, 10 m, 3 m, and 1 m resolutions (Figure 4). The spatial extents are listed in Table 1. Hillshaded oblique perspectives of the regions are shown in Figure 5.

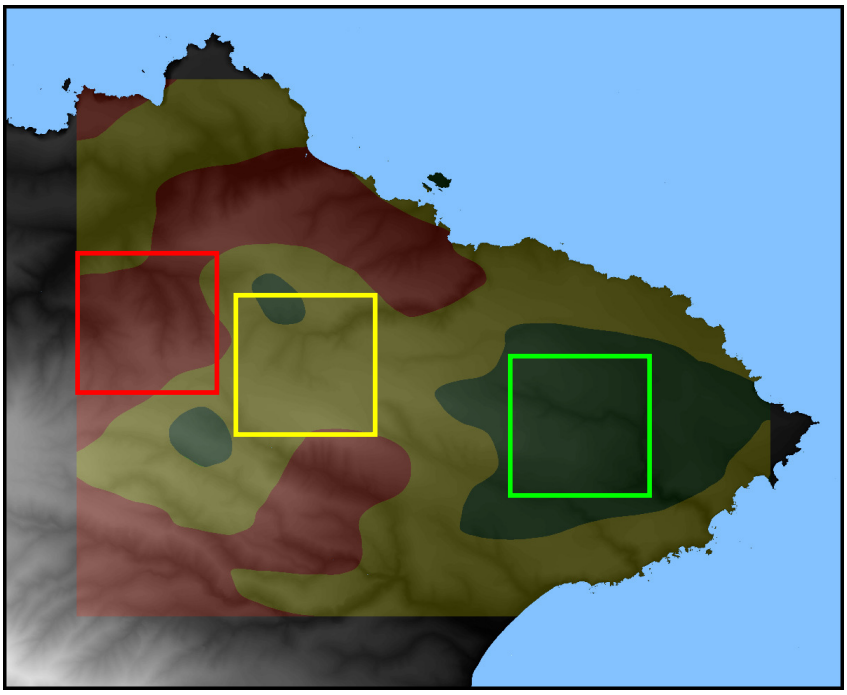


Figure 3: Three terrain type classifications of high, medium, and low variability, shown in red, yellow, and green, respectively. The selected terrain type regions are shown with three correspondingly colored squares.

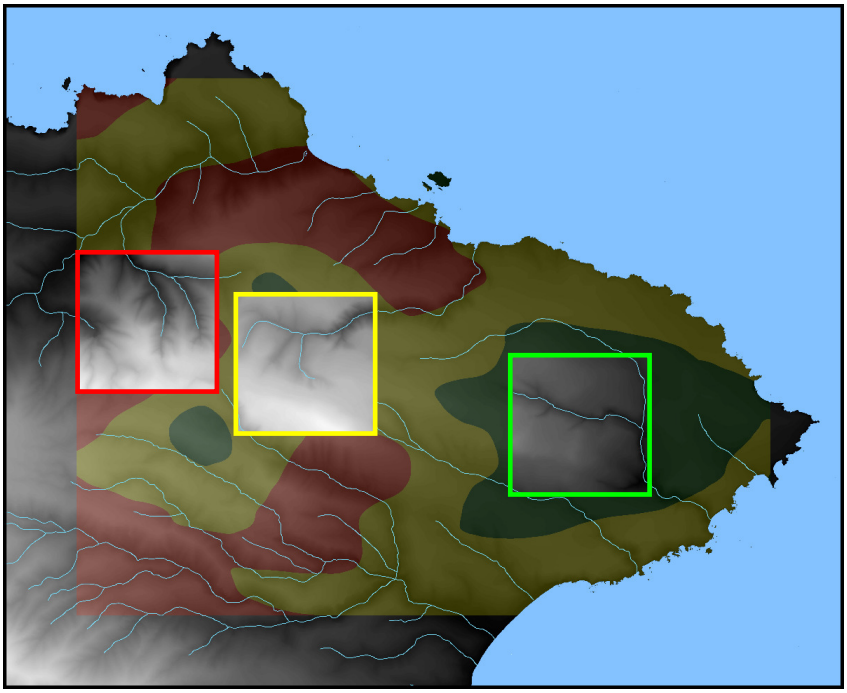


Figure 4: DEMs of three selected terrain type regions, representing high, medium, and low variability, bounded by red, yellow, and green squares, respectively. USGS National Hydrography Dataset stream locations are shown in cyan.

2.2.4 Surface network algorithm

To extract surface network features we first employed the foundational Fowler-Little method (Fowler & Little, 1979). The process finds features using two sets of steps, one set for ridge- and peak-related features and another for course- and pit-related features.

The algorithm performs a moving window analysis with a four-cell neighborhood including the central cell, one to the right, one below, and one diagonal to the right and below. For the set of ridge- and peak-related features, cells are considered not-ridge candidates if they are the lowest elevation in the four-cell neighborhood. The rest are ridge candidates. By a comparison of ridge candidates, those that are lower than their neighbors are ridge passes. Starting from each pass, a ridge-climbing process labels ridge

Terrain Type	Easting Min. (West)	Easting Max. (East)	Northing Min. (South)	Northing Max. (North)
High	262,547	263,447	3,768,993	3,769,893
Medium	263,567	264,467	3,768,723	3,769,623
Low	265,337	266,237	3,768,333	3,769,233

Coordinate System: UTM, Zone 11 N, NAD83

Table 1: Spatial extents of high, medium, and low variability terrain type regions.

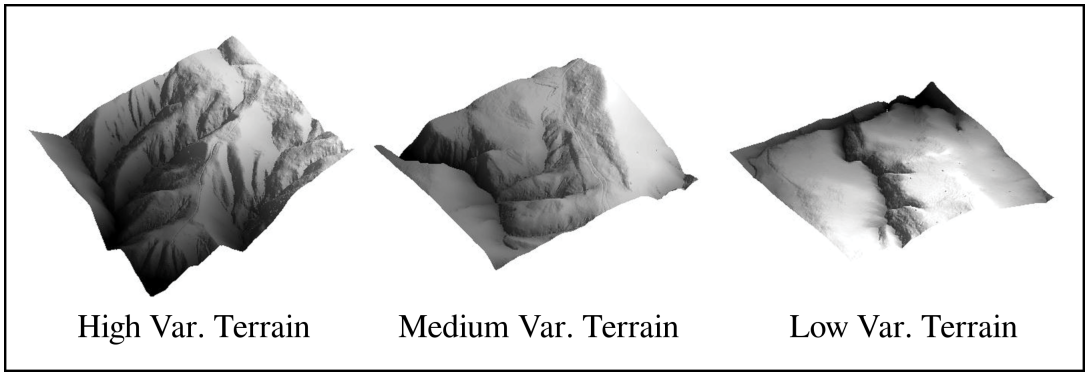


Figure 5: Hillshaded northwest perspective view of the high, medium, and low variability terrain type regions.

cells by following each neighboring ridge candidate with the highest elevation. When a cell has no higher ridge candidate neighbors, it is labeled a peak. An inverse process, with respect to elevation, finds the course- and pit-related features.

2.2.5 Courses and ridges with conventional software

Another approach to extracting courses and ridges is with common conventional methods that aim to determine the location of stream features in the landscape. Although the stream channels are more complex than courses in both reality and in concept, the two share similar characteristics as relatively low elevations between higher regions. In gen-

eral, algorithms meant to derive stream networks numerically approximate the locations of courses in DEMs.

Of the numerous stream extraction algorithms in existence, this study employed a group of closely-related implementations of the D8 single-flow algorithm (O'Callaghan & Mark, 1984). The common workflow for finding streams in a DEM is to fill pits, compute both downslope flow direction and upslope flow accumulation areas for each cell, and apply a global thresholding value to extract the cells with the highest flow accumulation.

Courses and ridges were extracted with methods implemented in a range of popular commercial, free, and open source software packages. The D8-related methods were performed with ArcGIS (ESRI, 2012), GRASS (GRASS Development Team, 2012), LandSerf (Wood, 2007), MapWindow (MapWindow Open Source Team, 2008), and SAGA (SAGA Development Team and User Group Association, Bock, Conrad, Koethe, & Ringeler, 2008). The workflow processes and algorithms used for each conventional software program are shown in Table 2. To compare the methods before thresholding, the differences between each pair of flow accumulation maps for each terrain type and each feature class were computed and locations with the top 95% of differences were extracted (see examples in Figure 6 in 2.3 "Results and discussion"). Finally, to extract course and ridge features, a threshold of the highest 95% of flow accumulation values was used to extract courses. Ridges were extracted with the same procedures, but with the inverted DEMs.

2.2.6 Peaks, pits, and saddles with basic algorithms

To supplement the courses and ridges found with conventional software, peaks, pits, saddles were also extracted. Unlike courses and ridges, the five conventional software

Software	**Menu Items**	**Commands** *(documented algorithms in italics)*
ArcGIS	Spatial Analyst Tools	Fill Flow Direction *(Jenson & Domingue 1988)* Flow Accumulation *(Jenson & Domingue 1988)*
GRASS	Raster / Hydrologic Modeling	Depressionless map and flowlines *(Jenson & Domingue 1988)* Flow accumulation *(Mitasova & Hofierka1993)*
LandSerf	Analyse	Pit removal Flow accumulation
MapWindow	Watershed Delineation / Advanced TauDEM Functions	Fill D8 Flow Direction *(Garbrecht and Martz 1997)* D8 Contributing Area
SAGA	Terrain Analysis / Preprocessing Hydrology	Sink Removal Catchment Area (Parallel) *(O'Callaghan & Mark 1984)*

Table 2: Conventional software programs, menus, and commands employed for computing flow accumulation areas.

programs did not all include methods to identify peaks, pits, or saddles. Instead a 3 x 3 cell moving window "Min-Max" method was used to extract peaks, pits, and saddles from each terrain type DEM using the Python programming language (van Rossum, 2010) and the Geographic Data Abstraction Library (GDAL, 2014). The central cell was labeled as a pit if it was the minimum or a peak if it was the maximum. The central cell was labeled a saddle if it was a maximum of cells across one (horizontal, vertical, or diagonal) linear axis and a minimum in the orthogonal axis.

2.2.7 Multi-scale membership values

The four resolutions of the binary feature maps (of the presence or absence of each feature) were averaged to create multi-scale maps, with cell membership values from 0 to 1, for each terrain type, feature class, and extraction method. Before averaging, coarser resolutions were resampled to the finest resolution. For each terrain type region, this resulted in six multi-scale maps for each of the ridges and courses: five for the conventional stream extraction methods and one for the Fowler-Little method. This also resulted in two multi-scale maps for each of the peaks, pits, and saddles: one for each corresponding "Min-Max" method and one for the Fowler-Little method. In total, for each of the three terrain types, there were 18 multi-scale maps: six course, six ridge, two peak, two pit, and two saddle maps (Figure 7 in 2.3 "Results and discussion").

2.2.8 Multi-scale RMSE differences

For each feature class in each terrain type, the pairwise differences between the methods was computed as the root mean squared error (RMSE) of the multi-scale feature class membership values with Equation 1, where x and y are membership values of corresponding multi-scale feature class maps for any two methods, r is one of the total R rows, c is one of the total C columns, and (r, c) indicates a particular raster cell in the map by its row and column location. Two identical maps would have zero RMSE.

$$RMSE = \sqrt{\frac{\sum_{r=1}^{R}\sum_{c=1}^{C}\left(x_{(r,c)} - y_{(r,c)}\right)^2}{R \times C}} \tag{1}$$

The pairwise RMSEs were grouped into three sets of tables corresponding to the three terrain types. For ridges and courses, the RMSE data were represented as node-and-link graphs using the Python NetworkX package (NetworkX Developers, 2014). Links represent the inverse RMSE to visualize the degree of similarity with thicker lines. Also, for an analysis of each feature type with respect to terrain variability, the RMSE differences in each of the tables were summarized as an average and graphed.

2.2.9 Multi-feature membership values

The multi-scale results produced feature class membership values as fuzzy fields. Every location in the region has the potential of belonging to every feature class, and maybe more likely one feature than another. The multi-scale membership values of the five features were therefore compared at each 1 m cell. For each terrain type, three maps were created from this comparison: one for the dominant feature, one for the second-ranked feature, and one showing the difference between the highest two membership values as a metric of uncertainty.

The feature with the highest membership value is considered dominant at the location. However, there is still some degree of uncertainty. If the dominant feature has a considerably larger membership value than any other feature then there would be less uncertainty than if another feature has a similar or equal membership value. To represent the uncertainty of the dominant feature, the confusion index of Burrough et al. (1997), based upon the difference between the membership values of the two top-ranked features at any location, was considered a straightforward metric. With feature class

membership values ranging from zero to one, the confusion index, also ranging from zero to one, is the difference between the membership values of the dominant and the second-ranked feature, subtracted from one.

For example, consider one cell with a membership value between 0 to 1 for each of the five features. If that cell had membership values within the set {ridge, course, peak, pit, saddle} of {0.50, 0.25, 0.75, 0.00, 0.25}, the cell is found to be dominantly characterized as a peak across all scales of analysis with a likelihood of 0.75. However, the confusion index is relatively high with a value of 0.75. The difference between membership values of the dominant feature, 0.75 for a peak, and the second-ranked feature, 0.50 for a ridge, is a difference of 0.25. Subtracting 0.25 from one yields a confusion index value of 0.75, which is relatively large in the range of zero to one. This indicates a high degree of uncertainty about the dominance of the peak for that cell since the membership value of the ridge is nearly the same.

2.3 Results and discussion

2.3.1 Algorithm differences: Course-related flow accumulation differences

Maps of flow accumulation differences between each of the five extraction processes allow for a visual evaluation of the largest differences between related D8-based algorithms. When using stream extraction algorithms that involve thresholding of flow accumulation there is some bias to consider. Features in not locally but globally high elevations (with low quantities of flow accumulation) are not as often represented as those in globally low elevations (with high flow accumulation).

Figure 6 shows examples of the course-related flow accumulation difference maps: the differences of GRASS vs. LandSerf on the 10 m high variability terrain (left), GRASS vs. LandSerf on the 30 m medium variability terrain (right). The thresholded differences are overlain on the DEMs for visual inspection of the context of locations identified. These are just two examples, but they illustrate that, regardless of terrain type or scale of analysis, the algorithms' results differ most in locations associated with the features of interest, such as courses or ridges.

Such differences indicate that a portion of the uncertainty in feature identification is associated with the various algorithms' implementations of the semantics of the feature characteristics. Only slight differences exist in the closely-related algorithm definitions of the feature properties and relationships, but major effects are evident in the most important locations.

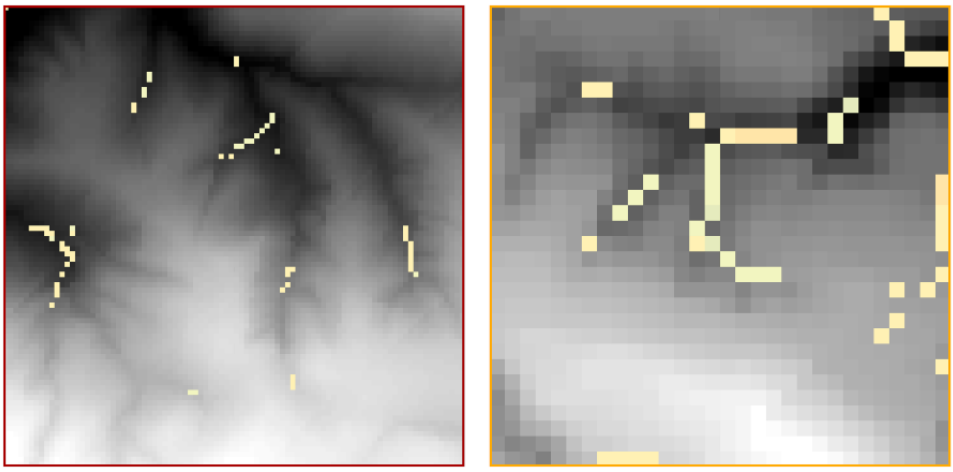


Figure 6: Examples of the course-related flow accumulation difference maps: the differences of GRASS vs. LandSerf on the 10 m high variability terrain (left), and GRASS vs. LandSerf on the 30 m medium variability terrain (right).

2.3.2 Multi-scale membership values

Multi-scale membership values are shown for each feature class in Figure 7. Higher membership values are represented as darker shades and indicate a higher likelihood that a location exhibits characteristics of the corresponding feature across the scales of analysis. Although broad cell resolutions have a major visual effect, all resolutions are weighted equally and high multi-scale membership values exist only in locations where the defining characteristic is also found at finer resolutions.

The location of features extracted from the data is dependent upon generalized data (i.e. pixels assigned discrete values representing variation over an area) and is also dependent upon the scale of analysis. The scale-dependent effects of each resolution are evident in the combined multi-scale maps, and suggest that entirely different surface networks exist at different scales. In turn, the complexity of multi-scale surface networks influences surface-related terrain processes. One implication, for example, is that if a location exhibits characteristics of a course across many scales, then that location has a high potential for convergent flow. On the other hand, a mix of features across different scales makes the terrain-based processes more difficult to ascertain.

The spatial extents of course and ridge features suggests that results can be generally separated into two groups. The first group consists of ArcGIS, SAGA, and MapWindow, which produce finer linear features than the second group of GRASS and LandSerf. Curiously, although GRASS is D8-related, it produces visibly different results than the other D8 methods. This is likely due to differences in specific implementations of the pit filling or flow direction algorithms. If a higher flow accumulation threshold was used for GRASS, the spatial extents of the features would be reduced, likely improving the match

to the first D8 group. LandSerf is a terrain analysis package with specialized semantics, so differences to the other D8 methods are expected. Finally, the Fowler-Little results have characteristics that bridge between, or augment the results of the two main groups. It produces moderately linear features, like the first group, but there are numerous small features so much of the region is filled, like the second group. The numerous mixed features likely arise from the local neighborhood analysis of the Fowler-Little method, which effectively differentiates adjacent cells into different classes. Also, there is no thresholding step, which tends to extract regions of contiguous cells, with high flow accumulation values, as the same feature class.

An issue that impacts ridge extraction arises directly from the pit filling procedures used in the D8 methods. The final results, and the intermediate results (not shown), indicate that pit filling of inverted terrain data is not suitable for identifying the ridge class. Enduring landscape features, which resist erosion, can maintain relatively large hill and peak structures in the landscape. Such endurant peaks and hills are not mirror images of pits and dales, which gravity fills with eroded sediment. When elevation data are inverted to perform stream extraction methods for ridges, this asymmetry affects the results. The endurant peaks become vast pits and any pit filling processes lead to large areas of terrain, inside the rim of the depression around the pit, that are masked from subsequent morphometric analysis. The flow direction leads from the highest flow accumulation entry points and exits the depression at the pour point, instead of following any critical line forms inside the rim. Ultimately, this process of filling vast pits produces poor representations of ridges. Specialized semantics are required to more accurately extract each of the course and ridge classes.

The pit, peak, and saddle features shown in Figure 7 have notably smaller spatial

extents than the course and ridge features. This is reasonable since they are essentially critical points, not meandering lines. Although not the only locations extracted, for both the Min-Max and Fowler-Little methods, there is a clear pattern of peaks along ridges and pits along courses. This pattern indicates the strong association between the respective high (peak and ridge) and low (pit and course) landscape features. Saddles are associated with both sets of extrema. This suggests that saddles supplement the two sets, in the sense of building continuity of the critical lines. For example a ridge line might climb to a peak, then descend to a saddle, and rise up again to a peak, and so on.

The Fowler-Little method extracted a considerably larger number of peak, pit, and saddle features than the Min-Max method. For saddles, this is partially due to the limited number of unique saddle form definitions in the Min-Max method. However, the initial identification of course and ridge candidates of the Fowler-Little method likely leads to an abundance of features found in all three classes. Before following ridges or courses from passes, allowing up to three ridge or course candidates in a four-cell neighborhood is less strict than the Min-Max approach of allowing a single feature in a six-cell neighborhood. Also, the course and ridge passes of the Fowler-Little method were combined into a doubled set of saddles, approximating the number of peaks and pits combined.

Overall, the variety of results is an effect of both feature extraction semantics and scale. Also, a representation of ridges, courses, peaks, pits, or saddles as fields of membership values across scales implies varying degrees of uncertainty of the degree to which the character of any location matches the terrain feature definitions.

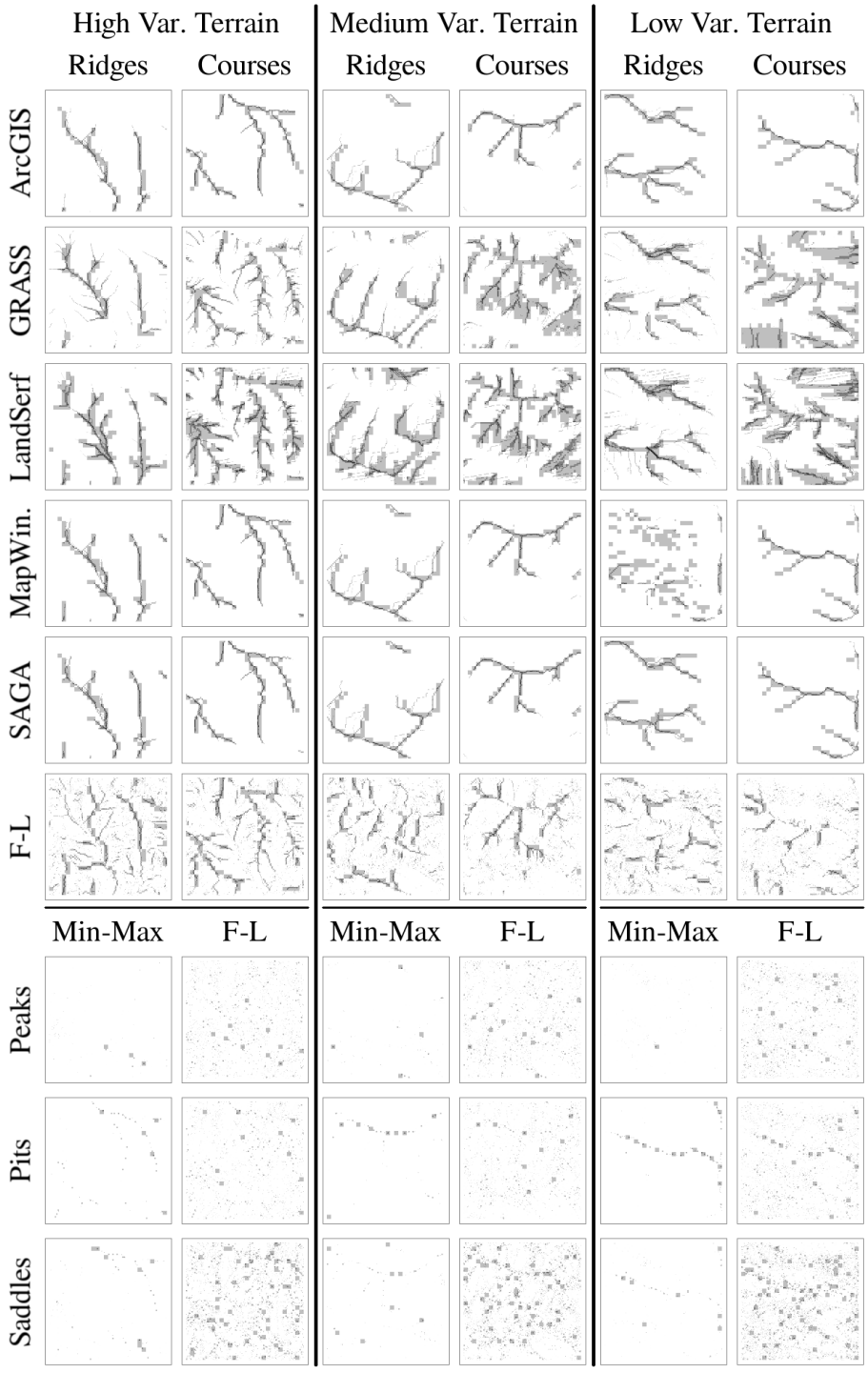


Figure 7: Multi-scale membership value maps for each terrain feature and method. Darker shades represent higher membership values across scales.

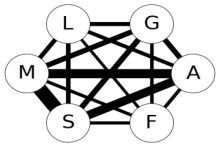
2.3.3 Multi-scale RMSE differences

As a quantitative comparison between the extraction methods, the pairwise RMSE differences between maps of multi-scale membership values were calculated for each feature class in each terrain type. Figure 8 shows three sets of tables, one set for each terrain type. The RMSE values of ridges and courses are also represented with node-and-link graphs. The nodes represent the methods, indicated by the first letter of their name (A:ArcGIS, G:GRASS, L:LandSerf, M:MapWindow, S:SAGA, F:Fowler-Little). Links represent the inverse of the RMSE differences so thicker lines represent more similarity. The significantly thick lines in the graphs represent major similarities, which correspond to the notably small RMSE values in the tables, shown in bold font.

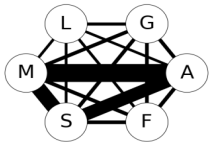
The RMSE node-and-link graphs in Figure 8 present two major patterns. First, for both courses and ridges, the ArcGIS, MapWindow, and SAGA group have considerable similarities, as noted previously with the multi-scale maps. Also as in the maps, MapWindow's unexpected confusion over ridges in low terrain variability is evident by smaller links compared to other links of that pattern. On the other hand, the LandSerf and GRASS group, and the Fowler-Little method each have relatively little similarity with any other method. So, the quantitative RMSE differences do not readily capture the general visual similarities of broad, region-filling features previously noted in the maps. This result suggests that, in addition to a quantitative evaluation, a visual review of mapped features is an important and complementary step to perform. The second major pattern in the Figure 8 node-and-link graphs is that results of the methods in the first group of ArcGIS, SAGA and MapWindow match much better for courses than for ridges, likely due to the pit-filling concerns for ridges already discussed.

High Variability Terrain

Ridges



Courses



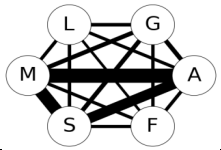
Ridges	A	G	L	M	S	F
A						
G	0.09					
L	0.12	0.11				
M	**0.05**	0.09	0.11			
S	**0.05**	0.09	0.11	**0.04**		
F	0.14	0.15	0.17	0.14	0.14	

Courses	A	G	L	M	S	F
A						
G	0.13					
L	0.18	0.15				
M	**0.02**	0.13	0.18			
S	**0.04**	0.13	0.18	**0.03**		
F	0.13	0.14	0.17	0.13	0.13	

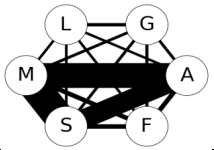
Peaks	Min/Max vs. Fowler Little:	0.06
Pits	"	0.05
Saddles	"	0.11

Medium Variability Terrain

Ridges



Courses



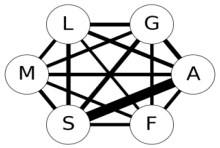
Ridges	A	G	L	M	S	F
A						
G	0.11					
L	0.15	0.13				
M	**0.03**	0.11	0.15			
S	**0.05**	0.11	0.15	**0.04**		
F	0.15	0.16	0.19	0.15	0.15	

Courses	A	G	L	M	S	F
A						
G	0.17					
L	0.19	0.14				
M	**0.02**	0.17	0.19			
S	**0.02**	0.17	0.19	**0.02**		
F	0.11	0.16	0.19	0.11	0.11	

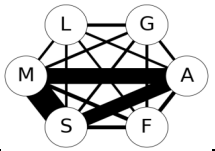
Peaks	Min/Max vs. Fowler Little:	0.06
Pits	"	0.05
Saddles	"	0.11

Low Variability Terrain

Ridges



Courses



Ridges	A	G	L	M	S	F
A						
G	0.11					
L	0.12	0.12				
M	0.13	0.14	0.15			
S	**0.05**	0.11	0.12	0.13		
F	0.15	0.16	0.18	0.15	0.14	

Courses	A	G	L	M	S	F
A						
G	0.17					
L	0.20	0.15				
M	**0.03**	0.17	0.20			
S	**0.03**	0.17	0.20	**0.02**		
F	0.12	0.18	0.21	0.12	0.12	

Peaks	Min/Max vs. Fowler Little:	0.07
Pits	"	0.06
Saddles	"	0.11

Figure 8: Pairwise RMSE differences between maps of multi-scale membership values for the five feature classes in three the terrain types. Values for ridges and courses are also represented as node-and-link graphs. Nodes represent the methods and thicker links represent more similarity.

Degrees of agreement between methods does not imply that any set is more accurate than another. In addition to similar approaches producing near matches, it is possible that slight differences in approaches result in considerable mismatch or, conversely, that vastly different approaches can identify similar feature locations. Both Figure 7 and Figure 8 show that the D8-based feature extraction processes can produce similar results, but are also sensitive to different implementations. Both related and diverse workflows can provide information about the locations of discernible terrain features. However, each definition of features is limited in some way and subject to semantic uncertainty. Comparing several methods both quantitatively and visually can provide objective evidence of slightly different process results and guide the refinement of methods to identify and address nuanced characteristics not previously included in any single algorithm.

2.3.4 Terrain differences: RMSE summaries

To evaluate the multi-scale map differences across the three terrain types, the values in the RMSE difference tables were summarized into two graphs in Figure 9. The top graph consists of two sets of three box and whisker plots each that summarize the RMSE differences; one set for ridges and another for courses. The bottom, middle, and top lines of the boxes represent the first, second, and third quartiles and the whiskers represent one standard deviation from the mean. Overlain on each of the box and whisker plots are small circles representing the source data, the fifteen RMSE values, which correspond to the pairwise differences of the five methods. Trend lines across the three terrain types connect the mean values (with dashes) and median values (with dots). The bottom graph presents trend lines, across the three terrain types, of the RMSE difference patterns between the two extraction methods each for peaks, pits, and saddles.

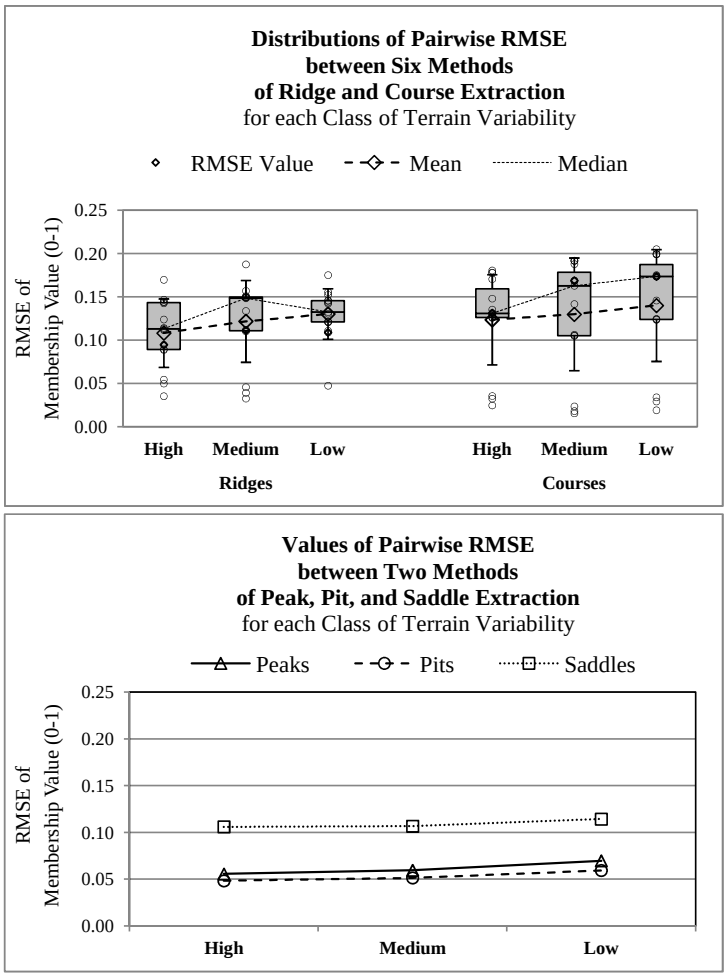


Figure 9: Summary graphs of the pairwise RMSE differences between methods for the five feature classes with respect to the three terrain types. Top: box plots for ridges and courses. Bottom: line graphs for peaks, pits, and saddles.

Some general issues are noteworthy in reviewing the summaries of pairwise RMSE differences. First, lower RMSE values are not necessarily more accurate. The close matches between results of methods can be from high or low quality representations. Second, differences are affected by both the spatial extent and the range of membership values. A large number of mismatched pixels of broad (as opposed to fine) features might have a greater effect upon differences between maps than the intensity of membership values, which range from 0 to 1. Finally, it is helpful to consider the RMSE summaries (Figure 9)

along with the source RMSE tables (Figure 8) and multi-scale membership value maps (Figure 7).

There are similarities between the RMSE summary graphs of ridges and courses. The position of boxes, means, and medians along the y-axis of RMSE values show that the pairs of maps generally differ least in high variability terrain and that the differences increase as the terrain variability decreases. This suggests that there is more classifier confusion with the ambiguity of low variability terrain while feature forms are more discernable and crisper in high variability terrain. Another common pattern, as previously noted with the RMSE tables, is that each of the sets have outliers of noticeably small differences between the pairs of the ArcGIS, MapWindow, and SAGA maps.

The well-matching outliers result in contrasts between the ridges and courses. There are generally wider spreads of differences for the courses compared to the ridges. This is apparently due to opposing groups of methods that either match well or not, as also indicated by the relatively large separation between the high medians compared to means for the courses. Instead of separated groups of extremes, the ridges tend to have more evenly-distributed spreads of RMSE values that converge with lower terrain variability. Also, the presence of different RMSE distribution patterns (of spreads) from the same workflows, only finding ridges with inverted DEMs, supports the suggestion of the assymetry of terrain features.

Specific to ridges, the spread of RMSE values is widest in the high variability terrain. As terrain variability decreases, the spread converges, but to larger differences. Compared to ridges in the high variability terrain, the medium variability terrain ridges of LandSerf and GRASS are broader, leading to more differences. In the low variabil-

ity terrain, ridge differences are greatest, with the additional fragmentation seen in the GRASS, the Fowler-Little, and the (anomalous) results of MapWindow.

For courses, this pattern of increasing differences with lower variability terrain also generally exists, except for the agreement between the ArcGIS, MapWindow, and SAGA pairs. With decreasing terrain variability, the courses of LandSerf and GRASS are broader and Fowler-Little increasingly fragment. Differences between the variety of LandSerf, GRASS, or Fowler-Little results increasingly diverge from the well-matching ArcGIS, MapWindow, and SAGA group.

Considering the peaks, pits, and saddles graphs, there are noticeably larger differences for saddles. This is due to differences in approaches mentioned in the multi-scale results: the Min-Max method including a limited set of saddle definitions, the Min-Max method allowing a single critical point in a six-cell neighborhood, and the Fowler-Little method allowing an initial set of three ridge or course candidates in a four-cell neighborhood. Most importantly, with regard to differences across terrain types, is that the pattern of larger differences for low variability terrain appears again. The trend is slight, but it exists for all three features, as it did for courses and ridges.

2.3.5 Multi-feature membership values

One goal was to represent all of the surface network features on one map. However, a problem arises if multiple feature classes are represented as overlapping at any location. This can occur if the feature types were extracted independently. For example, the SAGA method used separate, independent processes to find each of the features. There is a chance that several feature classes are found at a single location. In contrast,

the Fowler-Little method is a deterministic approach, with only one feature possible at each location. With a multi-scale approach, however, the independent and deterministic methods both generate fuzzy fields of membership values for each feature. With this multi-scale fuzziness, there is potential for each location to exhibit characteristics of every feature.

Figure 10 includes the multi-feature membership value maps showing the 1st- and 2nd-ranked features at each location, and the differences between the two. Ridges, courses, peaks, pits, and saddles are represented with magenta, cyan, red, blue, and green, respectively; differences are in gray. Darker shades represent larger magnitudes, for either membership values or differences. Two example sets and close-up details of the high variability terrain type results of the SAGA and Fowler-Little methods are shown in Figure 11.

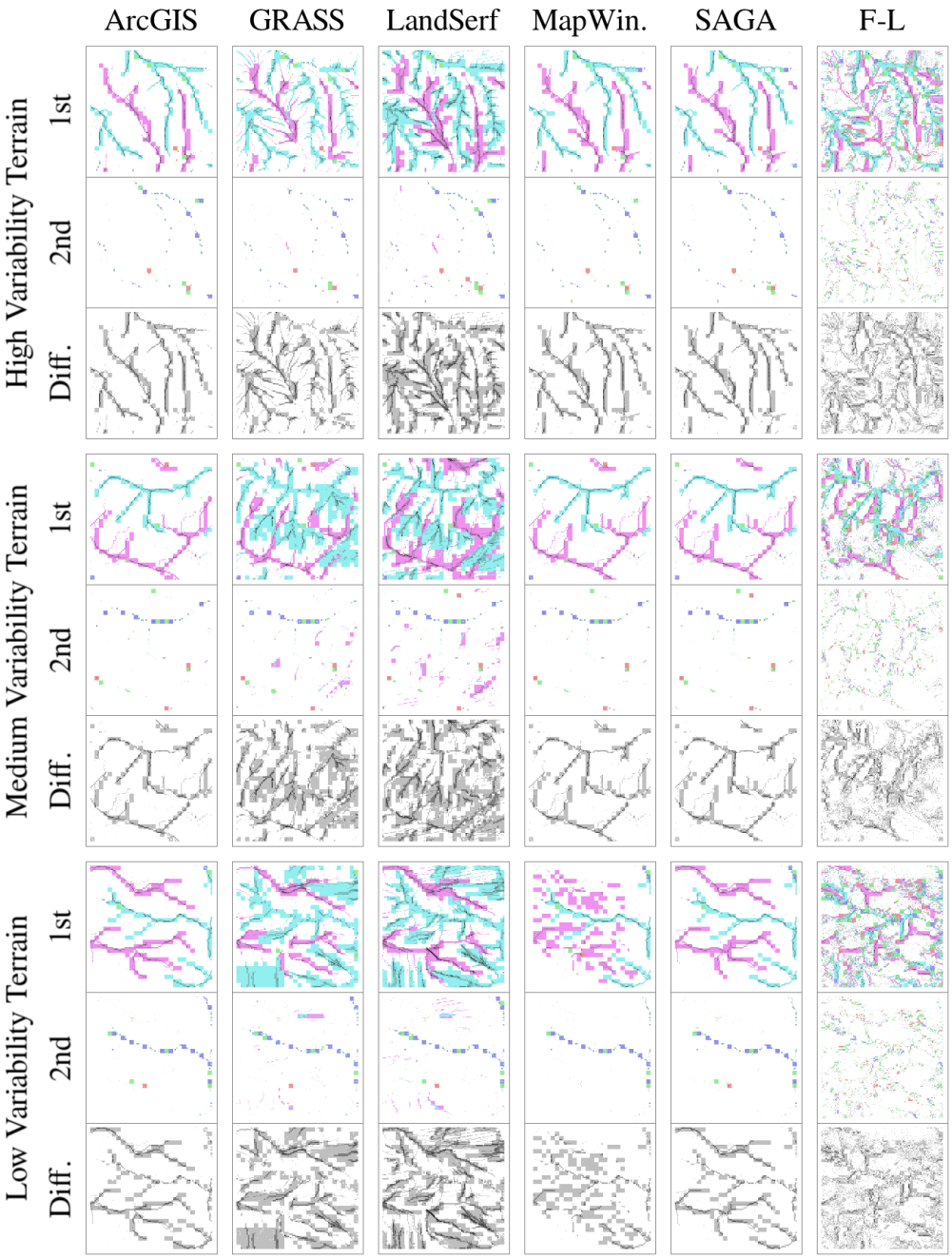


Figure 10: Multi-feature membership value maps showing the 1st- and 2nd-ranked features at each location, and the differences between the two. Ridges, courses, peaks, pits, and saddles are represented with magenta, cyan, red, blue, and green, respectively; differences are in gray. Darker shades represent larger magnitudes, for either membership values or differences.

In reviewing the full set of results shown in Figure 10, the similarities within and differences between results of the two software groups are quickly discernible. Also, the major pattern of course and ridge dominance is inherited from the lineage of flow accumulation and threshold processes previously described.

The dominant features are increasingly crisp as terrain variability increases. In terrain with high variability, surface network features have finer configurations than the broad and spatially extensive features in terrain with low variability, exemplified by GRASS, LandSerf, and Fowler-Little. However, feature crispness also involves the intensity of membership values. The high variability terrain has generally darker shades, representing high membership values, compared to terrain types with less variability. The lack of crispness in low variability terrain may be due to thresholding of courses and ridges as the top 95% flow accumulation. A threshold potentially results in "swamped" representations of broader spatial extents than a geomorphometric approach like that of the Fowler-Little algorithm. Patterns of broad features with low membership values are apparent in the low variability terrain with smoother surfaces and fewer distinctive forms than in the surfaces with high variability.

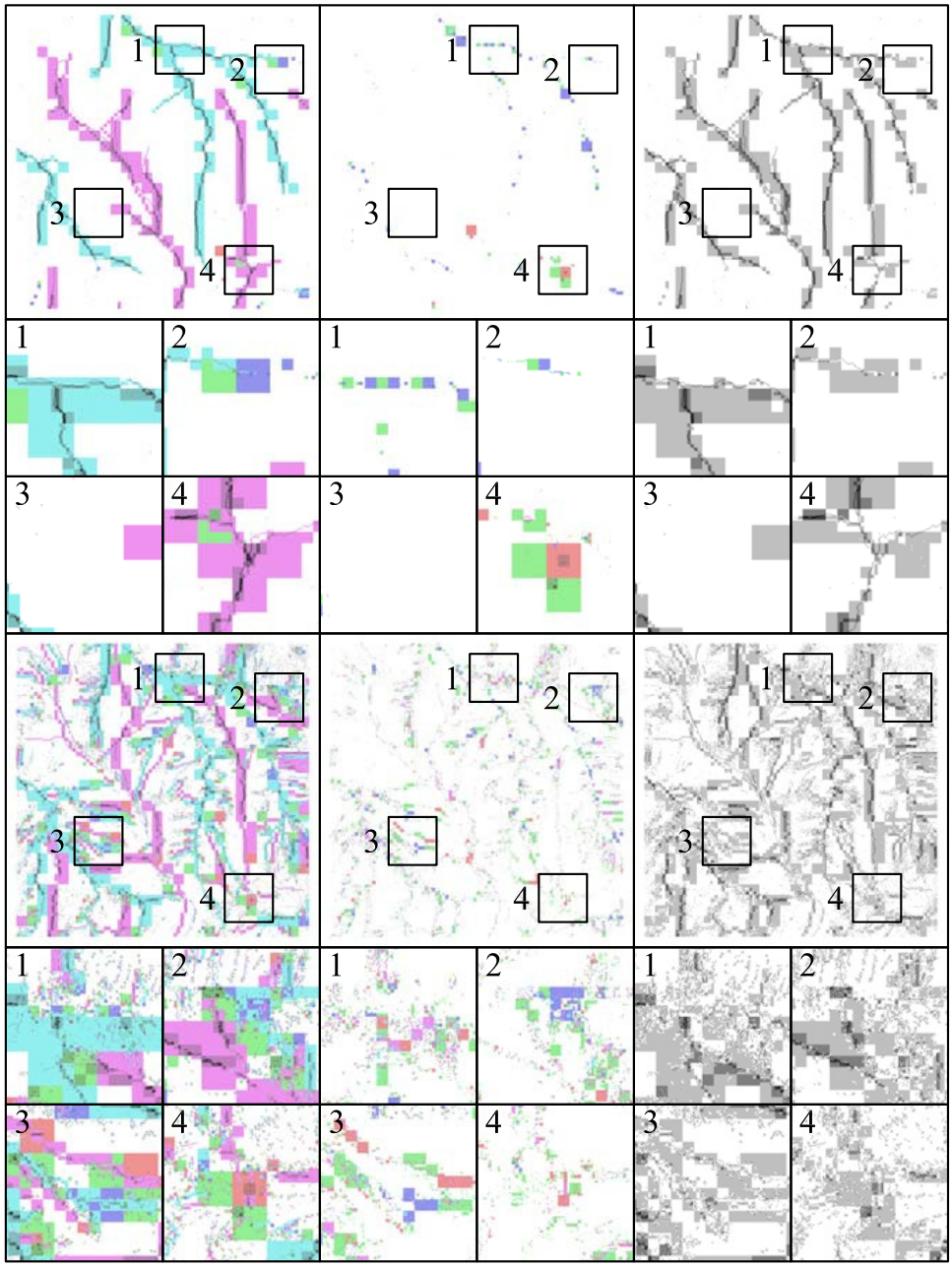


Figure 11: Two example sets and close-up details of the multi-feature membership value maps in the high variability terrain type from SAGA (top two rows) and Fowler-Little (bottom two rows). The first row of each set shows, from left to right, the 1st- and 2nd-ranked membership value features and the differences between the two. The second row of each set shows close-up details. Ridges, courses, peaks, pits, and saddles are represented with magenta, cyan, red, blue, and green, respectively; differences are in gray. Darker shades represent larger magnitudes, for either membership values or differences.

In Figure 11, portions of the close-up details 1 and 4 of SAGA and 2 of Fowler-Little have examples to show that locations with less uncertainty are generally in locations having high membership values for the dominant class coupled with low membership values for the 2nd-ranked feature. Much could be said about the overall differences among all of the multi-feature results, but especially so with regard to the close-up details. The SAGA method, with D8 thresholding and limited Min-Max algorithm, provide much less information about potential terrain surface features. As shown in both the overall maps and the details labeled 2 and 3, SAGA produces relatively discrete and isolated features compared to the region-filling "featureness" from the Fowler-Little algorithm. While there are few examples of correspondence for the 1st highest membership values, perhaps only courses in detail 1, there is better correspondence between the methods for some of the 2nd highest membership value features, like the locations of pits in detail 2 and a peak in detail 4.

Another notable pattern is that similar features, with respect to local elevation, occasionally alternate in adjacent locations and between the 1st- and 2nd-ranked feature maps. In details 1 and 2 of the SAGA 1st- and 2nd-ranked features, the course, saddle, and pit features alternate, as do the ridge, saddle, and peak features in detail 4. This pattern of related features alternating is also apparent in the Fowler-Little results, but the picture is much more complicated. Even when the overall maps of both methods appear to represent similar dominant features, there are considerable differences in the details of the Fowler-Little's non-thresholded approach.

2.4 Conclusion

An analysis of the uncertainty of terrain feature extraction was performed. The particular feature classes, related to surface network theory, included peaks, pits, saddles, ridges, and courses. Various extraction methods were employed to produce multi-scale and multi-feature maps. Raster DEMs in four spatial resolutions were used along with six workflows, five involving conventional extraction methods available in free, open-source, and commercial software, and an implementation of a well-known surface network algorithm. The effects of terrain variability were also considered across three levels.

Whether various methods aim to extract similar features, based upon different properties and relationships, they often produce conflicting results. The results of conventional extraction workflows showed the largest differences at important locations, nearby terrain features of interest. Whether extracting each individual feature independently (i.e. separately, with potentially overlapping classes) or all features deterministically (i.e. with unique classes at each location), the multi-scale map of each feature exhibits fuzziness that has the potential to overlap with that of other features. By ranking and extracting features based upon multi-scale membership values at each location, two sets of multi-feature maps were produced. One each of the dominant and the second-ranked features. A third map represented the difference between the top two membership values, which provided a measure of uncertainty similar to an established metric. These three maps show that most of the regions exhibit some degree of "featureness." Some locations exhibit a strong signal of a single, specific terrain feature class. Other areas contain information about two or more feature classes. The variation and uncertainty were partially related to the semantics of competing methods and to the multi-scale fuzziness.

Terrain feature uncertainty was also related to the terrain variability. The comparison showed smaller differences with higher variability terrain. The ambiguity of low variability terrain results in spatially broader and diminished signals of features. With high variability terrain, the various workflows clearly become more congruent with crisper features with respect to both spatial extent and feature class membership values.

The core point in considering multi-feature membership values is that each location exhibits a quantifiable degree of being any of the surface network features. Extraction algorithms are always designed to find a limited set of feature characteristics. Therefore, semantic uncertainty often exists between the concepts of complex, related features and the design of a classifier. Also, even if features are found across several scales of analysis, with high multi-scale membership values, it is also important to recognize that other feature characteristics can also be present. There can be entirely different surface networks at each scale, which increases the complexity of forms and surface-dependent processes. One way to narrow a search for locations with complex configurations and interactions is to find locations with a high degree of uncertainty.

This research supports previous work that suggests surface features associated with natural terrain can only be described in terms of uncertainty (Fisher et al., 2004). However, an approach that combines results for multi-scale and multi-feature analyses provides the means to better understand the terrain. Regardless of which extraction methods are used, the conclusion about terrain feature uncertainty still applies. Although terrain feature extraction incurs quantifiable uncertainty, the overall conclusion is that employing various extraction methods and progressing through multi-scale and multi-feature analyses provides insights into terrain feature uncertainty by increasing information and confidence about which terrain features exist at any given location.

Acknowledgments

This research was supported under a grant from the USGS Center of Excellence for Geospatial Information Science, through the Californian Cooperative Ecosystem Studies Unit as part of work supporting The National Map and the National Spatial Data Infrastructure. We are grateful to Dr. Lynn Usery and the USGS Center of Excellence for Geospatial Information Science for this support. A portion of this research is based on data and processing services provided by the OpenTopography Facility with support from the NSF (Award Numbers 1226353 & 1225810).

3 Spatial outlier detection

3.1 Introduction

In the previous section, the performance of geographical feature extraction was impacted by surface variability, scale, and extraction method. Due to those three factors, the detected location of a feature was subject to some degree of uncertainty. The previous tests on real terrain involved representations, incomplete estimates of the true characteristics of the surface.

Feature extraction results are typically compared to ground-truth data to determine the accuracy of the process. However, the high resolution lidar data sets included as part of the analysis were arguably the best quality of data available, especially with regard to capturing both a fine level of detail and a comprehensive coverage of a study region. Ultimately, this led to a high degree of complexity with regard to the configuration of features. Even with the fine resolution, samples of the true surface remain implicitly sparse and subject to measurement error. The representation of sampled real terrain contributed to diverse results, but so did surface variability, scale of analysis, and extraction methods.

In order to better isolate the effects of variability, scale, and extraction method, a controlled experiment was conducted. Of the set of surface network feature types, one was selected as the subject of focus, peaks. The point with the highest value compared to the local surroundings is presumably a straightforward feature to find. Results would likely also inform analyses of the inverse, or opposite, class of pits. Compared to points of mixed extrema (i.e. saddles) or the meandering and branching linear properties of ridges and courses, the class of peaks was considered the simplest feature to find.

As local extrema, peaks have the characteristics of spatial outliers. In general terms, outliers are rare observations that have significantly different values compared to the remainder of a data set (Barnett & Lewis, 1994). However, if data are spatially-referenced, then spatial outliers can exist. (From here forward, the term outliers refers to spatial outliers unless otherwise stated.) Values of spatial outliers are not necessarily rare. Rather, they have significantly different attribute values compared to their local spatial neighborhood (Shekhar, Lu, & Zhang, 2003).

Spatial outliers are important for understanding nature because they might represent rare places of interest, such as sites with characteristics in need of preservation or with hazards. Although such anomalies are often removed for the process of modeling general trends, the identification of spatial outliers, and why they occur, remain current research goals (Cressie & Wikle, 2011).

The concept of spatial dependence (Tobler, 1970), implies that the influence of phenomena can be observed to extend across space. The similarity of a characteristic in a spatial neighborhood eventually transitions to more variability at a distance. The change in the structure of variance across space can perhaps follow the form of a spatial decay function, as illustrated in Figure 12. It is notable that the similarity in character, or value, of nearby locations presents challenges for the use of general statistics with spatial data because the assumption of the independence of data is violated. However, statistics exist for spatial association, to describe both the general character of spatial phenomena (Moran, 1950; Getis & Ord, 1992) as well as local instabilities (Anselin, 1995). The characterisitcs of spatial stationarity, particularly with regard to local anomalies, is considered by this experiment.

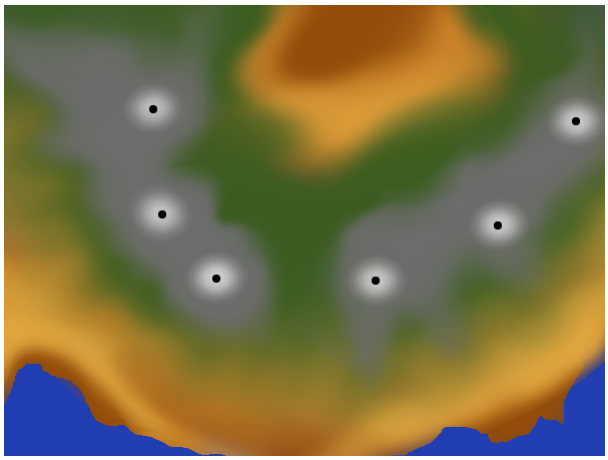


Figure 12: Visualization of spatial dependence around points. Colors represent different attribute values that transition from high spatial autocorrelation near each black point to less similarity (i.e. more variation) farther away.

The Gaussian function was selected as the model of a simulated spatial outlier. The function has a point of largest magnitude. In other words, the form has a maximum, or a peak. It also has continuous, curved transitions toward its base. Spatial patterns may exhibit such variation either directly, such as the terrain in Figure 13, or through transformation. The Gaussian function is a common form for modeling spatial dependence, such as with a variogram. Measurement errors and the Central Limit Theorem are also related to the Gaussian form. Consider the process of finding the position of a static object with a series of measurements assumed to be independent and affected only by random error. An estimate of the location from the many observations might exhibit the pattern of a probability density function similar to a Gaussian distribution: clustered near the true location with the highest density, which smoothly decreases with larger distances, as illustrated in Figure 14.

The research question of this experiment is: how do existing spatial outlier detection

Figure 13: Gaussian shaped hill in the natural landscape.

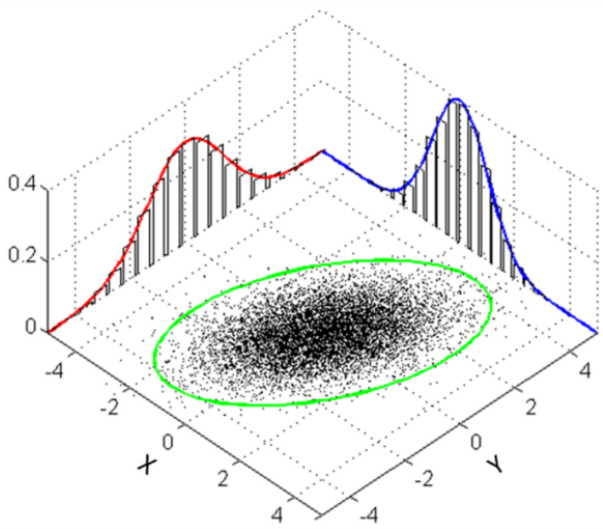


Figure 14: A Gaussian probability density function approximates the degree of clustering of random points (source: Wikimedia Commons, CC0 1.0 Universal Public Domain Dedication, by user Bscan).

algorithms perform with respect to outlier shape and the scale of analysis? The tests involve a known outlier shape, Gaussian, with varying height and width, represented by raster data sets of varying resolution, produced with various assignment operators.

Even though spatial outliers are commonly considered discontinuities compared to their spatial neighborhood, they may still exhibit smoothness at finer scales. For example, because of the local smoothness of the Gaussian function, the spatial outlier may or may not be detected. The result might depend upon the scale of analysis and the size of

the anomaly. The shape of the simulated outlier was controlled by varying parameters of the Gaussian form. Along with the potential for an added constant value, the function has three parameters: the center, spread, and amplitude. For this experiment, the center was held constant. The only Gaussian parameters varied for the tests were the amplitude and spread (termed height and width from here forward).

For this controlled study across various scales of analysis, two stages of data generation were necessary: the generation of a "source" data set with (relatively) "super-fine" resolution, and the generation of data sets to be tested. For the source data, a computational simulation of an outlier with a Gaussian form requires that values be computed at discrete locations based upon the mathematical function. A Gaussian function estimation, as a super-fine resolution raster grid, served the purpose as a source data set, from which the coarser resolution raster grid "test" data sets were computed.

One consideration in generating test raster grids was the assignment operator employed. The operator assigns a singular value to each test raster cell based upon a set of many spatially coincident data values of the super-fine resolution source raster. The reason various operators were evaluated was to better understand the effects of data transformation on the accuracy of representing a peak. The smoothness of super-fine resolution data is reduced by the transformation process of aggregation into coarser resolution raster cells. As illustrated in Figure 15, various operators will each have different degrees of impact on the tested outlier's smoothness and the subsequent spatial outlier detection processes. An accuracy assessment can quantify that impact.

If an outlier's spatial structure, its size, matches the scale of analysis, its value is most clearly different than neighboring values. However, if there is a mismatch between the

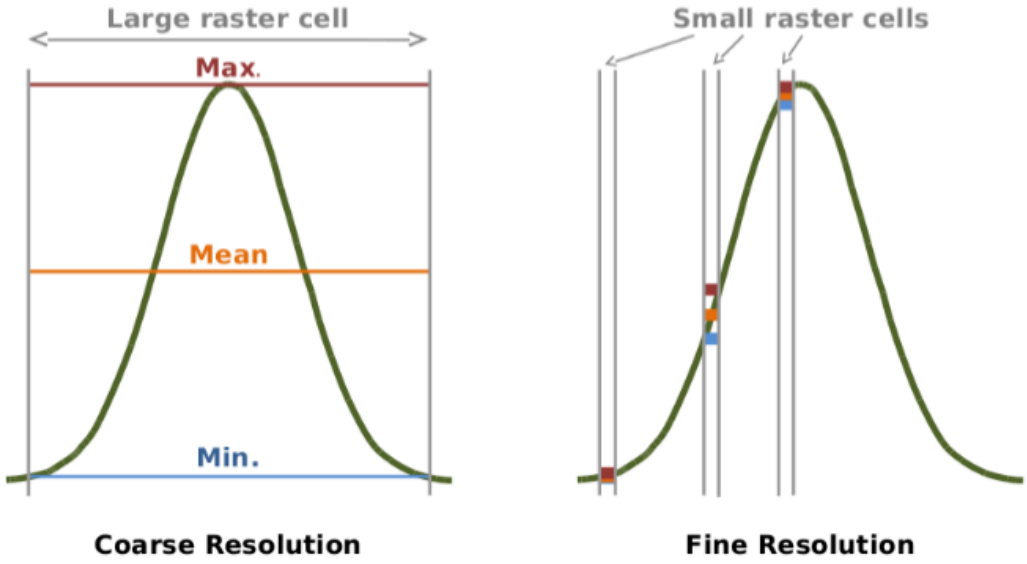


Figure 15: Illustration of the impact of the assignment operator and the grid resolution. Due to assignment, in large grid cells, the signal of a spatial outlier might be missed. Due to resolution and assignment in small cells, outliers can be missed by comparison to neighbors where there is a major transition (i.e. along the side) or where there are similar values (e.g. near the top or base).

size of the outlier and the scale of analysis, the signal of the outlier can be missed due to masking or swamping. Masking occurs if the outlier is smaller than the scale of analysis. For example, a unit of analysis, such as a raster cell, might be assigned a value using an aggregation function applied to many observations located within the cell's spatial extent. An average of those values can "mask" the presence of a rare extreme value. In other words, the minority contribution of the rare value will not be well-represented by the average. The other scale mismatch, when the unit of analysis is much finer than the size of the outlier, can lead to swamping. If the spatial extent of the anomalous phenomena exceeds the resolution of the data, the region near the top of the outlier might be representated by many similar values that "swamp" across numerous contiguous cells. A spatial outlier detection process might miss the broad top region of similar values. With either case of a scale mismatch, an outlier detection method that tests neighboring values

for high contrast can fail.

Numerous computational detection methods have been developed to identify the location of spatial outliers (Aggarwal, 2013; Chandola, Banerjee, & Kumar, 2009). Spatial outlier detection processes commonly require the definition of a neighborhood around any point of analysis and a comparison function between attribute values of that point and neighboring values, as illustrated in Figure 16.

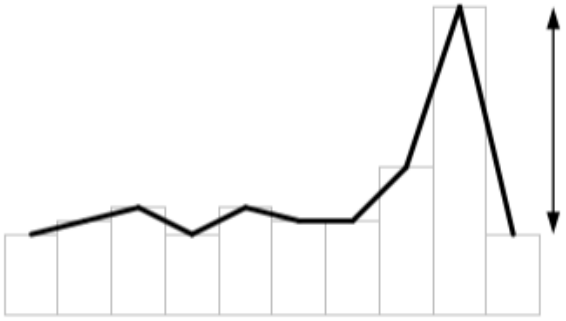


Figure 16: Illustration of the general spatial outlier model of common detection algorithms, to find a relatively extreme attribute value compared to those in a given neighborhood.

When introduced, the algorithms are often "field-tested" with complex real data. However, similar to the analysis in the previous section on surface network features, there are typically no ground-truth data for performing an accuracy assessment of the detected outliers. In contrast, this study evaluated three established techniques against controlled reference data of simulated outliers to discern breakdown characteristics of the three selected detection methods.

The first detection method was the non-iterative *Median* algorithm of Lu, Chen, and Kou (2004). Developed as a multiple attribute spatial outlier detection method, it

involves finding the median value of each data point's respective neighbors. It includes an assumption that the distribution of the differences from respective neighborhoods is Normal. Probabilities of each point are calculated using its Mahalanobis distance in each variable. A comparison involves an arbitrary threshold to label extreme values as outliers.

The second method was the $Z-test$ algorithm of Shekhar et al. (2003). This method involves representing the data as a Moran scatterplot and calculating a regression line from the points. The residual of each point is determined and the distribution of all residuals is assumed Normal. The z-score probabilities of residuals are used for comparison to an arbitrary threshold to label extreme values as outliers.

The third method was the spatial local outlier method ($SLOM$) algorithm of Chawla and Sun (2005). A goal of this algorithm is to suppress the labeling of outliers in highly unstable areas, where the notion of outlier is not meaningful. First, the maximum value of distances between the point of analysis and its neighbors is removed from the computation, similar to the computation of a trimmed mean. Then the mean of distances between the point of analysis and each remaining neighbor is calculated. The mean is used in a count of the number of times it is crossed by the ordered sequence of differences between each neighbor and the point of analysis. The neighborhood variability is used as a factor to inversely adjust the final $SLOM$ value.

As is typical of many spatial outlier detection methods, the algorithms produce a value, or outlier score, for each datum that represents the magnitude of the result of the comparison function. That value essentially represents the "outlierness" of the datum. Beyond the application of the comparison function, another factor in discriminating between "normal" regions and spatial outliers involves an arbitrary threshold, selected as

appropriate for the topic of study. In this controlled experiment a range of thresholds were evaluated to collect a comprehensive indication of the detection performance without inducing bias due to any specific threshold per method. From each spatial outlier detection process on each data set, numerous sets of labeled outliers were created from the outlier scores using the range of thresholds at each percentile of 0 through 100.

As an accuracy assessment, each set of results was compared to a reference outlier as represented by the majority of the area under the Gaussian function. The true positive rate (TPR), or hit rate, and false positive rate (FPR), known as the Type I error, were both calculated as accuracy metrics. With these two metrics, two other metrics can be calculated if necessary, the true negative rate and the FPR, known as the Type II error. In order to compare methods with respect to the height and width Gaussian parameters, the grid resolution, and the assignment operator, graphs of each were produced with both the mean and median TPR and FPR of all the other variables combined.

3.2 Methods

3.2.1 Software

Two free and open-source software were employed in the experiment. The generation of the outlier simulations and the analyses of the results were performed with the statistical software R (R Development Core Team, 2013). The ELKI software provided the algorithms and operations for the spatial outlier detection (Achtert, Kriegel, Schubert, & Zimek, 2013).

3.2.2 Gaussian shape

Attribute values of the raster grids represented a two-dimensional (2D) Gaussian shape. The Gaussian function has the following parameters: a is the height (i.e. amplitude), b is the center, c is the width (i.e. standard deviation, or spread), and d is a height constant. The independent variables x and y are the coordinate locations across the two respective spatial dimensions in which the Gaussian function value was computed. Equation 2 is the general form of the Gaussian shape and Equation 3 is the function employed to compute the 2D Gaussian value at each source grid coordinate.

$$f(x) = ae^{-\frac{(x-b)^2}{2c^2}} + d \tag{2}$$

$$f(x,y) = \sqrt{\left(ae^{-\frac{(x-b)^2}{2c^2}} + d\right)\left(ae^{-\frac{(y-b)^2}{2c^2}} + d\right)} \tag{3}$$

Permutations of the Gaussian parameters specified 45 unique shapes. Parameters for both of the x and y dimensions were varied equally to create spatially symmetric shapes. The height parameter a was incremented from one though nine, in steps of two (five values). The width parameter b was incremented from one though nine, in steps of one (nine values). The center c and the height constant d were both held at zero for all permutations.

The one-dimensional representation in Figure 17 illustrates boundary cases of the permutations if either the Gaussian height or the Gaussian width was held at the boundary value (i.e. minimum or maximum), while the other varied. The first frame shows the thinnest outlier shape, varying across the full range of heights. The second frame shows the tallest outlier shape, varying across the full range of widths. The third frame shows

the widest outlier shape, varying across the full range of heights, and the fourth frame shows the shortest outlier shape, varying across the full range of widths.

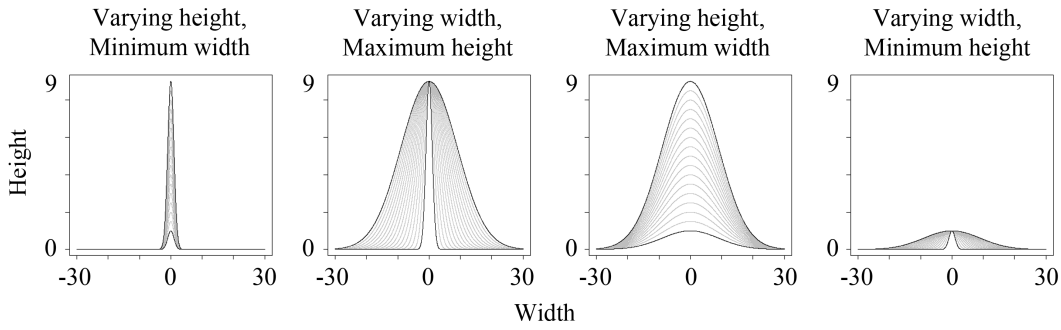


Figure 17: One-dimensional representation of varying the Gaussian shape parameters.

3.2.3 Source grids

All of the raster data sets were square grids that measured 162 units (not cells) on each side with a coordinate domain in each dimension of -81 through 81 units. The 45 super-fine resolution source grids, each containing values calculated from a singular Gaussian function permutation, were 810 x 810 cells each.

3.2.4 Test grids

Coarser resolution test grids contained spatially generalized representations of the source data, as any sampled values might be observed for real geographical variables. The test grids had resolutions with odd numbers of cells along each axis, which ranged from three cells through 81 cells, in steps of six (14 resolutions). Corresponding to each test grid, related grids had a matching resolution (e.g. for grids containing reference labels, outlier scores, outlier labels, and accuracy assessment labels, each described below).

Considering only the 45 unique Gaussian shapes represented at each of the 14 test grid resolutions, there would have been 630 test grids. However for each of these permutations, three cell value assignment operators were separately employed to create three test grids, one grid for the *minimum*, one for the *maximum*, and one for the *mean* operator. Cells of each of the three test grids represented the result of applying one of the three assignment operators to the sets of intersecting source grid cells. This process created a total of 1,890 test grids (45 Gaussian shapes $\times$ 14 resolutions $\times$ 3 assignment operators).

3.2.5 Neighborhoods

All spatial outlier detection processes require a defined neighborhood, to which the location of analysis is compared. Neighbors of each test grid cell were identified as any adjacent cells in either the vertical, horizontal, or diagonal directions. This was generally the queen's case of eight adjacent neighbors, except for edge cells, which have five neighbors, and corner cells, which have three neighbors.

3.2.6 Outlier detection algorithms

To each of the 1,890 test grids, three spatial outlier detection algorithms were applied separately. The algorithms, as implemented in the ELKI software, belong to the *outlier.spatial* collection of methods and are listed below:

$CTLuMedianAlgorithm$: the non-iterative $Median$ algorithm of Lu et al. (2004),
$CTLuZTestOutlier$: the $Z-test$ algorithm of Shekhar et al. (2003), and
$SLOM$: the $SLOM$ algorithm of Chawla and Sun (2005).

3.2.7 Outlier score grids

Applying the three algorithms to each of the 1,890 test grids produced 5,670 outlier score grids. Cells of these grids each contained an outlier score, which is a numerical value computed by the comparison function of each detection algorithm. It is a method-specific measure of outlierness at any cell location. For the three methods evaluated, a larger magnitude of outlier score indicates a greater likelihood that the location has spatial outlier characteristics.

3.2.8 Outlier label grids

In practice, and as appropriate for any particular domain of study, a threshold value is arbitrarily specified for application to the outlier scores. Any cell with a score that matches or exceeds that value is labeled as a "detected" outlier. However, because different algorithms produce different ranges of scores, a comparison of various algorithm results with identical thresholds is not suitable.

For this analysis, the three detection methods were quantitatively compared based upon a range of threshold values. Each threshold value was applied to each outlier score grid to create a grid of outlier labels. Ranging from the minimum to the maximum of each outlier score grid, a sequence of 101 equally-spaced threshold values were computed to represent percentiles from zero to 100. This comprehensive approach includes thresholds that potentially label outlier cells in any number from zero cells to all cells in the grid. Aggregating results of outliers labeled with all the thresholds, later, provides a general response metric of the outlier detection performance, or how frequently it labels outlier cells.

3.2.9 Reference label grids

For the accuracy assessment, reference grids were created to compare with the detection results of labeled outliers. The extent of a reference outlier was defined as three standard deviations from the center of the Gaussian shape. Therefore, cells of the reference grids were labeled as reference outlier cells if they intersected a circle with a radius three times the Gaussian width parameter (i.e. three standard deviations) from the center.

3.2.10 Accuracy assessment grids

This controlled study enables a comparison of a grid of outlier labels (o) to be compared to its related grid of reference labels (r). The comparison produces a grid of accuracy assessment classes (a). The four possible classes, true positive (tp), false positive (fp), true negative (tn), and false negative (fn), are determined for each corresponding cell (i) by the conditions in Equation 4:

$$a_{\mathrm{i}} = \begin{cases} tp \text{ if } o_{\mathrm{i}} = TRUE \text{ and } r_{\mathrm{i}} = TRUE \\ fp \text{ if } o_{\mathrm{i}} = TRUE \text{ and } r_{\mathrm{i}} = FALSE \\ tn \text{ if } o_{\mathrm{i}} = FALSE \text{ and } r_{\mathrm{i}} = FALSE \\ fn \text{ if } o_{\mathrm{i}} = FALSE \text{ and } r_{\mathrm{i}} = TRUE \end{cases} \tag{4}$$

3.2.11 Accuracy assessment metrics

The two accuracy assessment metrics calculated were the true positive rate (TPR) and the false positive rate (FPR). The metrics are determined from the total counts of the true positives (TP) and the false positives (FP), from the accuracy assessment grids, and the total counts of the reference positives (P) and the reference negatives (N), from the reference grids. The TPR was calculated with Equation 5 and the FPR was calculated with Equation 6.

$$TPR = TP/P \tag{5}$$

$$FPR = FP/N \tag{6}$$

3.2.12 Accuracy assessment graphs

To compare the effects of each algorithm with respect to each variable tested, the TPR and FPR of all other variables were aggregated. Both the mean and median TPR and FPR values were computed for each variable tested: the grid resolution, the Gaussian height, Gaussian width, and the assignment operator. As mentioned previously, all of the outlier label thresholds were included in the process of computing the accuracy assessment grids, the TPR, the FPR, and the aggregated TPR and FPR mean and median values. Including both the mean and median values provides information about the skewness of the distribution of each set of results, based upon the property of the mean being skewed further toward extreme values than the median.

3.3 Results and Discussion

Results of the experiments enabled both quantitative and qualitative analyses. Graphs of the TPR and the FPR represent the quantitative accuracy assessment metrics. Additionally, numerous sets of raster grids were produced representing sources, reference labels, outlier scores, and thresholded outlier labels. A set of examples of such raster grids were selected for a visual qualitative analysis.

Figure 18 presents the mean and median TPR and FPR values for each of the three algorithms. In each of the graphs, accuracy assessment values derived from all outlier label thresholds are aggregated across all variables, with respect to each singular variable tested. Therefore, each graph illustrates a highly generalized response observed for each variable.

3.3.1 False positive rates

Nearly all of the FPR results are very low. This is expected with a controlled study of a grid values representing only a single Gaussian outlier structure embedded in an otherwise constant, zero-valued background. With only the presence of a single isolated outlier embedded in the grids and no other variation, there was low potential for confusion between the outlier and the background. The only notable pattern is a very high mean FPR value, indicating poor performance, for the $Z-test$ algorithm in grids with very few cells. A possible reason for this is that this algorithm labels outliers based upon z-values of residuals compared to a regression line, which was likely underspecified with just a few cell values, an unlikely situation in practice.

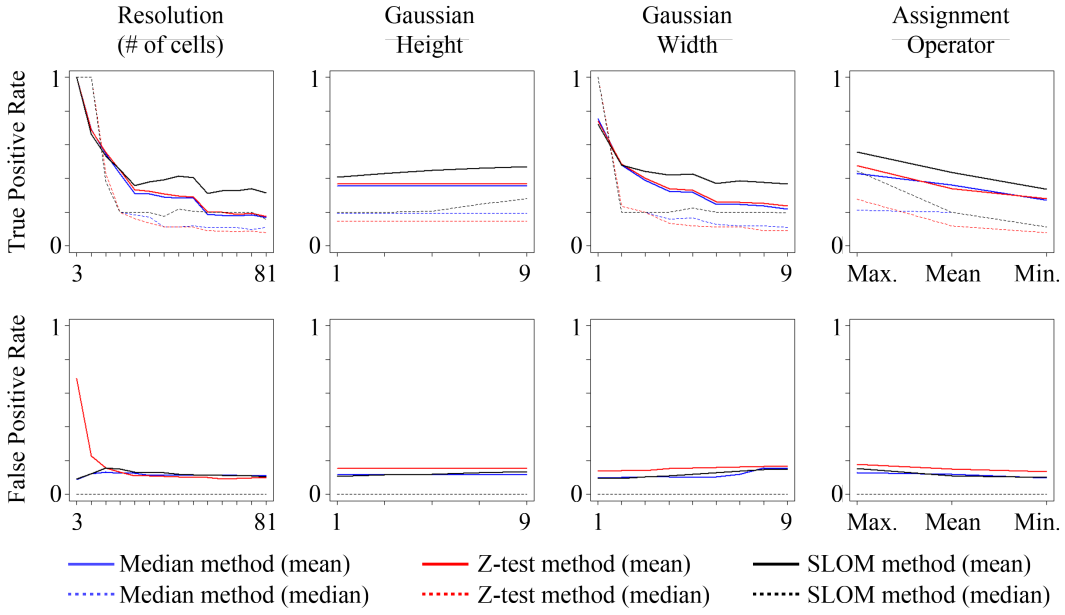


Figure 18: The mean and median of the true positive rate (TPR) and the false positive rate (FPR), with respect to each tested variable: grid resolution, Gaussian height, Gaussian width, and the assignment operators *maximum*, *mean*, and *minimum*.

3.3.2 True positive rates: assignment operator

Compared to the FPR results, the TPR graphs contain more information about the detection of the simulated outliers. The Gaussian shape has considerably larger values near its core than the farther values that asymptote toward zero. Therefore, each of the assignment operators sufficiently represented the outlier for detection. Spatial outliers are best represented by assignment operators that match the sign of the outlier, the direction of the anomalous values on the attribute scale relative to its neighborhood. For example, extreme values of positive-valued outliers are best represented by the *maximum* assignment operator. Although the *minimum* assignment operator resulted in a detectable representation its performance was the lowest of the three operators. Representing an anomaly's positive attribute values by an application of the *minimum* operator results in cells that have minimized or "dampened" values. Aggregating multiple values with

the *minimum* operator results in a representation that misses the peak attribute value of a positive-valued anomaly, even if it was sampled. The approach of matching the assignment operator to the sign of the outlier values is beneficial for representing anomalies for obvious reasons. However, a more common practice is to use the *mean* operator, which would mask the effect of locally extreme values having either positive or negative signs. If it is critical to locate outliers, two separate analyses with the *maximum* and the *minimum* will likely provide better results than a single analysis involving the *mean* operator.

3.3.3 True positive rates: Gaussian height

Also in regard to finding extreme values, a surprising result was that the Gaussian height had little effect. Evidently, regardless of height, the algorithms detected any outliers at nearly the same level of performance across all of the other variables tested. Only the $SLOM$ algorithm had increasing TPR values with increasing heights, but the change was only slight. As only a single outlier was represented, detection was influenced more by presence than by the degree of attribute value extremeness. If the neighborhood contains additional variation, then the height of the outlier is expected to have a greater affect with regard to differentiating the structure of the outlier from any surrounding variability.

3.3.4 True positive rates: Gaussian width

It was the width variable that strongly affected the detection of the Gaussian shape. The smallest width of one unit resulted in a much higher TPR. At that value the reference outlier has a radius of three units and a diameter of six. Doubling that size the mean TPR decreased from about 0.75 to about 0.50, the largest drop. There is a continual

decrease in mean TPR, with a leveling of performance at about a Gaussian width parameter value of 6 units, corresponding to a diameter of 36 units. Beyond that, the $SLOM$ algorithm maintains a mean TPR of about 0.20 and the other two methods have lower performance of about 0.13, decreasing slightly. The median TPR has a similar, but exaggerated, pattern that indicates near peak performance with the smallest Gaussian width of one and a lower performance of about 0.20 at a width of two. With larger widths, the median TPR stays about level for $SLOM$ and decreases slightly for the other two algorithms. In general, broad outliers are problematic for the spatial outlier detection methods, with high rates of performance across all other variables only attained with the smallest of the outlier widths.

3.3.5 True positive rates: grid resolution

The grid resolution is also related to the effects of outlier scale upon the detection performance. Put simply, coarser resolution performed best. Both the mean and median TPR have perfect performance of one with only three cells in both the x and y dimensions. Regardless of the other variables, any outlier present is detected if the value is assigned to just the center of three cells. For the finer resolution of 9 cells, the median TPR remains at one, but some of the aggregated variable results reduce the mean TPR to about 0.70 for all algorithms. The major decrease in mean TPR continues to about 0.18 at a resolution of 27 cells. A divergence then occurs between the $SLOM$ method and the other two until a resolution of 51 cells. The $SLOM$ increases slightly to about 0.20, while the other two remain about constant. There is a large drop for all mean TPR values at the next resolution of 57 cells. $SLOM$ decreases to a mean TPR value of about 0.18 and the other two methods decrease to around 0.15, at which point all of the methods stabilize. Similar to the TPR pattern of increasing the Gaussian width parameter, when

aggregated across all variables, the increase in grid resolution results in a clear decrease in spatial outlier detection performance.

3.3.6 Qualitative results

As illustrated in Figure 19, an interesting outcome of the spatial outlier detection methods was the variety of spatial patterns of labeled outlier cells. To visualize an example of the initial data, included in the upper-left corner of the figure is a test grid with Gaussian values along with its corresponding grid of reference labels to the immediate right. Again to the right, an outlier scores grid provides an example of an intermediate output from a spatial outlier detection algorithm. After the application of thresholds to the outlier scores, grids containing labelled outlier cells were produced. Completing the top row of the figure are examples of the *Median* algorithm outlier label grids. The second and third rows contain examples of the $Z - test$ and $SLOM$ outlier label grids, respectively.

3.3.7 Visually complex results, difficult to retrieve original feature

Figure 19 illustrates how a simple, smooth Gaussian function can result in complex patterns of labeled outliers. If one were to "detect" such a complex pattern of cells as outliers, it would be quite a reach to intuitively or computationally derive the original distribution. The diverse geometrical patterns are partially due to the effects of the data structure of sampled values aggregated at discrete locations of a regular grid. In addition, the patterns were also due to a mismatch between the semantics of the algorithms and the problem of interest. The algorithms are designed to find individual cells that are relatively extreme compared to a given neighborhood. The problem of interest involves varying the sizes of outliers and the scales of analysis, to address the scenario of

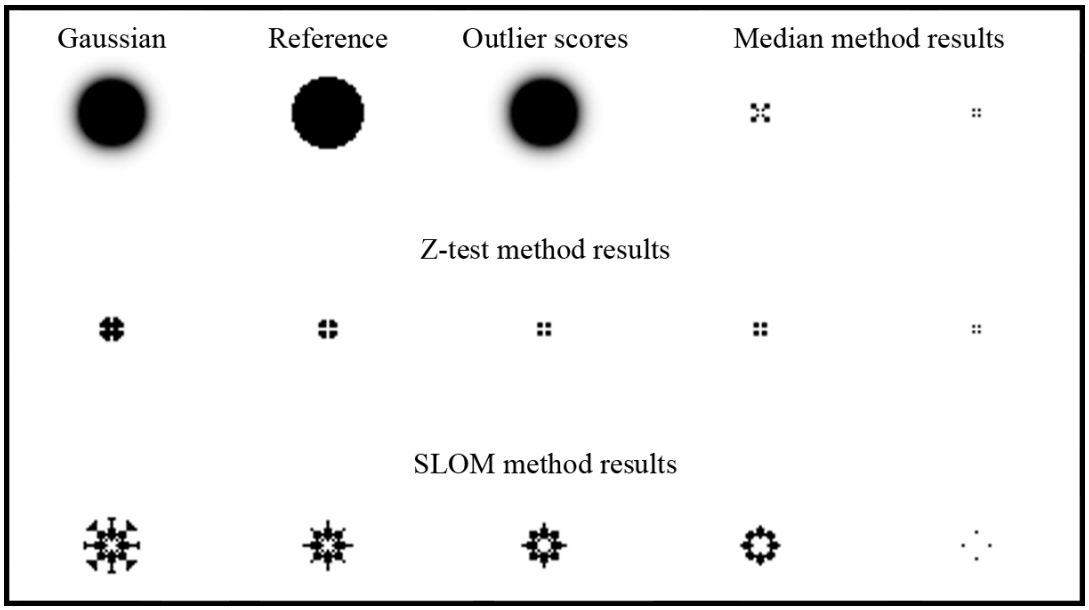


Figure 19: Examples of grids of Gaussian values, reference labels, outlier scores, and selected outlier labels. Darker tones indicate higher Gaussian or outlier score values, or the presence of either the reference or the outlier class.

an outlier "swamping" or spanning across numerous sampled locations. A key pattern illustrated by the outlier label grids is that the center of the initial outlier, having the largest magnitude of the attribute values, rarely resulted in a detected outlier. While the sloping sides of outliers were often represented, in curious patterns, detecting the outlier tops is a clear problem.

3.4 Conclusion

This controlled experiment was conducted in order to better isolate the effects of variability, scale, and extraction method on the process of identifying spatial outliers. The subject of focus, spatial outliers, are similar to the peak class of surface network features as they are positive-valued local extrema. The analysis might also inform the inverse case of negative-valued pit-like outliers. In considering the uncertainty related to the outlier

size, shape, and variability across space, a concern is the concept of how the analysis of such features across a range of scales is influenced by spatial dependence (Tobler, 1970).

To test representations of spatial outliers with varying degrees of spatial dependence and uncertainty, simulated outliers were modeled with the Gaussian function. Different shapes of outliers were controlled with two parameters, the height and width. Samples from super-fine resolution "source" data were aggregated into coarser resolution raster representations of outliers. To further investigate variability and scale issues arising from the aggregation process, two additional variables were included, the raster cell resolution and the assignment operator. In total, there were four variables employed for the spatial outlier simulations: the raster grid resolution, Gaussian height, Gaussian width, and the assignment operator. Applied to grids representing the simulated outliers were three standard spatial outlier detection methods: the *Median* method of Lu et al. (2004), the $Z - test$ of Shekhar et al., and the $SLOM$ method of Chawla and Sun (2005).

The quantitative detection results are varied. The outlier height had little influence on the results. With a single outlier in isolation, the presence of any value was sufficient for detection. Regarding cell assignment, the best detection of outlier extremeness was found with the best representations, by matching the *maximum* operator to positive-signed values.

The spatial outlier detection algorithms performed well for outliers represented by one or a few locally extreme values compared to a given neighborhood. That characteristic, for which the detection methods were designed, is illustrated in Figure 16. In other words, if the scale of the outlier matches the scale of analysis (i.e. spatial resolution), then the tested detection methods perform well.

As designed, the detection methods tested are not suitable for a mismatch between the scale of the spatial outliers and the scale of analysis. In this experiment, the break down of classifiers was exemplified in scenarios with broad spatial outliers, having large widths that span many grid cells. This key finding is also related to the qualitative results. Among a diversity of complex labeled outlier patterns, the top regions of the outliers were always missed. This is due to the spatial dependence of the outlier tops, particularly of wide outliers with similar values spanning, or swamping, multiple cells. To address this issue, a multi-scale outlier detection method is suggested, one that includes specifications related to parts of an outlier, such as a top region.

4 Spatial dependence of anomalies

4.1 Introduction

4.1.1 Previous work

In the previous chapter common spatial outlier detection methods were evaluated with a focus on the issue of the scale. There were issues related to a mismatch between the scale of analysis, or the spatial resolution, and the scale of the outliers. Wide outliers, with broad spatial extents spanning across the space represented by multiple observations, had a characteristic missed by the common detection algorithms. The top regions near and including the location of the peak value were rarely detected due to the relative similarity of those values, or high local spatial autocorrelation.

4.1.2 Research question and implications

This section is focused on the following research question: is high spatial autocorrelation associated with wide spatial outliers? It is an arguably counterintuitive concept, that an anomaly is partially composed of an area with high spatial dependence. If true, an implication is that a conceptual model of a spatial outlier might include a component of high spatial dependence. That measurable and quantifiable characteristic might be beneficial for identifying locations of wide outliers in spatial data.

4.1.3 Controlled study, outliers in a field

In the previous chapter, controlled data sets of isolated outliers mitigated the problematic complexities of identifying spatial outliers in real data. However, toward a future goal of

detecting wide spatial outliers, an important aspect of this study involves a comparison between spatial outliers and their spatial context. Specifically, the extended aim is to investigate the potential of patterns of high spatial autocorrelation to distinguish spatial outliers from their surroundings.

As a means of conducting the controlled experiment in this section, simulated background fields were created, in which various sets of spatial outliers were embedded. The degree of spatial autocorrelation is not only measured in the region influenced by the spatial anomaly, within the outlier, but also across the variation of the background field.

This controlled study, as in the previous chapter, involved simulated outliers of known Gaussian shapes, with specified values for the height and width parameters. As opposed to a single isolated outlier for each pair of Gaussian parameters, sets of outliers were created across the extents of a field. Outlier locations were selected randomly with only two restrictions. First, a gap between outliers was imposed so they would not overlap or influence their immediately surrounding background field. Second, outliers were allowed to fill only a minority of the area.

Each set of outliers was embedded into a variety of background fields, which were created by means of unconditional Gaussian simulation (Gómez-Hernández & Journel, 1993; Pebesma, 2004). The cells of the fields contained values assigned by Gaussian probabilities and a model of spatial variability. The geostatistical model of spatial variability was based upon a variogram structure with variable values specified for the nugget, partial sill, and range parameters.

4.1.4 Alternative approach, significantly different spatial autocorrelation

There are potentially many approaches to statistically evaluate whether high spatial autocorrelation is common to wide outliers. For example, one alternative is to test whether the degree of spatial autocorrelation observed within wide outliers, in the top region, is significantly different than at random locations. Since the locations and properties of simulated outliers are known, the degree of spatial autocorrelation could be measured and compared between those locations and random locations.

Although this experiment includes outliers of known location, for which the degree of spatial autocorrelation can be measured, the simulated fields also have the potential for containing outliers at unknown locations. The fields were designed to have regions that might have patches of similarity. Some of those patches might follow the hypothesized pattern of wide spatial outliers, a patch of similar values that are extremely different than the surrounding values. Therefore, it is expected that there is a potential for outliers at unknown locations throughout the simulated field.

The two groups to compare, values of samples at known outlier locations and those at random locations, would not be disjoint with respect to the presence or absence of an outlier. The random samples potentially include characteristics of outliers at unknown locations. A statistical comparison would be confused by this error of commission.

4.1.5 Selected approach, comparison of sampling results

Another approach was selected. A test was conducted to compare the performance of various sampling methods in "finding" spatial outliers. Based upon the region subset as two

classes of high and low local spatial autocorrelation, one sampling method incorporates, and gives preference to, regions with the characteristic of high spatial autocorrelation. The other methods are more random or regular in design. A description of these three components, the local spatial autocorrelation metric, the sampling by unequal probabilities, and the other sampling methods, along with an description to the performance evaluation proceeds below.

4.1.6 Local Moran's I, local spatial autocorrelation

The metric employed to indicate the degree of spatial autocorrelation was the Local Moran's I value (Anselin, 1995). Simply stated, the Local Moran's I is a measure of local similarity, or the spatial clustering of similar values. A higher Local Moran's I value computed for a local neighborhood indicates that the region includes data, nearby one another, that are more similar than a region with mixed values. In other words, the Local Moran's I indicates nearby values are spatially dependent upon one another.

The Local Moran's I is applicable to data having known attribute values and, importantly, topological relationships. For example, the data are typically embedded in some spatial reference system, although graph connectivity (e.g. for networks) or other relationship definitions are also possible. In the case of this experiment, the ordered structure of the raster grid provides the topological connectivities and relationships between each pair of cells.

There are two special requirements for calculating the Local Moran's I. First, the definition of a neighborhood is required. A neighborhood definition produces a list of neighbors by a criterion such as distance, adjacency, or network linkage. A limited num-

ber of neighbors based upon any such criterion is also common. In this experiment the neighborhood of any central cell of analysis is defined as the adjacent cells that share an edge or a vertex (i.e. the queen's case neighborhood).

The other requirement is a set of weights for the neighbors. The weights are to adjust the amount of contribution each neighbor has to the Local Moran's I value of each local neighborhood. This can be specified by a number of schemes such as equal weighting, binary assignment, standardized values, or distance-based. A row-standardized method, that sums to the number of neighbors, was employed in this experiment. It gives near equal weighting to the adjacent neighbors and provides an adjustment, in aggregate, for the minority of neighborhood with fewer neighbors, along the edges and at the corners of the raster grid.

Equation 7 is the general formula for the Local Moran's I. Equation 8 is the R package *spdep* implementation for the *localmoran* function, employed in this experiment.

$$I_i = (x_i - \bar{x}) \sum_{j=1}^{n} w_{ij}(x_j - \bar{x}) \tag{7}$$

$$I_i = \frac{(x_i - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2/(n-1)} \sum_{j=1}^{n} w_{ij}(x_j - \bar{x}) \tag{8}$$

where I is the Local Moran's I value for each point of analysis (i) computed from the set of attribute values (x) and weights (w) of each neighbor (j) in each neighborhood (k). The implementation in the R package *spdep* modifies the product of a) a datum's difference from the mean of the neighborhood and b) the sum of its neighbors' weighted differences from the mean, by standardizing that product by the variance of all neighborhoods, a constant value, following Sokal, Oden, and Thomson (2008).

4.1.7 Local Moran's I, probability sampling

The Local Moran's I value was used in a probability sampling approach. The foundation of this approach is based upon random sampling methods. It is only different to the equal probabilities of simple random sampling, by the application of unequal selection probabilities. In this experiment, the selection probabilities are specified by a sampling factor for each datum that adjusts the probability of sampling that datum, relative to the other data. This sampling method (indicated by $LocalMoran's I$ from here forward) employs just two probabilities, one for each class of local spatial autocorrelation, determined by a $LocalMoran's I$ classification boundary. The value of the probability for the higher class of spatial autocorrelation was set relative to the lower class by means of the $LocalMoran's I$ sampling factor. The $LocalMoran's I$ sampling factor, therefore, intentionally induces bias in the sampling.

In this experiment, two sampling factors were specified, each applied to either of two classes of cells. First, the Local Moran's I value was calculated for each and every raster cell in the region. Then, based upon an arbitrary Local Moran's I value, the cells were partitioned into two classes of high and low Local Moran's I values. The $LocalMoran's I$ sampling was performed with one of two sampling probabilities applied to each raster cell. The higher sampling factor was applied to cells of the high Local Moran's I class to increase the probability of sampling those locations. Similar to purposive sampling, the application of unequal probabilities is intended to find whether the locations in the class of high spatial autocorrelation represent portions, or parts, of the known spatial outlier features.

The results of the $LocalMoran's I$ sampling, with preference to locations having high spatial autocorrelation, were compared to the results of unbiased sampling. The intent of this comparison is to investigate whether the property of high local spatial autocorrelation occurs more frequently in outliers than at random. Of particular interest is whether locations of outliers are found more frequently with the $LocalMoran's I$ sampling than with unbiased sampling.

The $LocalMoran's I$ sampling is the only method with unequal selection weights. Similar to purposive, judgmental, or critical case sampling, the unequal weights were intended to sample at a higher frequency from a particular class, that of high spatial autocorrelation. If samples were found coincident with spatial outliers at a higher frequency with $LocalMoran's I$ sampling than with any sampling with equal weights, an inference is that the spatial outlier features have a component of high local spatial autocorrelation within their spatial extents.

4.1.8 Three unbiased sampling strategies

For the comparison, three other unbiased sampling strategies were selected. Each of these methods employ equal selection probabilities across the sampling frame. As such, they are often used for statistical analyses that assume independence of the observed attribute values. As described by Berry and Baker (1968), they are: *random*, spatially-stratified random (called $stratified$ from here forward), and systematic (called *regular* from here forward). *Random* sampling is simply a selection of arbitrary locations throughout the spatial extents. $Stratified$ sampling is intended to avoid the over-sampling and under-sampling of various regions that might occur with simple random sampling. To do so, the stratified sampling process first partitions the space into regular areas, such as a

square grid. Then, in each area, an equal number of random samples are taken. With a *regular* sampling strategy, samples are taken across the space at equal interval distances. Examples are shown in Figure 20.

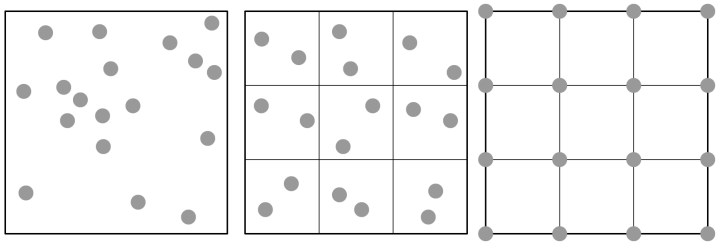


Figure 20: Of the four sampling strategies compared, the three selected unbiased methods are, from left to right, *random*, *stratified*, and *regular*.

4.1.9 Performance evaluation, two metrics

Two quantitative metrics are employed for the comparison of the sampling strategies. The first is the proportion of the samples on outliers. This is the ratio of the count of sampled locations coincident with outliers to the count of total samples. This metric indicates the degree to which a sampling strategy "finds" outliers. This is the main comparison with regard to indicating whether spatial autocorrelation is associated with outliers.

However, there exists the chance that any outlier is sampled multiple times and another is not sampled at all. To investigate whether a sampling method finds a high number of outliers another metric is also employed. The second metric is the proportion of unique known outliers found by samples. This is the ratio of the count of samples coincident with unique outliers to the count of total outliers. This provides an indication of whether spatial outliers consistently have a characteristic of spatial autocorrelation.

Examined together, the two metrics allow for a more comprehensive analysis of whether spatial autocorrelation is associated with outliers.

In addition to any spatial autocorrelation in the structure of the outlier, the experimental design also includes fields exhibiting varying degrees of spatial autocorrelation. This enables a controlled evaluation of whether the spatial autocorrelation of the outlier is distinctive with respect to its spatial context, as embedded in fields having different patterns of variability.

Objective results are provided by the quantitative metrics. However, by chance, the values of any single sampling process might have selected locations with unusual characteristics. Instead of relying on one realization of the process for each set of outlier and field parameters, multiple sampling iterations were conducted to strengthen a conclusion about the relationship between the background field, the embedded outliers, the sampling strategies, and the resulting metrics. This was to balance an aim of finding outliers with small samples with a need to collect sufficient information for comparing patterns of generalized of results, as mean values with confidence intervals.

4.1.10 Expectation of comparison results

This experiment compared results of the $LocalMoran'sI$ sampling to $random$, $stratified$, and $regular$ sampling. Although the unbiased methods are expected to sample outliers occasionally, the $LocalMoran'sI$ sampling preference for high spatial autocorrelation is expected to sample outliers more often. If the $LocalMoran'sI$ method consistently samples spatial outliers more often than the other methods, it indicates that wide spatial outliers partially consist of regions with relatively high spatial autocorrelation.

4.2 Methods

4.2.1 Overview

This experiment was conducted on numerous synthetic realizations of fields with different patterns of spatial variability. Embedded, or mixed, into the variable fields were spatial outliers of known location and Gaussian shape. The Local Moran's I value was computed at each cell from its immediate neighborhood and used in one of the four sampling strategies compared. The $LocalMoran'sI$ probability sampling, with higher selection probabilities for high Local Moran's I values, was compared with equal selection probability *random*, $stratified$, and *regular* sampling methods. The comparison covered two quantitative metrics, the proportion of samples on outliers and the proportion of unique outliers found. For each mixed field, 100 sampling iterations were conducted for a generalized analysis of the means and confidence intervals of the two metrics with respect to each sampling strategy.

4.2.2 Software and general data

The generation of the outlier simulations and the analyses of the results were performed with the statistical software R (R Development Core Team, 2013). Three R packages were particularly important for this experiment. The *gstat* package enabled the creation of the background fields, the *spdep* package was employed for computing the Local Moran's I values across the fields, and the *sp* package provided various sampling methods.

Each data set represented a field as a two-dimensional raster grid structure with both

spatial extents ranging from zero to 100 units. The resolution was one unit per cell, creating grids of 100 × 100 cells.

4.2.3 Outliers

Simulations of multiple spatial outliers were created across the fields. Outlier values were assigned using the Gaussian function as generally described in the previous experiment, in section 3.2.2 "Gaussian shape" on page 49. The only differences were the specified Gaussian height and width parameter values. The Gaussian height parameter values were incremented across the set {10, 20, 30}. The Gaussian width (i.e. the spread, or standard deviation) parameter values were incremented across the set {1/3, 2/3, 3/3}. It is worth noting the size of the reference outliers. With a reference outlier defined as three standard deviations radially from the center, the outlier width parameters result in reference outlier diameters six times larger with extents in the set of {2, 4, 6}, respectively. The three values each for the two parameters specified 9 Gaussian shapes.

Although the placement of multiple outliers across the extents of the field was mostly by random selection, locations were controlled by two restrictions. First, outliers were not allowed to overlap. A gap of one reference outlier width (diameter) was enforced around each outlier. The second control of outlier placement was with regard to the total area covered by outliers. Only a minority of the area, 10%, was allowed to represent known outliers.

For the performance evaluation, outlier reference grids were created for each configuration of placed outliers. The creation of outlier reference grids generally follows the methods of the previous experiment (described in section 3.2.9 "Reference label grids" on

page 53). The extent of a reference outlier was defined as three standard deviations from the center of the Gaussian shape. Therefore, cells of the reference grids were labeled as reference outlier cells if they intersected a circle with a radius three times the Gaussian width parameter (i.e. three standard deviations) from the center. However, there were two extensions added for this experiment. First, a larger field included multiple outliers. Second, cells of each reference outlier were labeled with a unique outlier identifier to enable the computation of the unique outliers found performance metric, described later.

4.2.4 Fields

Background fields were generated by unconditional Gaussian simulation. Differing levels of spatial autocorrelation were induced by employing exponential variogram models with variable nugget, partial sill, and range parameters. The R package *spdep* was employed to specify a variogram model with the function *vgm*. The model form was of the type *Exp* (exponential). Each permutation of parameter values for the nugget {3, 6, 9}, the partial sill {10, 20, 30}, and the range {10, 15, 20} specified the 27 variogram models.

Simulations of fields were produced with the R package *gstat* and the function of the same name. With each of the 27 variogram models, an *nmax* (maximum number of neighbors) of eight, and a *beta* (height constant) of zero specified 27 gstat objects. Each of the objects were submitted to the *predict* function to create 27 field simulations.

In the computational loop of parameter value permutations, each iteration created a set of outliers and a field. The outliers were embedded into the fields by addition. It is important to clarify that the outliers did not replace values in the field. Rather, all of the variation existing in the original field at outlier locations remained in the resulting

mixture. Figure 21 illustrates an example of mixing the outliers into the fields, producing a mixture grid. All permutations of parameters for the 9 outlier shapes and 27 field variations resulted in 243 mixture grids.

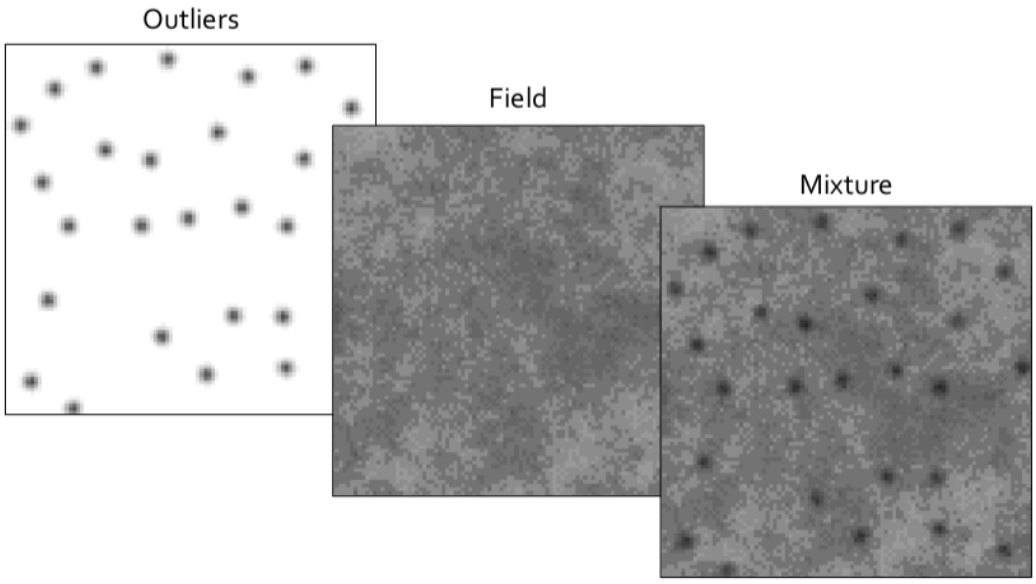


Figure 21: Example of mixing the outliers into the fields, producing a mixture grid.

4.2.5 Local Moran's I sampling method

The mixture grids were each used as the basis for all variations of the sampling processes. One of the sampling strategies, the $LocalMoran'sI$ sampling, incorporated information about local spatial autocorrelation. The metric employed to indicate the degree of spatial autocorrelation was the Local Moran's I value (Anselin, 1995), discussed in section 4.1.6 "Local Moran's I, local spatial autocorrelation".

The R package *spdep* was employed for computing the Local Moran's I values for each cell of each mixture grid. The $cell2nb$ function created a listing of neighbors for each cell

generally following the queen's case (eight cells), except along edges (five cells), and at corners (three cells). The *nb2listw* function assigned weights for each neighbor of each mixture grid cell. By the default W method, row-standardized values were assigned so all of the weights summed to the number of neighbors. The weights and the mixture grid values were submitted to the *localmoran* function to compute the Local Moran's I value at each cell location, thereby creating a Local Moran's I grid corresponding to each mixture grid.

The $LocalMoran's I$ probability sampling was performed with the *sample* function of the *base* R package. The function includes a *prob* parameter to specify the probability that each cell might be sampled relative to other cells in the data set.

For the $LocalMoran's I$ sampling each cell was assigned one of two sampling factors per its membership in a class of either low or high spatial autocorrelation. A $LocalMoran's I$ classification boundary parameter was employed to differentiate and subset the values into two classes, creating a Local Moran's I grid. To cover an effective range for this parameter, various quantile values were specified, incrementing across the set {0.7, 0.8, 0.9}, corresponding to percentiles in the set {70th, 80th, 90th}. From each Local Moran's I grid three Local Moran's I classified grids were created. Local Moran's I grid values above the Local Moran's I classification boundary quantile value were labeled as the class of high spatial autocorrelation, the remainder of cells were labeled as the class of low spatial autocorrelation.

The low spatial autocorrelation cells were each assigned a sampling factor of one. To cover an effective range for the Local Moran's I sampling factor, values were incremented across the set {10, 50, 90}. By means of applying these three Local Moran's I sampling

factors, from each Local Moran's I classified grid three Local Moran's I probability grids were created. The Local Moran's I sampling factor was thereby specified to potentially increase sampling of cells with high spatial autocorrelation (in separate tests) 10, 50, or 90 times more than the low spatial autocorrelation cells.

Due to these two variables, the Local Moran's I classification boundary and the Local Moran's I sampling factor, the number of potential results again increased. From the 243 mixture grids, and then the 243 corresponding grids of Local Moran's I values, three values for each of these two variables increased the number of permutations by a factor of nine to 2,187 variations.

4.2.6 Three unbiased sampling methods

The three unbiased sampling strategies, compared to the $LocalMoran'sI$ sampling, were the $random$, $stratified$, and the $regular$ strategies. The R package *sp* provided these three methods through the *spsample* function. Each of these three sampling methods was performed by specifying them for the *type* parameter, indicated by their name denoted here.

4.2.7 Samples and performance

The sampling performance was tested with small samples, such as 1%. However, due to a concern for sampling rare events and to ensure adequate sampling of the small outliers a range of values were tested. The sample size for all four strategies varied across percentage values in the set {1, 2, 3}. Including three sample size values with the previous set of permutations, the total permutations of all variables was 6,561.

As previously mentioned, any individual sample might observe unusual patterns in the data, or possibly miss rare occurrences of spatial outliers. For this reason, 100 iterations of each sampling method were performed on each of the 6,561 permutations of variable values.

For each iteration of each sampling method, two performance metrics were computed. The first metric, the samples on outliers, is the ratio of the count of sampled locations coincident with outliers to the count of total samples. Any sample locations that intersected a reference outlier were counted. The second metric, unique outliers found (by samples), is the ratio of the count of unique outliers coincident with sample to the count of total outliers. Any sample that intersected a uniquely labeled outlier in the reference grid was counted.

For the 6,561 variable permutations, 100 iterations of each of the four sampling processes were performed. From the iterations of each sampling method, the mean and 97.5% confidence intervals of the two metrics were calculated for the comparison.

4.2.8 Summary of parameters, permutations, and metrics

Table 3 shows a listing of the three values for each of the eight variable parameters, resulting in 6,561 permutations. Applied to each permutation were 100 iterations of each of the four sampling strategies. From each set of 100 iterations, the mean and 97.5% confidence intervals were calculated for each of the two performance metrics.

Values	Parameter
{1/3, 2/3, 3/3}	Outlier width
{10, 20, 30}	Outlier height
{3, 6, 9}	Field nugget
{10, 20, 30}	Field partial sill
{10, 15, 20}	Field range
{0.7, 0.8, 0.9}	Local Moran's I classification boundary (quantiles)
{10, 50, 90}	Local Moran's I probability sampling factor
{1, 2, 3}	Sample size (percents)

Table 3: Variable parameters and values for this experiment.

4.2.9 Visualize a small set of results, grids

A visualization was created to present a means of reviewing a small set of results. The metrics of proportions of samples on outliers and unique outliers found are represented graphically in grids. Each square grid was composed of nine cells (3×3). Darker tones in the cells indicate higher values, or higher performance. In each grid, along the x-axis are increasing values of the outlier width parameter and increasing sample sizes are along the y-axis, however, any pair of the variables are possible for this type of visualization. Grids for each of the four sampling methods and for each of the two performance metrics were created.

4.2.10 Visualize the entire set of results, nested loop plots

Nested loop plots (Rücker & Schwarzer, 2014) were employed to present the entire set of 6,561 results of each performance metric for each sampling method. The preparation process involves a regression analysis to rank (by magnitude) and sort (by sign) each of the variables by their degree of correlation with the variation in the dependent variable, the results. The y-axis is the value of the metric. For each sampling method the mean and the 97.5% confidence interval of the 100 sampling iterations were graphed. The x-axis is

partitioned, in an ordered sequence, by the number of values of each variable, starting with the variable with the highest correlation with the variation of the results. A legend indicates the order of variables (ordered with the largest partition given to the variable with most correlation), their values (three each), and their sign (increasing or decreasing).

4.3 Results and Discussion

4.3.1 Example of sampling with Local Moran's I

The following example is to illustrate the $LocalMoran'sI$ probability sampling approach. A binary classification was performed on a mixture grid (of outliers embedded in a variable field). The classes were subset with a $LocalMoran'sI$ classification boundary value. Figure 22 is an example of a binary classification based upon a boundary value representing the 50th percentile of the range of Local Moran's I values. Cells with low spatial autocorrelation, up to and including the 50th percentile are white. Cells with higher Local Moran's I values are gray.

A sampling factor was applied to each cell based upon its class. Cells with low spatial autocorrelation were assigned the base value of one. The larger $LocalMoran'sI$ sampling factor was applied to the cells with high spatial autocorrelation in order to increase the probability of sampling cells with high spatial autocorrelation.

Figure 23 illustrates an example comparison of samples (colored dots) on outliers (light blue semi-transparent circles) by all four sampling strategies: *random* (green), $stratified$ (blue) , *regular* (red), and the $LocalMoran'sI$ probabilistic sampling (black). For visualization purposes, the sample sizes in this example are small. As a precursor to

Figure 22: Example of the Local Moran's I binary classification, based upon the 50th percentile boundary value.

computing the performance metrics involving proportions of samples, counts were taken of the number of samples on outliers ("hits") and unique outliers found ("unique") for each of the four sampling methods of the example.

The counts of samples on outliers and of unique outliers found both clearly indicate that the $LocalMoran'sI$ sampling found more wide outliers than the other three sampling strategies. That is an indication that spatial outliers might be found by their high spatial autocorrelation. That finding is further strengthened by a visual inspection. Every wide outlier in the example has a region with high local spatial autocorrelation (in gray) that is near the center of the outlier, corresponding to the region around the peak value. That every outlier has this pattern, with a patch of cells in the class of the top half of Local Moran's I values across the entire field, indicates that parts of wide outliers can be found by this characteristic, even relative to their spatial context in a variable field. The

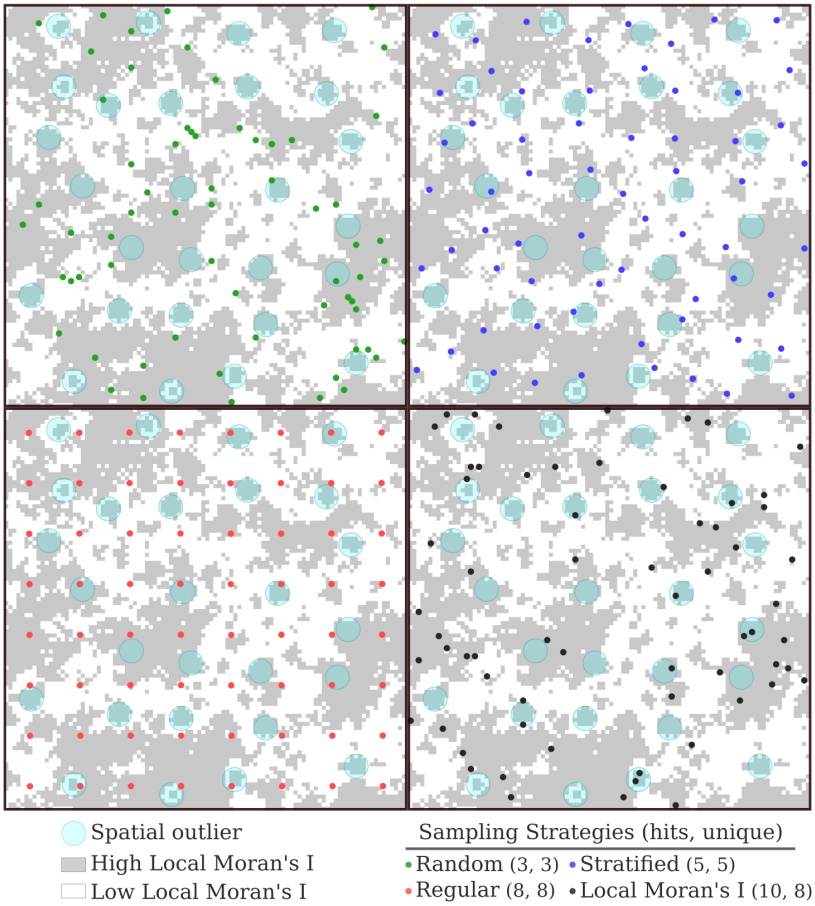


Figure 23: Example comparison of each sampling strategy's performance of samples on outliers (hits). The $LocalMoran'sI$ probabilistic sampling increases samples in regions with high spatial autocorrelation and finds more outliers.

visual example provides an illustration to guide interpretation of the experimental results.

4.3.2 Visualization of a small set of results, grids

A visualization of a small set of the results is illustrated in Figure 24. Values for both of the metrics, samples on outliers and unique outliers found, are represented by shades of gray. Darker tones in the grid cells indicate higher proportions. Highlighted with yellow boxes are performance results with consistently high values from the $LocalMoran'sI$

method, as compared to the other methods. For each metric of each sampling method, only two of the eight variables are represented, sample size and outlier width. With three values for each variable, the shades of the nine grid cells provide a visual cue for the comparison of the performance of each method in response to the change in two variables.

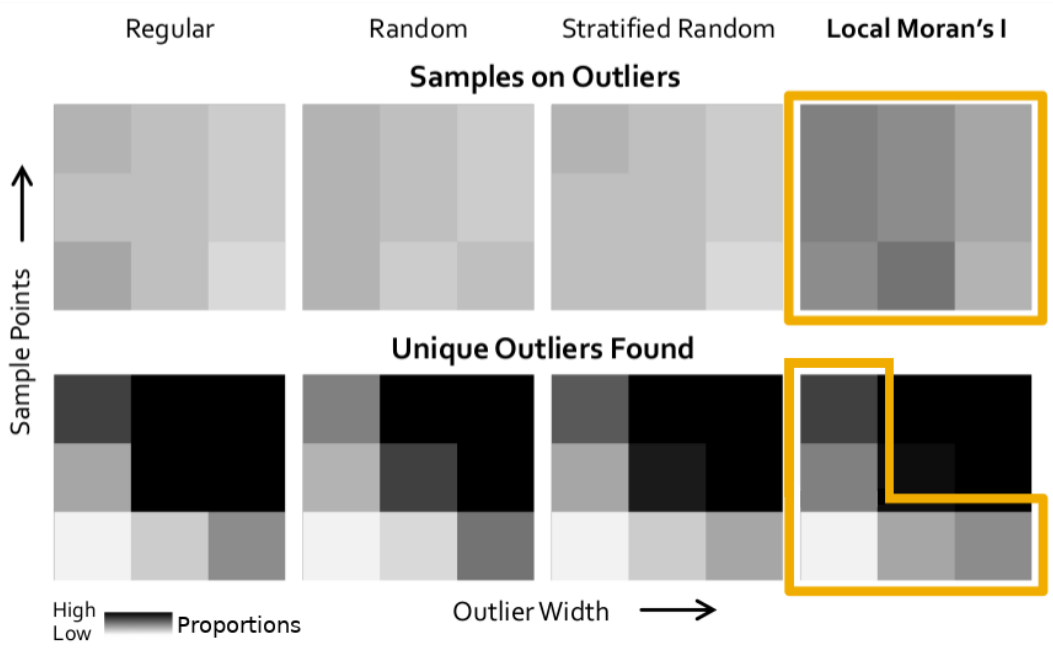


Figure 24: Example of visualizing a small set of the results for both of the metrics, proportions of samples on outliers and of unique outliers found. Darker tones in the grid cells indicate higher proportions. Highlighted with yellow frames are performance results with consistently high values from the $LocalMoran's I$ method, as compared to the other methods.

The small set of results further support the higher performance of the $LocalMoran's I$. The figure shows results from all four sampling methods. In the first set, representing proportions of samples on outliers, the grid cells for $LocalMoran's I$ are all darker than the corresponding cells of the other methods. Evidently, the inclusion of information about the characteristic of high local spatial autocorrelation into the $LocalMoran's I$ sampling increased its performance rates. Regardless of the sample size or outlier width,

sampling with the $LocalMoran's I$ method finds outliers at a higher rate than the other three methods.

It can be inferred that outliers consistently include in the class of cells representing regions of relatively high local spatial autocorrelation. This finding is dependent upon the contrast in variation between the outlier and the field, based on values of their respective parameters. However, the darker gray cells in this visualization specifically represent underlying quantitative values of the central patches of local spatial autocorrelation within outliers noted in the previous $LocalMoran's I$ example.

The second set of results are for unique outliers found. This is important with respect to finding higher numbers of the known outliers. With near black in the four upper-right cells of all the grids any differentiation is difficult to generalize. This pattern is simply an indication that the methods are successful at finding high numbers of the wide outliers with larger sample sizes. However, that is particularly so for the *regular* and the $LocalMoran's I$ methods.

Outside of that quarter, there are two patterns of differentiation possible with regard to either smaller samples or smaller outliers. First, for all sample sizes of only the small outlier widths, both the *regular* and the $LocalMoran's I$ methods again have higher performance, with the $LocalMoran's I$ slightly better. Second, the $LocalMoran's I$ also outperforms all other methods for small sample sizes. This indicates that the information about high local spatial autocorrelation included in the $LocalMoran's I$ method is beneficial for finding outliers consistently throughout a data set, regardless of variation in their spatial context; it does so even with small sample sizes. Although the spatial autocorrelation in an outlier is relative to the field and the field variability might mask

outliers, searching regions of high spatial autocorrelation consistently yields higher numbers of outliers than the sample methods that employ equal selection probabilities.

4.3.3 Visualization of the entire set of results, nested loop plots

Two nested loop plots were generated based upon the methods of Rücker and Schwarzer (2014). The nested loop plots show all of the results from the nested loops in the computational process, or all permutation results. Figure 25 presents the nested loop plot of all 6,561 permutation results of each of the four sampling strategies, with respect to the proportion of samples on outliers. Similarly, Figure 26 presents the nested loop plot for the proportion of unique outliers found. Graphed on each plot are the results of the 6,561 permutations. The values graphed are the mean and the 97.5% confidence intervals of each metric, calculated from 100 sampling iterations of the four sampling methods: *regular* (red), *stratified* (blue), *random* (green), and "Local Moran's I" probability sampling (gray).

With the result value indicated along the y-axis, the results of all the parameter sets are ordered along the x-axis in nested partitions. Basically, each partition level across the x-axis represents one of the variables, which are ordered by the magnitude and sign of correlation between a variable and the results. The x-axis is first partitioned by the number of values of the variable having the largest degree of correlation to the result values. Each of the largest partitioned sections are then partitioned again to represent values of the next highest correlated variable, and so on. Each of the partitions at any level represents results, with respect to a single value of that variable, which are ordered by the sign of the correlation, either increasing for positive correlation or decreasing for negative. A graphical legend included on the graph indicates each variable, its values,

and the x-axis span and signed order of results in each level of the sequence of partitions.

There is abundant information contained in the nested loop plots. Analysis of the plots can potentially involve any combination of subsets of the variables or their values. In Figure 25, which presents the proportion of samples on outliers, two patterns of results are readily visible. First, a sawtooth pattern, increasing in magnitude from left to right, represents the $LocalMoran's I$ results. With respect to the y-axis magnitude, this method has performance values fluctuating between about 10% to 50%. The values are significantly higher than those of the other three methods, all hovering around or below 10%. Because, for this metric, the *random*, $stratified$, and *regular* methods all perform relatively equally, their graphed lines overlap. This second overlapping and relatively linear pattern makes it difficult to discern the distinct color representing each of these three common sampling methods.

Although not the major pattern, there is a counterintuitive result. The $LocalMoran's I$ sampling results in a higher proportion of samples on outliers with higher $LocalMoran's I$ classification boundary percentiles. Increasing the classification boundary value reduces the percent of area covered by the high Local Moran's I class. It might be expected that a smaller class area would lead to less area in the outliers and therefore fewer samples. However, an explanation might be that the central regions of the wide outliers remain represented by that class, due to relatively high spatial autocorrelation. They are, therefore, sampled with the high selection probabilities from a smaller proportion of cells. The results of increasing samples on outliers might be due in part to the sampling strategy, but also in part to a relatively high degree of local spatial autocorrelation maintained within the extents of the outliers.

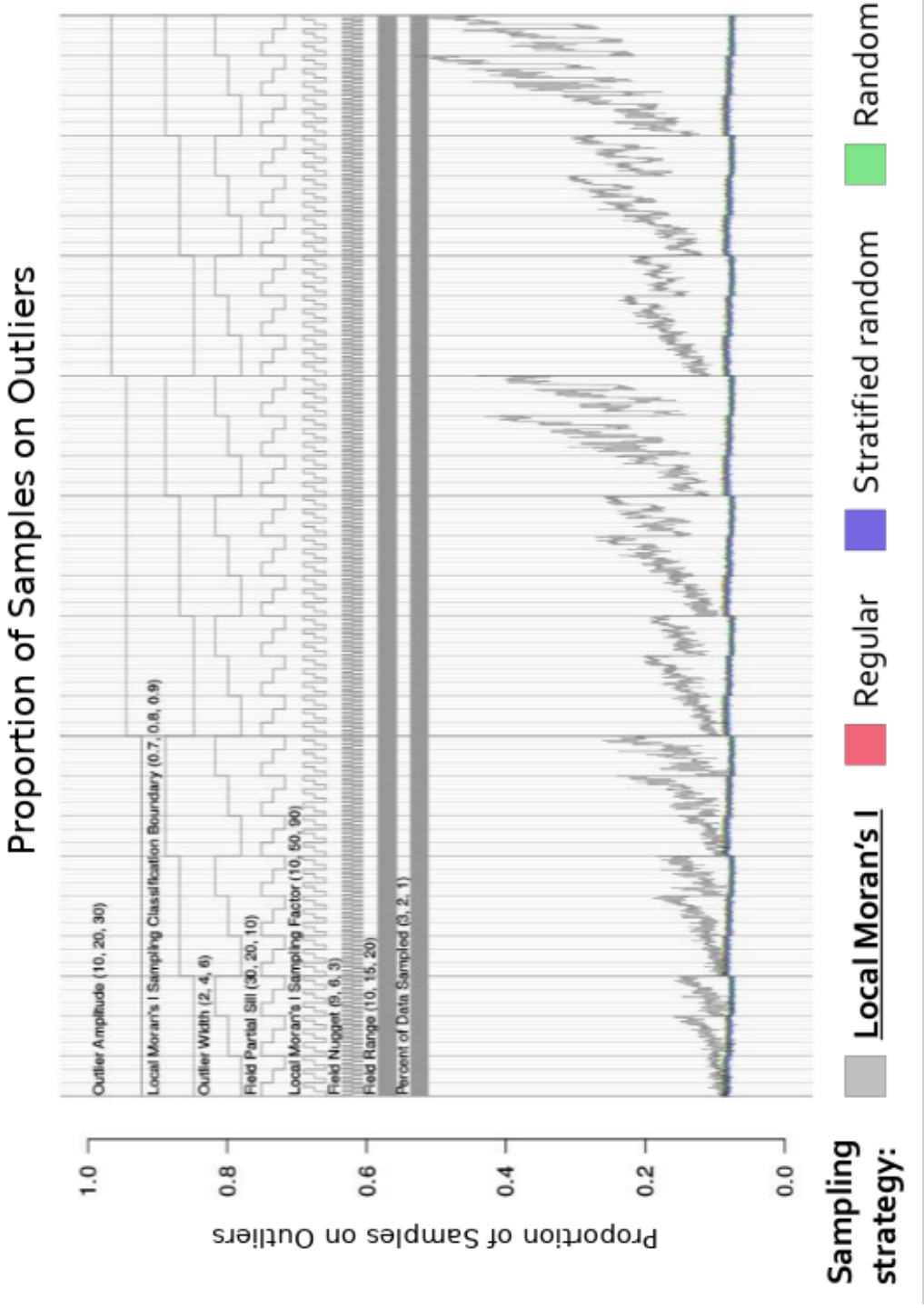


Figure 25: Nested loop plot of results of each sampling strategy, with respect to the proportion of samples on outliers.

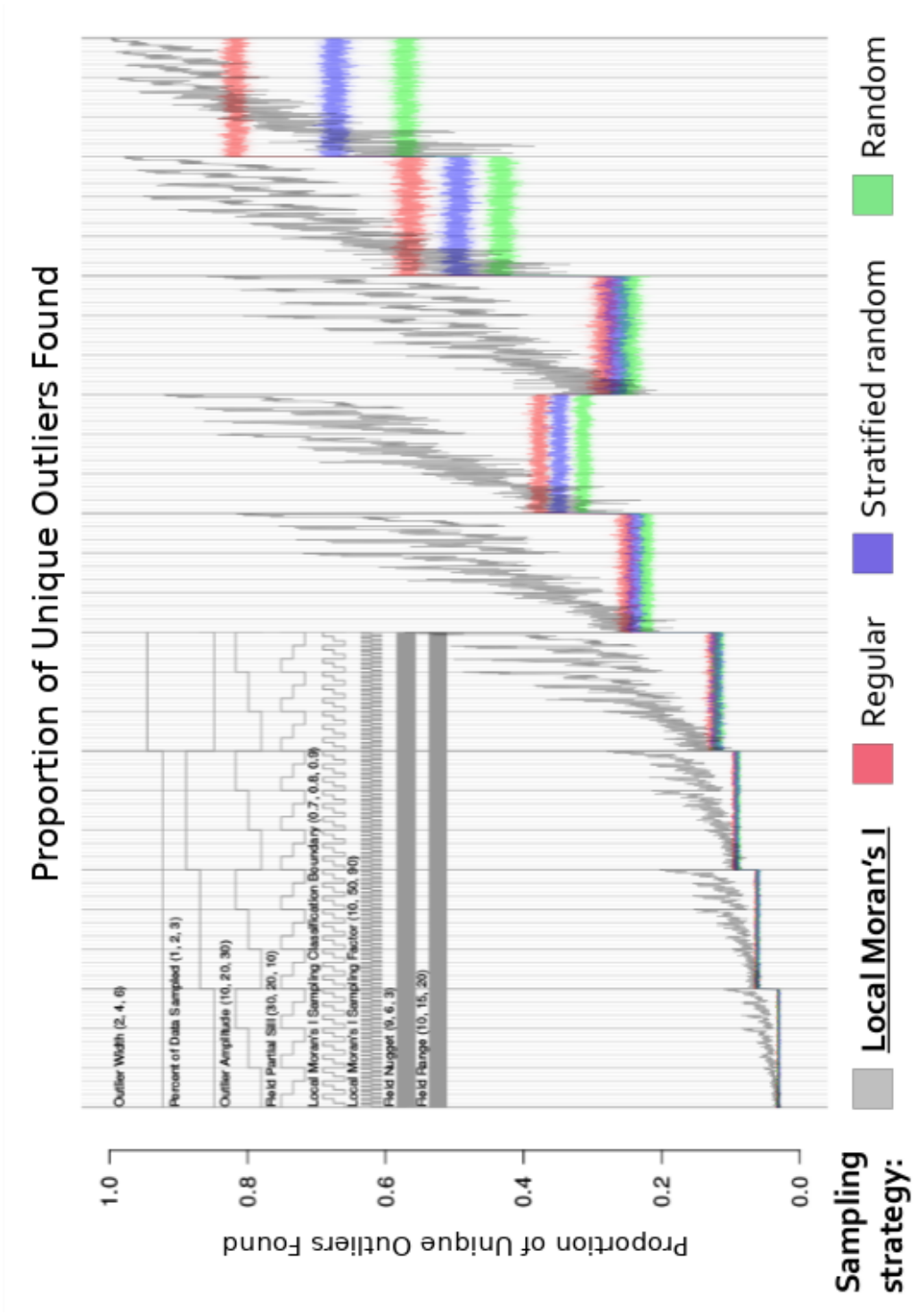


Figure 26: Nested loop plot of results of each sampling strategy, with respect to the proportion of unique outliers found.

In Figure 26, which presents the proportion of unique outliers found, the four sampling methods produce distinctive patterns of results. The sawtooth response of $LocalMoran's I$ results is again present, but is smoother and generally increases from below 5% to up around the maximum of 100% of unique outliers found. The three other sampling methods show more response for this second metric. There is a near constant linear pattern across most variables below the first two partition levels. However, the linear bars of y-axis variation are thicker with respect to increasing outlier width. In addition, the linear bars have higher values in response to both increasing outlier width and the increasing sample size. Across the three common sampling methods is a consistent ranking, from highest to lowest performance, of *regular*, $stratified$, and *random* sampling.

Some overlap exists with the three common methods until some divergence of results is evident with scenarios of higher sample size or outlier width. In those same two scenarios, the $LocalMoran's I$ has some overlap with the other methods, specifically with lower outlier height, except where the lower outlier height maintains its largest potential contrast with the lowest partial sill. The major pattern to note, however, is that the $LocalMoran's I$ generally finds a higher proportions of unique outliers, especially with the smaller sample sizes.

A notable result is that both plots show an increase in $LocalMoran's I$ sampling performance with increasing outlier width. This is related to the initial research problem, as wide outliers with large areas of local spatial autocorrelation are often missed by the common spatial outlier detection methods. However, these results indicate wide spatial outliers can be found consistently by leveraging high local spatial autocorrelation. A model of a wide outlier might, therefore, include the pattern of spatial dependence within the extents of the anomaly to better represent components of a spatial outlier

over various scales.

Although the nested loop plots, or subsets of the results, provide abundant information, they also offer value for direct analysis of general patterns. Taking results of all permutations together, the most important pattern is directly relevant to the research question of interest. Both plots show that wide spatial outliers, spanning several spatially referenced observations, are associated with the characteristic of high local spatial autocorrelation. The $LocalMoran'sI$ sampling results generally have higher performance than other methods with respect to both metrics, the proportions of samples on outliers and of unique outliers found. Together, these two sets of results suggest that high spatial autocorrelation is associated with wide spatial outliers.

4.4 Conclusion

The previous chapter indicated that common spatial outlier detection methods failed to label as outliers the central, top region of wide spatial outliers, which span across the space represented by many observations. If the scale of analysis, or resolution, does not match the scale of a wide spatial outlier, the methods miss the central region around the peak value, the top of a wide spatial outlier. The hypothesis of this experiment is that a wide spatial outlier is partially characterized by a region of high local spatial autocorrelation. If so, this characteristic might be included in a model of a spatial outlier over various scales.

The approach taken in this experiment was facilitated by the simulation of outliers in a variable field. Each wide outlier, simulated with a Gaussian shape, perturbed a

variable background field. One benefit of simulating outliers is the known information about the shape and location of the simulated outliers. This mitigates problems related to the complexity of defining features in the landscape, which have ambiguous boundaries, complex patterns, and various other uncertainties. With random but known outlier locations, their characteristics can be compared relative to the variable background field.

A comparison was made between the sampling performance of different random sampling methods. One method, termed here the $LocalMoran'sI$ sampling, included higher selection probabilities at locations with high local spatial autocorrelation. Three other methods, $random$, $stratified$, and $regular$, all had equal selection probabilities over the sampling frame.

Leveraging information about high local spatial autocorrelation, the $LocalMoran'sI$ sampling method revealed a key characteristic associated with wide outliers. By increasing the probability of sampling cells of high spatial autocorrelation across a field, the $LocalMoran'sI$ method resulted in more samples on outliers and more outliers found. This suggests that the wide outliers are partially composed of regions of high spatial autocorrelation. Visual analysis revealed that the location of relatively high local spatial autocorrelation was in the central region around the peak value, the top part of the spatial outlier.

5 Multi-scale spatial dependence

5.1 Introduction

5.1.1 Previous work

The previous chapter provided evidence that high local spatial autocorrelation occurs in association with spatial outliers. Patches of high local spatial autocorrelation were evident in the central regions, or the tops, of known outliers. One factor affecting the degree of spatial autocorrelation was the variation of the outlier relative to that of the neighborhood. Another factor was the relationship between the scale of the outlier (its size) and the scale of analysis (the spatial resolution of the data). If a geographical feature is a local anomaly, at one scale it is an instance of local extrema, a spatial outlier entirely different than its neighborhood. At finer scale of analysis the feature exhibits patterns of various degrees of spatial dependence, which might indicate different parts of an outlier. While some parts of a spatial outlier have a high degree of variation, other parts have a high degree of spatial dependence. In this chapter, spatial outliers with a larger size than the scale of analysis are hypothesized to have three parts: a top, a side, and a base, which have distinctive properties compared to each other and the background field in which they are embedded.

5.1.2 Research question

The research question addresses whether distinctive scale-based characteristics of spatial dependence exist for parts of spatial outliers. For clarity, the investigation of "scale-based characteristics" does not mean a comparison of patterns of spatial dependence across individual spatial scales. Rather, it means an investigation of whether distinc-

tive multi-scale signatures exist for different parts of outliers. The multi-scale signatures compared take the form of a "distogram," a composite statistical summary, the general structure of which is presented in section 5.1.7 "Distogram description".

Extending research presented in the previous chapter, this experiment included two additional considerations in order to address the potential for more challenging conditions with regard to the availability of spatial data. One consideration was to work with a sparse random sample, as opposed to the support of a complete raster grid. The other consideration was to examine whether multi-scale signatures found in a variable of interest are also evident in correlated proxy variables. The reasoning for the proxy variables and the random sampling are presented next, followed by an introduction of the experiment, the spatial outlier parts, the multi-scale signature, and the various comparisons.

5.1.3 Proxy variables

There are situations in which information about the variable of interest is not directly available. However, there might be other correlated variables with patterns similar to those in the variable of interest. Since correlated spatial variables have corresponding spatial patterns, if an anomaly is evident in one variable, an anomaly might by spatially coincident in another variable. In cases where there is limited data of one variable then the data of another correlated variable potentially contains sufficient information to locate features such as those represented by the locally extreme values of spatial outliers. Patterns of spatial dependence indicating parts of outliers might also be identifiable and located by means of the correlated variable, depending upon the degree of correlation and the scale of analysis.

For this experiment, each variable field, represented by a simulated raster data set with spatial outliers embedded at known locations, was considered the original variable of interest. From an original data set, additional raster data sets were created to represent correlated ancillary variables. Gaussian conditional simulation was employed to generate raster data sets with attribute values spatially correlated to the original data sets. This process includes a means of controlling the degree of correlation. From each set of original data, several ancillary raster data sets were produced, each with a different degree of correlation to the original. The ancillary data sets represented variables correlated to the original variable across a specified sequence of target correlation values.

The process of Gaussian conditional simulation produces correlated data that are somewhat noisy. However, the Gaussian probabilities are applied uniformly across the region so the process induces generally regular patterns of correlation. This manner of simulation avoids irregular correlation and unequal correspondence across the data set. In other words, it avoids the potential for complexities of different variables that might result in different, uncorrelated patterns in some places. This enables a relatively equal comparison across numerous data sets with respect to the potential of spatially correlated ancillary data for finding outliers in a variable of interest.

5.1.4 Small sample

In the previous chapter, a complete raster was employed to represent a field. Supported by the entire raster, the local spatial autocorrelation was calculated at each and every cell, based upon similar queen's case neighborhoods of the eight adjacent cells (in all but edge or corner cases, which had six or three neighbors, respectively). For this experiment, in order to investigate situations where only sparse observations exist, a small random

sample of the field was employed for each test. This was to challenge the approach. The sparse sample represents a circumstance of having limited spatial information. The random sampling produces points that have a variety of spatial neighborhood configurations. Also, the results in the previous chapter show that random sampling more often missed spatial outliers compared to regular sampling. A small random sample is, therefore, a relatively challenging spatial configuration for finding characteristics of spatial outliers or their parts. If the parts are found to have distinctive characteristics with a small random sample then the approach likely works better with samples having more exhaustive or even coverage, such as a dense raster grid.

5.1.5 Experimental summary

The small random samples of the original and proxy data were used in an experiment concerning patterns of spatial dependence within spatial outliers, summarized as follows. At each of the sampled points a multi-scale signature of spatial dependence was computed as a distogram. The points were then assigned to classes based upon their locations relative to defined spatial outlier parts. Grouped by spatial outlier part class, the distograms of the points were generalized to produce one distogram model for each outlier part class. To mitigate the effects of chance and the potential of deriving an extreme case for any model, multiple iterations of the previous steps were performed. The final distogram models for each outlier part class were obtained by calculating the mean and confidence intervals from the multiple iterations. The final models were compared to ascertain whether distinctive patterns of spatial dependence exist for each spatial outlier part class. The comparison also included consideration of the various parameters in the experiment (i.e. parameter values for the field variogram, the outlier form, and the proxy variable target correlation).

5.1.6 Spatial outlier parts and patterns of spatial dependence

In the previous chapter, spatial outliers were examined with respect to the presence of local spatial autocorrelation. A spatial outlier is commonly considered a single extreme value when the scale of analysis matches the spatial extent of the outlier (Shekhar et al., 2003). At a finer scale of analysis, a wide spatial outlier, which spans across the space represented by multiple data, might be composed of sub-regions, or parts, that have various degrees of spatial autocorrelation. Based upon the research presented in the previous chapters, the central region, immediately surrounding the extreme value, is expected to have high local spatial autocorrelation representing the spatial dependence nearby the extreme attribute value. Just beyond that central region, a band of high variance is expected, representing a major transition from the extreme values to the background. Finally, another band of minor transition is expected, that blends the properties of the high variance region with the background field. Although various general terms for the point of extrema and the regions of transition might be considered, with positive-valued attribute values and perturbances, it is conceptually clear to call the three regions the top, the side, and the base of an outlier. Across the top, side, and base parts and into the background field, the ordered pattern of spatial dependence has the potential for identifying wide spatial outliers, with sizes larger than the spatial resolution of the data.

As in the previous chapters, this experiment employs spatial outliers simulated, as point anomalies, with the Gaussian function embedded in a variable background field. The spatial outlier form includes an extreme point and a transition to the background. The defined Gaussian shape of the outliers also enables a straightforward means of partitioning outliers. Outward from the center of an outlier, spatial boundaries at each of

the first, second, and third standard deviations defined its top, side, and base parts, respectively, as shown in Figure 27.

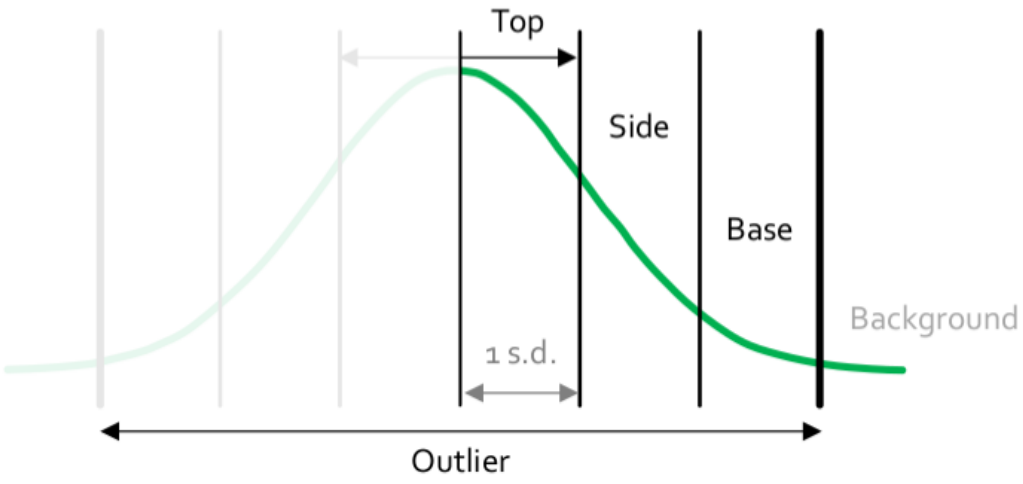


Figure 27: Classes of outlier parts and background, specified by spatial boundaries at the first, second, and third standard deviations.

The pattern of interest in this experiment essentially consists of local spatial autocorrelation that varies across different regions. Since calculating local spatial autocorrelation requires groups of data across space, the process of modeling such a pattern first requires a binning of data into groups ordered by distance. Then, the groups are summarized with a statistic representing the degree of local spatial autocorrelation. Composing the group values into a multi-scale signature of spatial dependence produces a special type of a general representational form: a distogram.

5.1.7 Distogram description

The concept of a distogram is based upon the common histogram, which summarizes groups of binned data. A histogram is a general, two-axes graph that represents groups

of binned data for comparison to one another. The method of partitioning the data into binned groups and the statistical operator for summarizing values of the binned groups are both arbitrary. There is no requirement that histograms present the group summary values in any order. However, it is sometimes informative to order the group values by their magnitude. Another option is to order the group values with respect to the scale of the variable used for binning. For example, data grouped into bins by an interval scale can be ordered by the sequence of interval midpoints or interval boundary values.

A distogram groups data into bins ordered by distance from a point of origin. For example, consider that a set of spatially-referenced data are points. From any point, an origin, the distance is measured to each of the other points. Distogram bins are defined by inner and outer distance boundary values without overlaps. Preferably there are no gaps between bins and they are defined to any appropriate limit, such as the largest point-to-point distance. Based upon its distance from the origin, each point is assigned to a unique bin. Finally, an operator is applied to each group of points and the resultant summary statistic is represented on a two-dimensional graph, with bins along one axis and the magnitude of the statistic along another.

Although the description above is of a distogram for a single point of origin, many identically-defined distograms can be combined, referenced together by their points of origin, and their statistics summarized again. For example, distograms of each and every point in a data set can be combined by applying an operator to average their summary statistics per bin.

The common term "bandwidth distance" typically refers to a boundary distance to constrain a region of analysis for an operation. Therefore, from here forward, the in-

ner and outer distance values of distogram bins will be called the inner-bandwidth and outer-bandwidth, respectively. Correspondingly, the bins will be called bands, nominally referenced by the outer-bandwidth value or by their index in the ordered sequence of bands outward from the origin. Assuming 2D Euclidean space, each band is an annulus, a region between two concentric circles, centered on the point of origin. The first band around the origin covers a circular area if its inner-bandwidth is zero.

The general distogram structure is a method of statistical summary and such a representation is familiar in the domains of spatial analysis and geostatistics. Two common structures can be considered special versions of distograms: the (semi-)variogram and the autocorrelogram. The variogram has y-axis values of the sum of squared difference between observations in the corresponding distance bin, with each bin normalized by dividing by the number of observations in that bin, and for the semi-variogram, by definition, divided by two (Matheron, 1963). The autocorrelogram depicts y-axis values of the autocorrelation coefficient of groups of points across a series of lag distances (Venables & Ripley, 2002). Although various versions of autocorrelation functions are used in an autocorrelogram, the autocorrelation values calculated for each distance along the scale also include the point of origin, at a lag distance of zero. In this experiment, the distogram was employed for the multi-scale signature of spatial dependence specified by a measure of local spatial autocorrelation, computed only from points within each distance bin. This is different than the variance at different distances, as found in the variogram. Also, since the Local Moran's I is computed solely by points in each bin, not including the point of origin, this is also different than the autocorrelogram. As a distogram is useful for representing any of a variety of statistics summarized per distance, it is helpful to refer to a distogram with some description of the specific statistical summary that it presents, such as a "distogram of averages", a "variance distogram" (or the com-

mon "variogram"), or a "Local Moran's I distogram," which was used in this experiment.

5.1.8 Distogram of Local Moran's I

For this experiment, a small random sample was taken from each raster data set yielding sets of points with attribute values. For each point, a multi-scale signature of spatial dependence was represented as a distogram. Specifically, for each point, the characteristic of local spatial autocorrelation was calculated as the Local Moran's I statistic of all neighboring points in each band of the distogram. Combining individual distograms of points belonging to each outlier part class yielded a generalized distogram model for each respective part class.

A choice was made for the definition of distogram bands with regard to the width of outliers. One alternative, not selected, was to hold the bandwidth distances constant across all test data sets, regardless of the outlier sizes. However, including a number of bands with sufficiently fine granularity to represent small outliers and also extend out to a range that covered the largest outlier and its surrounding neighborhood was infeasible for the available computational resources. The selected alternative was to define bandwidths for each simulation relative to the scale of the outlier, as specified by its Gaussian width parameter value. This approach enabled a comparison with respect to the various outlier sizes and the scale at which any distinction might be lost in the same sets of background fields.

The bandwidth distances were selected to potentially capture patterns of local spatial autocorrelation for the outlier parts and the background field. Of direct importance to the research question is a level of detail that might represent each of the top, side, and

base outlier parts. As Nyquist (1928) and Shannon (1949) describe, the minimal sampling distance required to represent a signal is one-half the magnitude of its wavelength. For a spatial feature of interest, that requires two points per feature from an origin. For example, consider the minimum sampling required to capture the (profile) form of a (symmetrical) mountain as a single point raised above the surrounding plains. Starting from one point of origin at the base on one side, two more points are needed, one at the peak and one at the opposite base. Following this condition of minimal sampling, two bandwidths were specified for each of the three outlier parts to capture their potentially distinctive patterns. The first six bandwidths, covering the extents of the three outlier parts, were equally spaced with distances of one-half of the outlier width parameter (i.e. one-half of the Gaussian standard deviation).

To potentially represent neighboring outliers and the background field beyond the extents of a known outlier and the first six bands, three more bands were included, for a total of nine bands in the distogram, as illustrated in Figure 28. The width of each of these three outer bands was equal to the diameter of an outlier (i.e. six times the Gaussian standard deviation, since the outlier radius was defined by the first three Gaussian standard deviations). If the point of analysis was at the center of an outlier, the seventh band would corresponded to the imposed gap between simulated outliers, the eighth would potentially capture a neighboring outlier, and the ninth would correspond to the mixed background field beyond. It is important to note that the distogram does not represent characteristics for outliers only (i.e. it is not to describe the space relative to the center of an outlier). Rather, the distogram is to represent characteristics of the space surrounding each datum, at any location in the field, whether coincident with a known outlier or not.

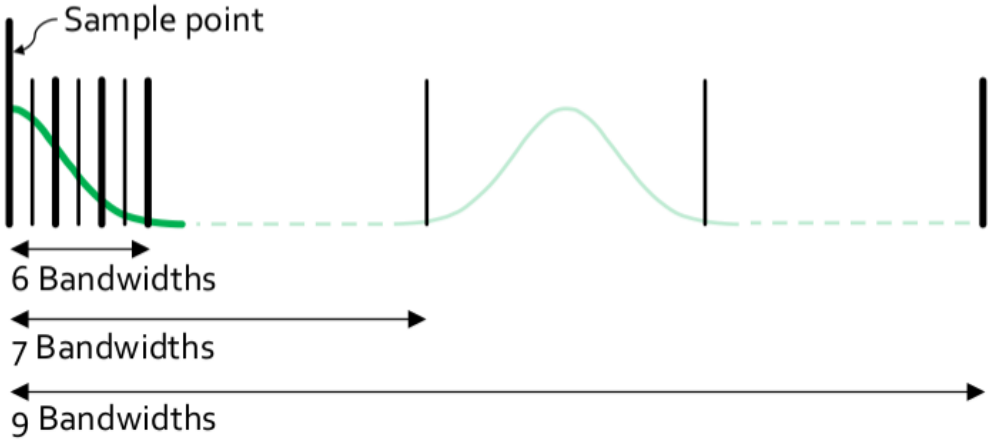


Figure 28: Distogram of Local Moran's I, showing the bands corresponding with an outlier, the imposed gap, a potential neighboring outlier and a mixed background beyond, if the origin point of analysis is at the center of an outlier.

The distogram for each point represents a multi-scale signature of spatial dependence, for which the measure of local spatial autocorrelation employed was the Local Moran's I statistic (Anselin, 1995). For each point of origin, values for each band of its distogram were assigned the Local Moran's I value computed from all neighboring points located within the distance range of that band. A key characteristic of this approach is that the multi-scale signature of spatial dependence represents local spatial autocorrelation across space with the non-overlapping bands of the distogram. The distogram values represent a sequence of disjoint adjacent regions, outward from the origin point of analysis.

Abundant research exists that evaluates local spatial autocorrelation across multiple bandwidths (C. Zhang & McGrath, 2004; C. Zhang, Luo, Xu, & Ledwith, 2008; Gerçek et al., 2011; Bone, Wulder, White, Robertson, & Nelson, 2013; Kedron, 2016; Yuan, Cave, & Zhang, 2018). However, the majority of such studies compare the separate results of varying only the outer-bandwidth, leaving the inner-bandwidth at zero (e.g. to define a circular area in Euclidean space). Setting the inner-bandwidth at zero includes

information of all points found from the origin to the outer-bandwidth. Varying just the outer-bandwidth tends to introduce a type of conflation with regard to the results of larger bandwidths, which include all of the information already included at smaller bandwidths. A comparison of results then involves some duplication of information. Such approaches are likely appropriate for the purpose of calibrating the outer bandwidth to find the scale at which patches of similar values occur. However, this type of conflation is problematic for the concerns of this experiment.

The aim of this experiment is to investigate patterns of local spatial autocorrelation occurring in regions at various distances from each point of analysis. Sequenced patterns of high and low local spatial autocorrelation would offset each other if the measured values from the regions are combined together. It is preferable to obtain the signals separately, in a partitioned manner, to identify a multi-scale signature of spatial dependence across several bands. By avoiding the generalization of mixing offsetting signals, the distogram models of each part might maintain distinctive signals compared to models of other parts.

5.1.9 Comparison of distograms for each part class

For each of the sparse points, a Local Moran's I distogram was computed. To compare distograms for each of the spatial outlier part classes, it is first necessary to identify which outlier part, if any, each distogram represents. As in the previous chapters, the simulations included Gaussian outliers centered at known locations and with known height and width parameters. Regions representing the outlier parts were identified based the center of each known outlier and its top, side, and base outlier part boundaries (at the first, second, and third standard deviations from the center, respectively). Each point, and its distogram, were labeled as belonging to an outlier part class or the background

class, depending upon the labeled region in which it was located.

To illustrate the labeling of points and distograms as belonging to a unique class of an outlier part or the background, three figures with schematic representations follows. In each of the figures the simulated source raster is underlain for visual reference. The shades of gray in the raster indicate the attribute value, with higher values in darker tones. There are eight relatively dark patches representing the embedded outliers.

Figure 29 shows a set of points, as yellow circles, representing an example of the small random sample locations, distributed across the field. These points represent the locations of original data from which Local Moran's I distograms are computed. Figure 30 shows an overlay of translucent red, green, and cyan colored bands representing the top, side, and base outlier part classes, respectively. The color of the sample points indicate their membership in either of the three outlier part classes or as a background point, in purple. Figure 31 shows the labeled points with all but one of the outlier's bands removed.

Also shown in Figure 31 is an example illustration of a profile of a topographic feature, a mountain, as an instance of a spatial outlier that has anomalous altitude characteristics compared to its surroundings. Hypothetical bands representing the top, side, and base regions of the spatial outlier are depicted graphically by colors along the bottom. The outlier part boundaries are indicated by the vertical black lines. Examples of sample points are depicted on the surface of the mountain, colored corresponding to their membership in one of the outlier part classes.

For each simulation, all individual point distograms were generalized into class-based distogram models. Specifically, for each class, the Local Moran's I values in each band

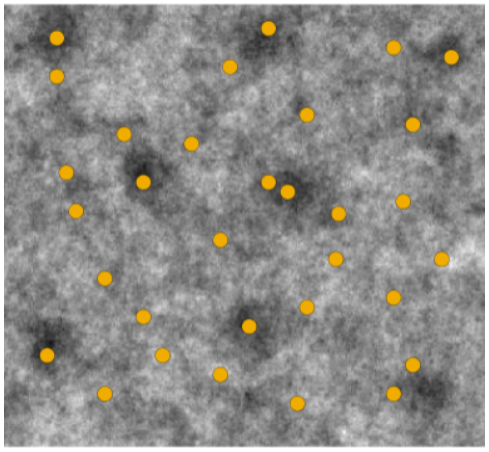


Figure 29: Illustration of random point locations (yellow) sampled across simulated field.

of the individual point distograms were averaged. Combining the patterns from multiple individual point distograms yielded a relatively stable estimation of a Local Moran's I distogram model compared to the distogram of a single point, which might represent an unusual circumstance specific to a single location.

At a higher level, the chance that any simulated raster was an unusual case was also addressed. Each unique set of outlier and field parameters produced a simulated raster representing one original variable, from which several proxy variables were also produced. In order to avoid the chance that the original variable was an unusual realization, multiple iterations of the simulations were performed and subsequently combined. From the multiple iterations, the mean and the upper and lower 95% confidence interval boundaries were computed for each distogram band of each outlier part class. The resulting summary distograms were considered the final distogram models used for comparison.

The final models were compared to address the research question of whether distinctive scale-based characteristics of spatial dependence exist for parts of spatial outliers.

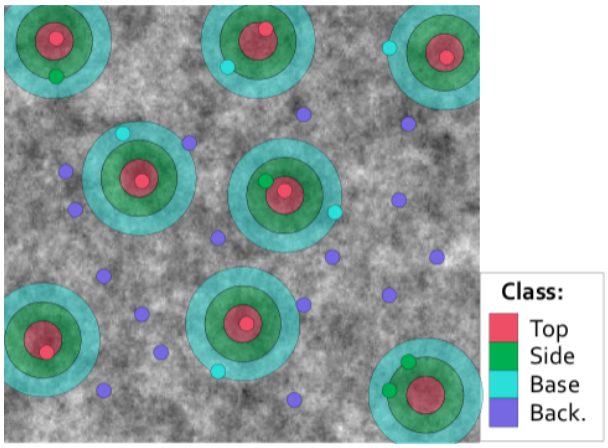


Figure 30: Illustration of the point samples classified into one of four classes of outlier parts (Top, Side, Base) or the background (Back.), based upon known outlier locations.

The approach specifies the general characteristics in question as a multi-scale signature of local spatial autocorrelation represented by a Local Moran's I distogram. The main comparison, ultimately based on only sparse point values of a small random sample, was whether any distinction was evident between the final distogram models of the four classes (i.e. the top, side, and base spatial outlier parts or the background). As the problem of identifying spatial outliers is dependent upon both scale and the patterns of variability of the anomalous features and the background, a core consideration was to evaluate the effects of the outlier height and width parameters in relation to the field nugget, partial sill, and range parameters on the ability to discern between part classes. Another consideration was the potential of the proxy variables, across the sequence of target correlations, for maintaining any distinction between outlier part classes.

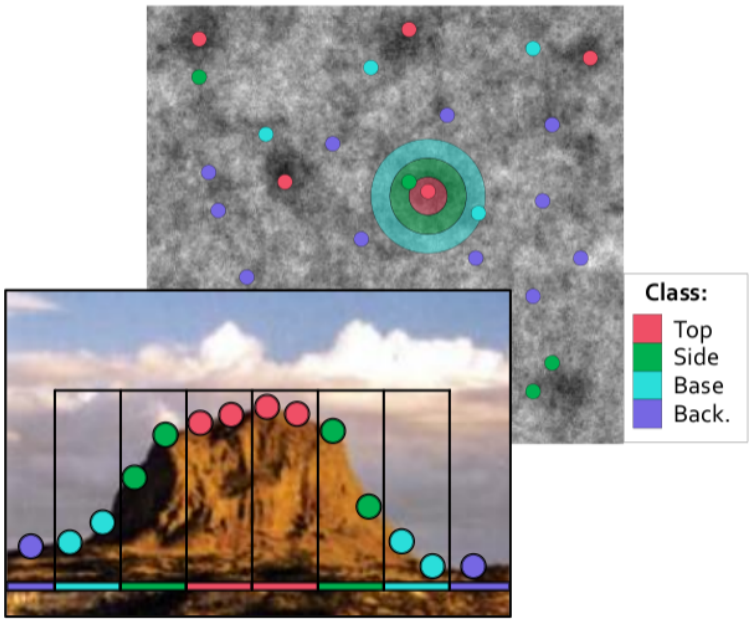


Figure 31: Illustration of classified points, including an example profile view of an anomalous topographic feature, a mountain, with points representing altitude values classified as one of the outlier parts or the background.

5.2 Methods

5.2.1 Overview

The methods for this experiment fall into three general stages: data generation, statistical modeling, and creation of visual representations for comparison. The data generation stage included consideration of the data structure, the generation of raster grids of original data and correlated data, and point samples representing a sparse data set for analysis. Various original data sets were created, involving 16 sets of different outlier shapes and 30 realizations each of 64 variable field parameter sets. From each original grid, of outliers mixed into a field, correlated data sets were generated, for a total of 5 related sets including the original. Small 1% point samples of the original and correlated data were the final data sets used for the statistical modeling. The experimental approach involved the statistical summary of patterns of spatial dependence around each point. To quantifiably

characterize this property, the Local Moran's I distogram values were computed for each point and aggregated per class of outlier parts or the background. To obtain more stable generalizations, the results of the 30 realizations of each variable field were aggregated. From the summarized Local Moran's I distogram values, graphs including the mean and 95% confidence intervals of all four classes were created to inform the analysis. Computational aspects of hardware, storage, and processing time are also included.

5.2.2 Computational system

Various computational hardware were employed for this experiment. A personal computer with the Windows 10 operating system was host to a Hyper-V virtual machine with the Linux Ubuntu operating system (version 14.04), upon which the majority of processes were performed. The virtual machine had access to a solid state storage drive and eight to twelve gigabytes (GB) of dynamic random access memory (RAM). Along with *bash* shell functionality it also included the R statistical software (version 3.4.3) with additional packages, such as *gstat*, *spdep*, *sp*, the *tidyverse* and its related *googledrive* interface.

A portion of the processing-intensive stages were performed on a remote cluster of twenty virtual machines, provided by the Center for Spatio-temporal Thinking and Computing and Applications. The machines each had the Linux Ubuntu operating system (version 16.04) and the same R software and packages as the personal computer listed above. Each virtual machine had access to 16 GB of RAM and 20 GB of storage. Due to the storage limitations, nearly 4 terabytes (TB) of source and result data were processed piecewise and transferred in both directions to cloud storage in Google Drive, provisioned by the University of California, Santa Barbara. Remote process management,

over *SSH* (Secure Shell) network connections, was enabled by both *bash* scripting and the *GNUScreen* software.

5.2.3 Raster grids

Each simulated data set represented a field as a 500 × 500 two-dimensional raster grid structure. Computational limitations necessitated that the data sets were small enough for efficient processing. However, to statistically characterize the outlier parts, a sampling frame that provides sufficient information was required. The simulated rasters were designed to provide at least a minimum amount of information required but in most cases the statistical support exceeded that. The spatial extents and the scale of analysis (i.e. resolution) required of a raster data set depended upon the scale of the feature (i.e. outlier, or outlier part) of interest. The grid characteristics, the outlier width, and the outlier placements were considered together as subject to four factors.

The first factor related to the scale of analysis. The minimum number of raster cells required for the smallest outlier part, the top, was estimated based upon a small sample size. To potentially sample 1 cell from the region of an outlier top, the one percent random sample employed for obtaining sparse points required a top part composed of at least 100 cells. However, although the Local Moran's I distogram was created from neighboring cells of all classes, it was considered preferable to also sample at least one neighbor in the top region to better estimate statistical characteristics of the top part. Therefore, a sample of at least two cells was the minimum needed for the top part, each is both a point of analysis and a neighbor for the other. A one percent sample expected to obtain two points required a minimum of 200 cells representing the top part. To derive the minimum outlier width, the radius of the top part can be estimated with the formula

of the area of a circle, $A = \pi r^2$, rearranged as $r = \sqrt{A/\pi}$, where A is the area and r is the radius. Assuming one unit of distance per cell, the radius of a top part that covers 200 square units (cells) is approximately 8 units. Therefore, the minimum radius required for a one percent sample to potentially obtain two point values from the region of the outlier top was considered 8 units (cells).

The remainder of the four factors affected the spatial extents of the rasters. The second, third, and fourth factors related to the preferred placement, density, and numbers of outliers in a field, respectively. The second factor, an imposed gap between outliers, restricted the otherwise random placement of outliers. To avoid overlaps of outliers and obtain consistent contrast between an outlier and the background field, the imposed gap was equal to the diameter of an outlier. The third factor reduced the proportion of area covered by outliers. To impose that only a minority of the area was covered by outliers, particularly for efficiency of locating outliers randomly in the available space, the proportion of area covered by outliers did not exceed 40% of the total area. The fourth factor was the selection of a minimum number of outliers created in any raster. Addressing this factor required balancing the need for data sets that are both large enough to characterize several outliers embedded in the same field but also small enough for efficient processing. The need for efficiency imposed an upper limit for the spatial extents, but a lower limit related to the width of large outliers. An arbitrary number of four outliers was selected. This was considered the minimum number of outliers that potentially provides sufficient information to characterize the general patterns of outlier parts from multiple instances in the same field.

To accommodate the four factors above, each of the two-dimensional raster data sets were created with the spatial extents of both axes ranging from zero to 500 units. The

resolution was one unit per cell, creating grids of 500 × 500 cells. This pair of spatial extents and resolution accommodated a minimal sampling of the smallest outlier part, a contrast between outliers and background, a computationally efficient placement strategy, and an arbitrary minimum number of outliers for the generalization of outlier part characteristics.

5.2.4 Outlier grids

The process of simulating outliers generally followed the methods of the previous chapter. Only particular parameter values were modified for the grid, the placements, and the outlier shapes. The new grid size was 500 × 500 cells, again with each cell again representing one square unit. The outliers were again placed randomly in the grid and separated by a gap of one outlier diameter. However, the new proportion of area covered by outliers was specified not to exceed 40% of the total area. Each permutation of the two Gaussian shape parameters having four values each were used to generate 16 outlier grids. The new (standard deviation) width values were in the set {10, 12, 14, 16} and the new height values were in the set {10, 20, 30, 40}. The 16 outlier grids are shown in Figure 32.

In addition to outlier grids containing (height) attribute values, 16 classified grids were also generated, each associated with an outlier grid. Based upon the locations and the width parameter of outliers in each grid, each cell in the associated classified grid was assigned to one of four classes. Cells within the first, second, or third standard deviation from an outlier center were assigned to the top, side, or base outlier part class, respectively. All other cells were assigned to the background class.

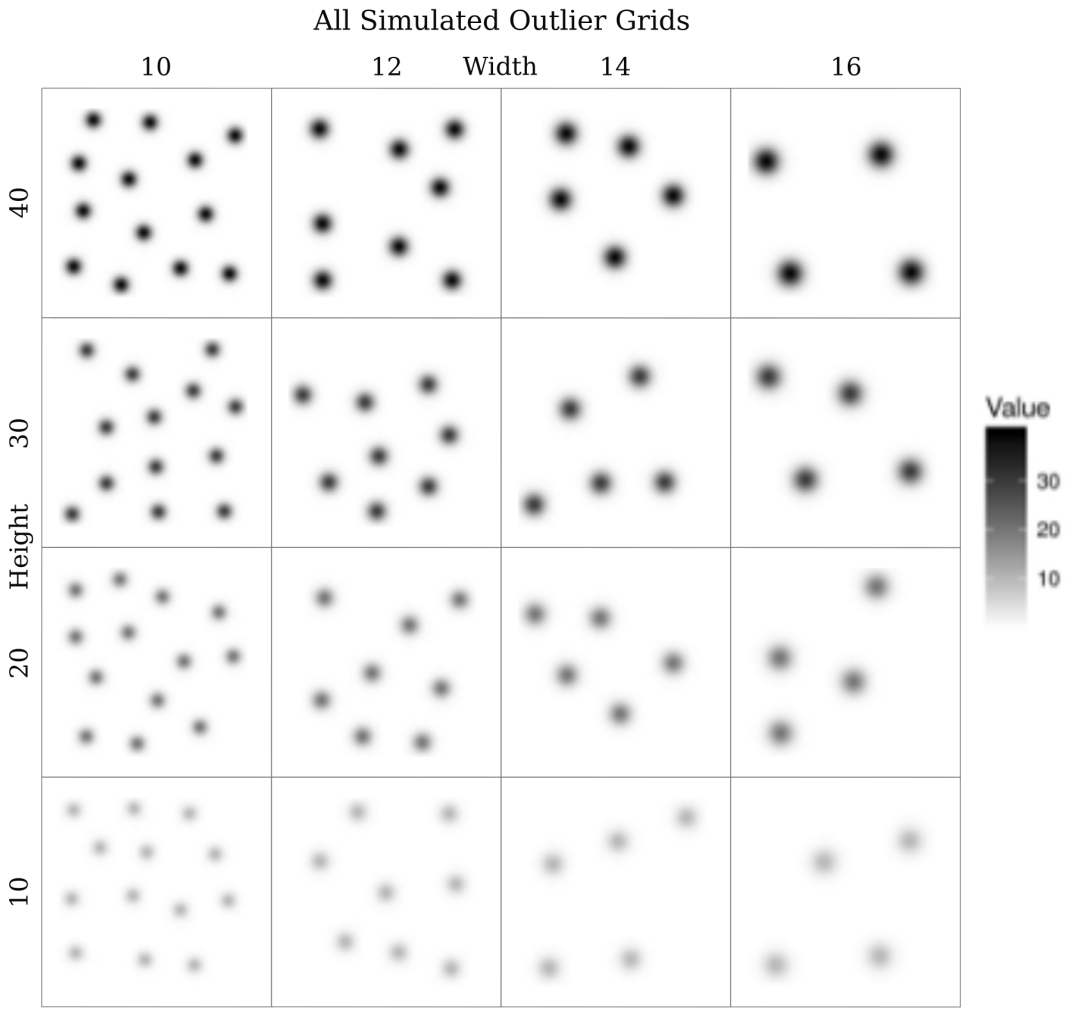


Figure 32: The sixteen sets of Gaussian outliers, which represent all permutations of four values each for the Gaussian height and width parameters.

5.2.5 Field grids

The process of generating fields by unconditional simulation generally followed the methods of the previous chapter. Only the grid and variogram model parameter values were modified. The grid size was 500 × 500 cells, with each cell representing one square unit. Each permutation of the new parameter values for the nugget {3, 6, 9, 12}, the partial sill {10, 20, 30, 40}, and the range {10, 20, 30, 40} specified the 64 variogram models.

The R package *spdep* was employed to specify a variogram model with the function *vgm*. The model form was of the type *Exp* (exponential). Simulations of fields were produced with the R package *gstat* and the function of the same name. With each of the 64 variogram models, an *nmax* (maximum number of neighbors) of eight, and a *beta* (height constant) of zero specified 64 gstat objects. Each of the objects were submitted to the *predict* function to create field simulations. Instead of just one realization, though, 30 iterations were performed. From each of the 64 permutations of field parameters, 30 realizations were generated, for a total of 1,920 field grids. Examples of the field grids, created with the two (highest and lowest) boundary cases of each of the three variogram parameters, are shown in Figure 33.

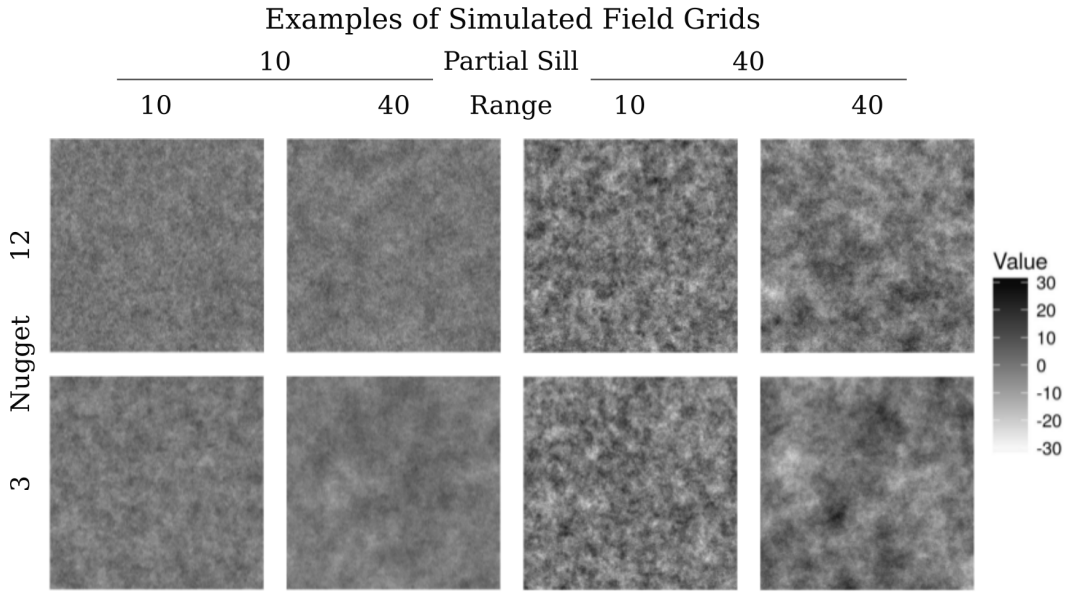


Figure 33: Examples of simulated fields created with unconditional Gaussian simulation, using the boundary cases of each of the variogram parameter values for the nugget, partial sill, and range.

5.2.6 Mixture grids

Departing from the methods of the previous chapter, the mixture of outliers into the field grids was controlled differently. Instead of unique outlier grids created for each mixture grid, the same 16 outlier grids were combined, by addition, into the various field grids. This was to enable a comparison across the various fields while holding constant the locations and shapes of each set of outliers. There were 64 permutations of the three field (variogram) parameters with four values each. Considering just one field realization for each permutation, if each of 16 outlier sets were embedded into each of the 64 fields, 1,024 mixed grids would result. This number represents the number of mixed (outlier and) field parameter sets, 1,024. Combining each of the 16 outlier sets into all 30 realizations of each of the 64 permutations of field parameters, generated 30,720 mixture grids. Each of the mixture grids are considered an original variable of interest. Figure 34 shows examples of mixture grids for the parameter value boundary cases of the smallest and largest combined parameter values for outliers (height and width) and fields (partial sill, range, and nugget).

5.2.7 Correlated proxy grids

From each of the 30,720 mixture grids of original variables, four correlated grids were generated with a process of conditional Gaussian simulation. The four target correlation values employed to generate the correlated grids were in the set $\{0.9, 0.8, 0.7, 0.6\}$. The resulting sets of one original and four correlated grids brought the total number of rasters to 153,600 grids, five times the number of original grids.

The process of creating a correlated grid involved three main steps. First, the original

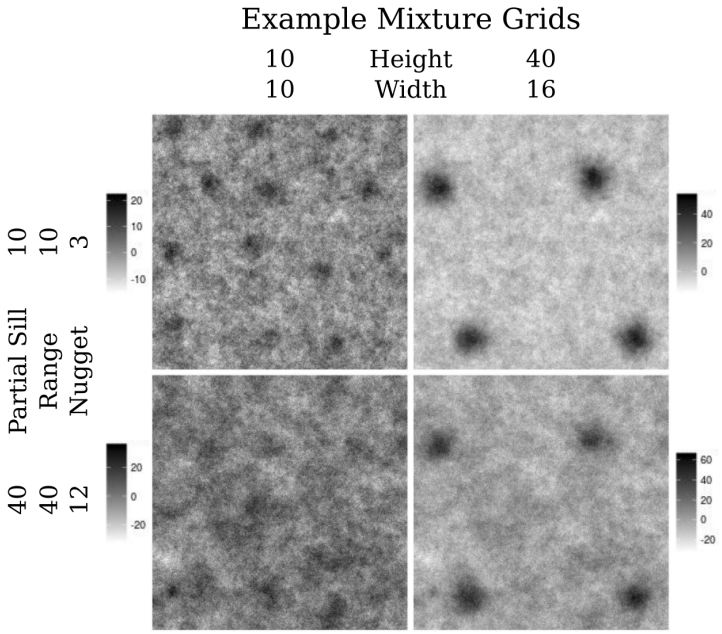


Figure 34: Examples of mixture grids of outliers in variables fields for the outlier and field combined parameter value boundary cases.

grid attribute values were adjusted with a "normal scores transformation". Second, a process of conditional Gaussian simulation created correlated values as normal scores. Third, the correlated normal scores were back-transformed to attribute values for the correlated grid. Several R functions from the *base* and *stats* packages were employed for manipulating the data and performing mathematical or statistical operations.

For the normal score transformation, the 500 × 500 grid of original attribute values was vectorized to a 250,000 × 1 matrix and the *qqnorm* function sorted the attribute values of the cells, ranked the values, and returned, for each original value, a new value assigned from the theoretical Normal distribution value of the same rank. Also, the original attribute values and normal scores were sorted and combined into an ordered list to serve as an index for the back-transformation step.

The Gaussian conditional simulation involved a 2×2 target correlation matrix with values of one unit in the diagonal cells and a specified target correlation value in the off-diagonal cells. A random sample of the Normal distribution was taken, with the *rnorm* function, to obtain uncorrelated normal scores in a 250,000 $\times$ 1 matrix, the same size as the original normal scores. Both the original normal scores and the uncorrelated normal scores were combined as columns into a 250,000 $\times$ 2 matrix. The standard deviations of both columns were calculated and the values were assigned to the diagonal of an otherwise zeroed 2×2 standard deviation matrix. A covariance matrix was determined by matrix multiplication of the standard deviation matrix, the correlation matrix, and the standard deviation matrix, again. A Cholesky factorization of the covariance matrix, performed with the *chol* function of the R *base* package, determined an upper-right triangular matrix. Matrix multiplication of the 250,000 $\times$ 2 matrix of (original and uncorrelated) normal scores and the upper-right triangular factor resulted in a 250,000 $\times$ 2 matrix, from which the second column of correlated normal scores was extracted.

The back-transformation step involved a linear interpolation, with the $approxfun$ function of the R *stats* package, between the original normal scores and the original attribute values to solve for a linear function. The function was applied to the correlated normal scores to return the correlated attribute values in a 250,000 $\times$ 1 matrix, which was reorganized back into a 500 $\times$ 500 grid. The measured R^2 correlation values between each original grid and each of its correlated grids do not exactly match the target correlation values, but they are reasonably close estimations for comparing across the range of target correlation values. Examples of an original grid of attribute values and the correlated grids are shown in Figure 35.

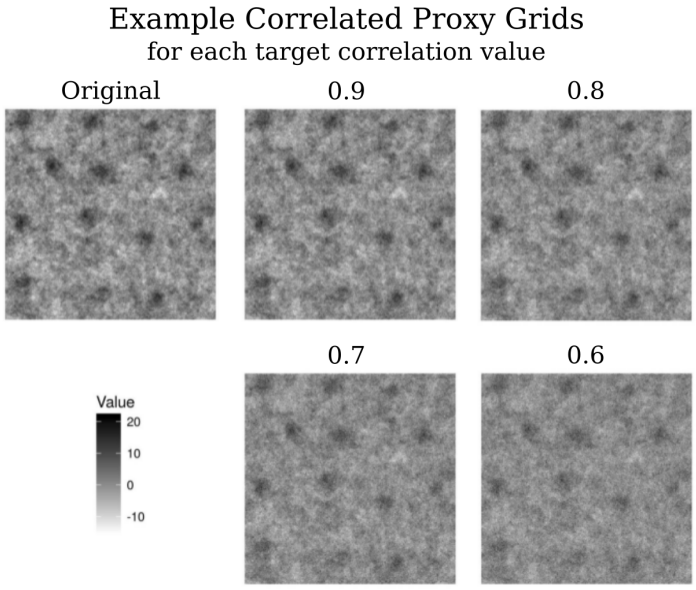


Figure 35: Examples of correlated proxy grids with target correlation values of 0.9, 0.8, 0.7, and 0.6, with the original mixture grid.

5.2.8 Sampled points

To address an aim of this experiment that the approach is applicable to limited and sparse data, a small random sample was taken from each of the raster grids. The R *base* package *sample* function was employed to sample 1% of the grid cells as spatially referenced 2D point locations. The attribute value and outlier class membership was taken from each sampled cell and assigned to the corresponding point.

There was a noteworthy adjustment to the sampling process. If a random sample was taken across the entire area of the grid, there is a chance of non-uniform sampling. In some regions samples might cluster and in other regions samples might be diffuse. This possibility can lead to an under-sampling of some outlier part classes, especially the smaller parts, such as the tops of small outliers. To mitigate this problem and to ensure a more uniform density of samples across the parts, the 1% sample was taken per class.

Sampling the classes based upon the number of cells for each class has the potential not to sum to a perfect 1% of the entire region. Therefore, the (smaller) number of samples for the outlier parts were rounded up and the remaining (larger) number of samples for the background were adjusted down to achieve a 1% sample, overall.

Although much of the analysis and results refer to a single set of points taken from each grid, there were actually two sets of samples taken. Reasons for the two sets were to provide redundancy and additional confidence in the results and conclusions. Both sets were subject to the same methods throughout the experiment, but the second set was only used for reference in case concerns arose with respect to any questionable result.

5.2.9 Local Moran's I distograms

For each sampled point location a Local Moran's I distogram was computed. Each distogram included a Local Moran's I value for each of nine adjacent, non-overlapping, distance-based bands. The point of origin for each distogram was the sample point location and the widths of the bands were relative to the size of the simulated outliers in each data set. The width of each of the inner six bands was one-half of the outlier (standard deviation) width parameter, corresponding to one-half of an outlier part. The width of each of the outer three bands was the width of an outlier diameter, six times the outlier (standard deviation) width parameter. Each band was defined by an inner-bandwidth and an outer-bandwidth that was used in the Local Moran's I calculation.

For example, consider the data sets representing the smallest outliers, with an outlier (standard deviation) width parameter of 10. The six inner bands are each 5 units wide and the three outer bands are each 60 units wide. The inner-bandwidth values

increment across the set {0, 5, 10, 15, 20, 25, 30, 90, 150}. The outer-bandwidth values increment across the set {5, 10, 15, 20, 25, 30, 90, 150, 210}. The inner-bandwidth and outer-bandwidth distance boundaries of the first band are the pair {0, 5}, and those of the ninth band are {150, 210}.

The Local Moran's I value for each band was computed from the attribute values of all neighboring points falling within that band, and excluded the point of origin. From the R package *spdep*, the functions *dnearneigh*, *nb2listw*, and *localmoran* were employed. For each point of each data set, the following process was performed to compute the Local Moran's I value for each band. With specification of the inner-bandwidth and the outer-bandwidth, the *dnearneigh* function identified neighboring points, with the neighbors of all points in a data set returned as a *nb* object. The *nb2listw* function converted the *nb* object to a *listw* object of the points' neighbors with assigned weights for each neighbor, specified as the B style of weights for binary inclusion of all the neighbors for each point. Finally, the *localmoran* function was employed to calculate, from the *listw* object, the Local Moran's I value for each point, based upon its neighbors' attribute values and their weights.

For each data set, the distograms of each point were aggregated by their class membership (i.e. top, side, base, or background). Grouped by class, the point distograms were summarized on a per-band basis. The average distogram for each class was calculated with the *t.test* function of the R *stats* package. Unless specified as an individual sample point distogram, from here forward the term "Local Moran's I distogram" refers to the averaged distogram model for each class and averaged, again, over the 30 simulation iterations for each parameter set. For comparison purposes, all four classes were represented on resulting distogram graphs. For each data set, a trio of distogram graphs

were created, one for each of the inner nine, seven, and six bands. This was to depict a range of results from the complete set to a focused view of details nearby the point of analysis.

In total there were 153,600 simulations (16 outlier parameter sets × 64 field parameter sets × 30 field realizations × 5 sets of original or proxy variables). However, since there were two sets of 1% samples from each simulation there was a total of 307,200 data sets. Considering the trio of distogram graphs, there were 921,600 graphs created, each with representations of the four class averages. If the four classes were represented on separate graphs, the count would have been 3,686,400.

The computation of Local Moran's I distogram values and graphs for each data set was performed on the cluster of 20 virtual machines at the Center for Spatio-temporal Thinking and Computing and Applications. Because previous processes for outlier and field simulations used all of the system RAM available, 16GB of RAM was allocated to each virtual machine. However the functions of the *spdep* and *gstat* packages were intermittently observed to utilize approximately 4 GB. Based upon preliminary testing, the processing was estimated to take about 3.2 days. However, the nearly 4 TB of data required bi-directional transfers to storage on the remote cloud of Google Drive. Network performance varied for the file transfers and, with all 20 virtual machines running simultaneously, the Local Moran's I distogram process duration was approximately 10 days. Running on just one of 20 machines, the duration might have been 200 days.

The next step was, to mitigate the potential for an unusual field simulation. The results of 30 realizations of each field parameter set were aggregated. Across each distogram band, the average and the 95% confidence interval were calculated with the *t.test*

function of the R *stats* package.

Considering just the first 1% sample, by aggregating the 30 realizations of field parameters, the number of 153,600 sets was reduced to 5,120 (16 outlier parameter sets × 64 field parameter sets × 5 sets of original or proxy variables). Accounting for the trio of distogram graphs of nine, seven, and six bands, there were 15,360 graphs created for the analysis, each with representations of the four class averages. Doubling that count to include the second 1% sample, for redundancy, the total of distogram graphs summarized across the field realizations was 30,720.

5.3 Results and Discussion

5.3.1 Overview

This experiment produced abundant results. The selection of outlier and field parameter values was intended to cover a range of scenarios. The outlier grids included features that ranged in shape from low to high and narrow to wide. The fields ranged from low variability to high variability, with respect to noise, total variance, and the general distance at which the variance had a major effect from each location. At one extreme, the parameter values produced outlier features that were clearly anomalous and of different character to their surroundings. At the other extreme, the values produced features that were visually imperceptible from the field in which they were embedded.

It can be argued that imperceptible features are not spatial outliers, that they do not exhibit extremely different characteristics than their neighborhood. It is common to label locations as spatial outliers only if they have locally anomalous qualitative or

quantitative characteristics. However, for this analysis, the term spatial outlier refers to any of the Gaussian-shaped features in the outlier grids. The outlier grids were intended to introduce locally anomalous features in variable fields. Added to the field grids, the outliers are defined features at known locations that perturb the field values, yielding the mixed grids. The perturbance effects range from minor to major, depending upon both the outlier and the field parameter values. To ascertain generalizable patterns, outliers and fields were generated across a range of variability to emulate scenarios that include clear anomalies as well as situations in which anomalies are difficult to identify.

Other than the extreme boundary parameter values, intermediate values were included to obtain an understanding of the transition from clear to imperceptible outlier features. However, considering that there were 1,024 outlier and field parameter value permutations, not all of the intermediate results are presented. Instead, visual reviews of selected sets of results guide the analysis across particular scenarios that result in the clearest multi-scale signatures or potentially provide insights about a breakdown of any general pattern.

The presentation of the results and a discussion progress through various aspects of the findings. Presented first are the results of the boundary parameter values, as sets of both mixed grids and distograms. General patterns of multi-scale signatures of spatial dependence are described, respective to each of the outlier and background classes. Investigations of confusing cases and transitions follow, with comparison and contrast to clear cases. Selected ranges of distogram bands are then examined to focus on consistently distinctive characteristics related to outlier parts. Finally, results from correlated proxy variables are presented with regard for their potential to locate anomalies in the original variable of interest.

5.3.2 Original mixture grids

Examples of the mixed grid results are arranged as a 2D array in Figure 36. The figure shows representations of the 32 permutations, the boundary cases, of the smallest and largest values of the two outlier and the three field parameter values. The figure presents one example of the 30 field iterations for each of the boundary cases. Of the sixteen sets of outliers shown in Figure 32, only four sets are represented in the 32 mixed grid boundary cases, the four permutations of the smallest and largest values of the outlier height and width parameters. The shades of gray in each mixed grid represents the attribute value range for each data set, from lowest (white) to highest (black). Since the range is not constant across all data sets, a comparison cannot be made based upon shade. However, the independent ranges enable the visibility of subtle details in smaller ranges.

The combination of outlier and field simulation parameters that generated the mixed grids relate to two general components of variability: the spatial extent and the attribute value. Patterns in the mixed grids change spatial extent, dependent upon the outlier width and the field range. Variability in the attribute value of the mixed grids is dependent upon the outlier height and the field nugget and partial sill. There is a large enough range of parameter values to cover a range of scenarios from background fields with clear outliers to fields with visibly imperceptible outliers, in the upper-right and lower-left quadrants of the mixed grid array, respectively.

Although there are several grids with easily discernible spatial outliers, the embedded outliers become lost when the variability of the field matches or exceeds that of the outlier. For example, the smallest outlier has a (standard deviation) width value of 10, creating

an outlier with a radius of 30. The largest field range radius is 40, which potentially creates patches larger than the smallest outlier. The variability of the attribute values is also important to the differentiation between outliers and the field. The smallest outlier height is 10. The largest combination of nugget and partial sill is 12 + 40, which equals a variance value of 52. Taking the square root of variance yields a standard deviation value of about 7.2. Assuming a Normal distribution, due to the Gaussian unconditional simulation, the outlier height of 10 has a z-score of about 1.38 and a one-tailed p-value of just over 0.08 meaning that over 8 percent of the simulated field values are expected to exceed the outlier height. The quantitatively-based expectations support the visual perception of low variability fields with clearly extreme outliers, in the upper-right quadrant of the mixed grid array, transitioning into a more confusing situation of small outliers disappearing into high variability fields, in the lower-left quadrant of the mixed grid array shown in Figure 36.

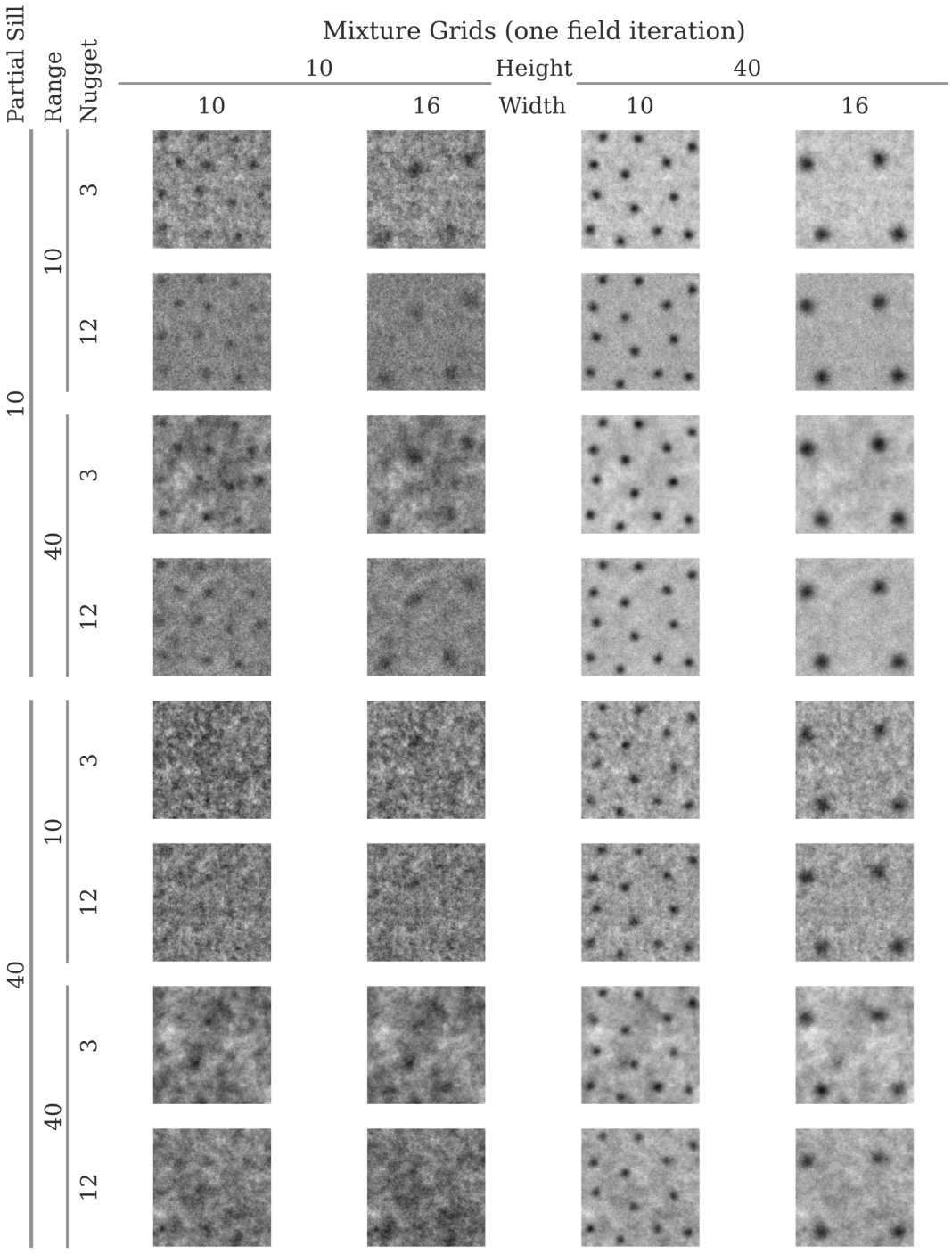


Figure 36: Example simulated mixture grids (one of thirty iterations) for each of the boundary outlier and field parameter values. Note that the grayscale values vary for each mixed grid, but each ranges from lowest (white) to highest (black).

5.3.3 Distograms, nine bands

Examples of the final distogram results are arranged as a 2D array in Figure 37. The figure shows representations of the 32 permutations, the boundary cases, of the smallest and largest values of the two outlier and the three field parameter values. Each distogram graph in the figure contains, for each outlier part and background class, the average and 95% confidence interval of the 30 distograms of the iterations of each field parameter set. The arrangement of distogram graphs corresponds to that of the example mixed grids in Figure 36. The arrangement was selected by the apparent importance of each parameter, with regard to the effects upon the output distograms. The ordering by parameter values, therefore, begins to provide insights about the properties of spatial outliers represented as Local Moran's I distograms.

Various alternatives were considered to select the order. One alternative for identifying important parameters was to use nested loop plots, as employed in the previous chapter. The requisite regression analysis assists in the identification of parameters that have the most influence on the Local Moran's I values of the final distograms. However, the best approach would be to apply that method for each band, as opposed to conflating characteristics of all of the bands. A problem arises if the regression results identify different parameters as important across the bands. There would potentially be difficulty in selecting a unique order of important parameters that works best across all the bands. Even more nuanced difficulties might arise if, for any band, multiple parameters are found with nearly the same degree of association. As a benefit, though, that situation could potentially justify flexibility in the selection. Consider a special circumstance that if a parameter is found to be top-ranked on some bands and second-ranked on others it could be used for the entire distogram. Ultimately, though, the series of distogram band values

presented too many difficulties for a quantitative selection by means of regression upon each band. However, there is still a potential for nested loop plots to be helpful for identifying the most important parameters for one band at a time if that is the focus of analysis.

Instead, the approach for ordering relied upon a visual inspection of the 32 final distogram graphs along with reasoning about several aspects of the distogram patterns. First, the distograms are arranged as a 2D array with the rows and columns corresponding to the field and outlier parameters, respectively. The columns have a major to minor order of outlier height, first, then width. An alternative would be to separate the column groups by width first, as a reasonable grouping is by the linear pattern of multi-scale signatures. Instead, the grouping started with height first, as more importance was given to grouping the columns by the clarity of the multi-scale signatures, particularly with respect to the breadth of the confidence intervals but also by the Local Moran's I magnitudes. Distograms of the higher outliers have the narrowest confidence intervals, which increases the distinctiveness between outlier part classes. Another reason was to group the distograms having the most confusion together, into the lower-left quadrant of the distogram array. Those distograms are associated with the lowest outliers and are embedded in the fields of highest variability, as discussed below.

The selection of row order also followed reasoning based upon visual review, however, considerations further involved two general parameter components, the attribute value and the spatial extent. Distogram graphs with various row ordering of the field parameters were reviewed and the selection made as follows. As seen in the distogram array, the nugget has least effect. This is evidenced in each column. There are similar graphs for each pair of nugget values across each of the four partial sill and range permutations. The similarity is with regard to both the linear pattern and magnitude of Local Moran's

I values. The nugget, therefore, although the actual minor of the three field parameters, is considered nearly negligible. The partial sill and range were selected as the major and the nominal minor parameters. This ordering were selected partially by following the magnitudes of the Local Moran's I values. In addition, another reason for this ordering was to assist the grouping of the most confusing distograms, of low outliers embedded in fields with the large patches of high values, into the lower-left quadrant of the distogram array shown in Figure 37. Finally, the partial sill and range order parallels to the outlier height and width order, with respect to the two general parameter components of attribute value and spatial extent.

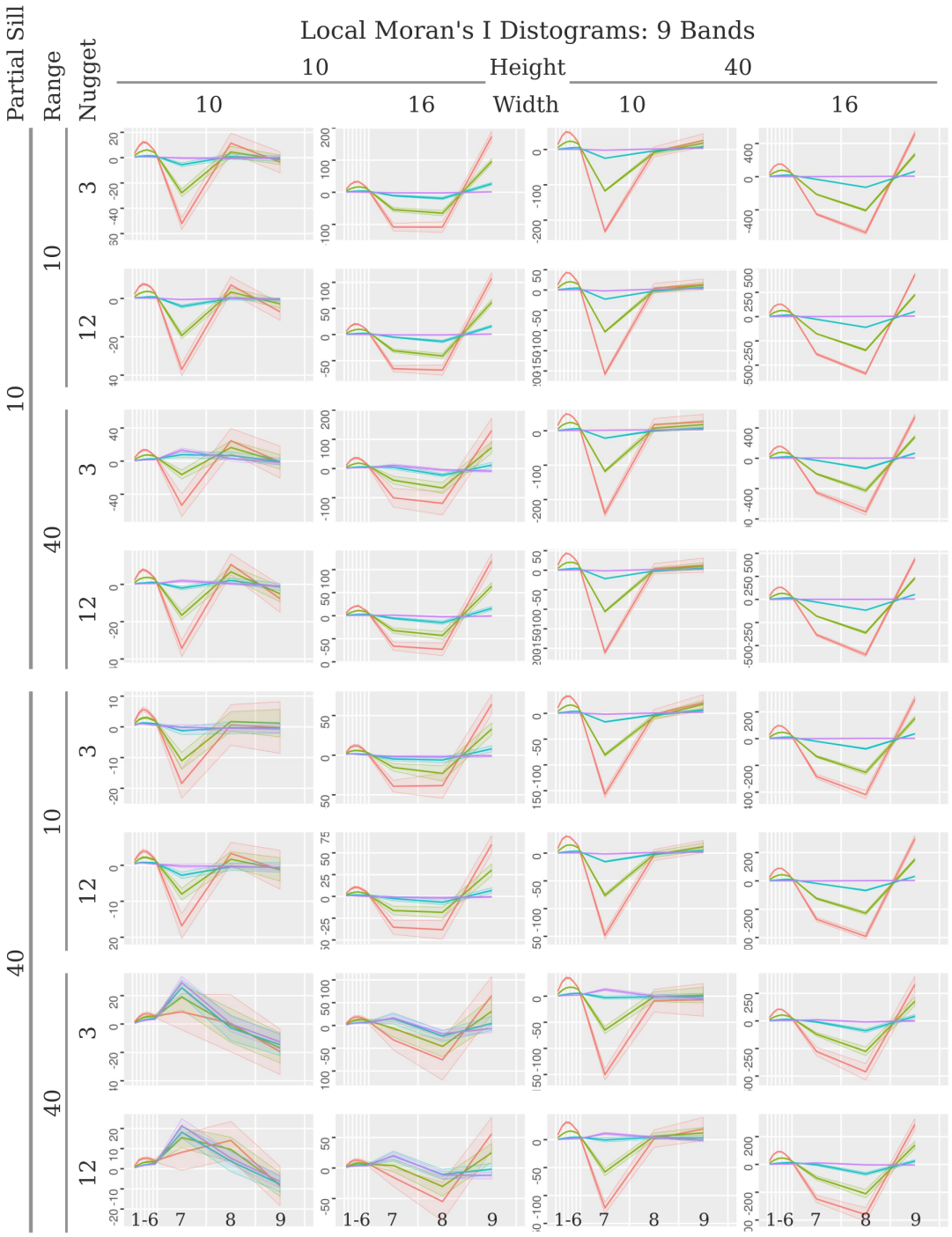


Figure 37: Local Moran's I distograms of boundary outlier and field parameter values. Showing nine bands, up to a distance that could include an outlier, a gap, a neighboring outlier, and the background field. The top, side, base, and background classes are colored red, green, blue, and purple, respectively, each with a central average value and semi-transparent color out to the 95% confidence intervals.

5.3.4 Background class

Before discussing patterns of outlier parts, there is reason to consider the distogram of the background field. The distograms of the background class have points of origin not coincident with an outlier. An expectation is that the Local Moran's I values of background distograms are responsive characteristics of the background field. The Local Moran's I distograms of the background class do not visibly change much and remain near zero, compared to the extreme values associated with the outlier parts. This apparently neutral pattern might be an artifact of the unconditional Gaussian simulation, which does include some degree of salt and pepper random noise that can dampen signals of patches of otherwise similar values. Also, the background class distograms have origins across the region, at various distances from outlier and field features. When aggregated, the various background distogram patterns likely offset creating the apparently neutral pattern of the Local Moran's I values closer to the typically expected Local Moran's I range of -1 to 1. This is opposed to the extreme Local Moran's I values of distograms with origins in regions of known spatial outlier parts. Measuring the background fields in isolation, without outliers, would reveal the effects of the variable field parameters upon the Local Moran's I distograms.

There is, however, at least one pattern evident for the background class in Figure 37. The background class has increased positive Local Moran's I values in band 7 in four pairs of graphs for which the field range parameter is 40, the parameters of which are shown in Table 4. This increase in local spatial autocorrelation generally occurs in fields with large patches of high values and with small outliers of low height and/or narrow width. The only exception is in the last row of the table, where the field of low values contains only low and narrow outliers. All together, these four scenarios represent situations in

which the small outliers have little effect compared to the field variability. So, patterns of high Local Moran's I values associated with the field parameters begin to appear in the distograms.

Local Moran's I, Band 7	Range	Partial sill	Height	Width
~25	40	40	10	10
~20	40	40	10	16
~10	40	40	40	10
~5	40	10	10	10

Table 4: Background class band 7 positive Local Moran's I values with associated field and outlier parameter values.

5.3.5 Outlier part classes

Distograms of the outlier part classes have points of origin coincident with an outlier. The Local Moran's I distograms are responsive to the presence of locally extreme values. The distogram graphs contain various linear patterns, connecting values across the series of distogram bands, that are dependent upon the outlier and field parameter values. The shape of the linear graphs is one of three aspects of the outlier part analysis. There are two other aspects important for comparison. Patterns of the distogram confidence intervals and Local Moran's I magnitudes are introduced before addressing the linear patterns, transitions, and specific, related outlier and field parameter values.

The breadth of the confidence intervals is considered a key aspect of ascertaining whether outlier part classes are distinctive to one another and the background. In scenarios with relatively large outlier height values that highly contrast with low field variance (i.e. partial sill and nugget) the confidence intervals are the smallest. The fine confidence intervals result in clear separation between outlier part distograms, which indicates that the characteristics of multi-scale spatial dependence are distinctive for each class. As

the height of outliers becomes nearly equal to the background field variability the outlier part confidence intervals expand and eventually overlap, indicating that the separation of classes breaks down.

The second aspect of the initial comparison concerns the magnitudes of Local Moran's I values. Across all of the 32 boundary case distogram graphs, the outlier parts have magnitudes much larger than the commonly expected Local Moran's I range of -1 to 1. Scenarios with outlier values that highly contrast with the background field have the most extreme Local Moran's I values. The tops are most sensitive and responsive to the perturbance of outliers embedded into the fields. In this regard, the tops are the most distinctive class, especially compared to the background. The sides and the base outlier parts have similar linear patterns as the tops but, in that order, the Local Moran's I values tend to approach the relative neutrality of the background.

The Local Moran's I distograms effectively capture two characteristics in one series of values. Nearby and distant neighbors in the top class exhibit the characteristics of the highest and lowest local spatial autocorrelation, respectively. That is especially the case when the point of origin is at or near the center of the embedded outlier. However, if a point in the top class is near the outer boundary of the top class region, the set of neighbors includes points in the side class. The major transition of attribute values in the side class neighbors increases the variance and decreases the Local Moran's I values in a distogram of a top point near the boundary.

The inverse situation occurs for points in the side class. If a point of origin in the side region is near the class boundary between the top and side, its set of neighbors includes points in the top region. Including the relatively stable attribute values of the top region

results in less variance and an increase in Local Moran's I values for those side points. Although a major transition of attribute values on the side of an anomaly might be expected to exhibit purely low local spatial autocorrelation, by aggregating some values of the top with those of the side, the side class has the second-ranked Local Moran's I distogram values for both of the characteristics of local spatial autocorrelation in the series (i.e. positive values near the point of origin and the negative values at larger distances). Like any general histogram, this effect of binning points aggregates and generalizes the summary values. Even though the side class distogram values do not represent purely low local spatial autocorrelation, their linear pattern is often distinct from, and occurs between, the patterns of the top and base classes.

The same situation continues for distograms of the base class, representing a minor transition of outlier values between those of the side and the background. Points in that class are partially influenced by both the major side transition and the background variability. As a result, the base class has the least spatial autocorrelation of the outlier parts and is most similar to the background.

It is worth noting that the Local Moran's I distogram values are affected by the relative area in each band. Assuming equal bandwidth, inner bands cover less area than outer bands. Therefore, fewer neighbors are used for the Local Moran's I calculation. Neighbors of a small set have more influence on the Local Moran's I value than neighbors of a large set. In the top class, nearby points have strong influence and produce high spatial autocorrelation. The base class also includes the top points, but in a band farther from the origin, which encompasses more area and a larger set of neighbors. The contribution of the top points and their characteristic of relatively high spatial dependence is diminished in the base class.

The magnitudes of Local Moran's I values are also affected by the variable values of the outlier and field parameters. Table 5 provides a ranked listing of the approximate change of the largest Local Moran's I value, relative to the parameter values and any special conditions. The table includes only approximations of the factor of change on the largest of Local Moran's I values in the distogram graphs, for a single band of the top class. It does not represent factors of changes across each and every band. Also, the ranking does not indicate the most influential parameters but rather the sensitivity of the Local Moran's I value across the parameter value ranges. In other words, a parameter might affect the Local Moran's I value, but not necessarily provide a major contribution to the result.

Factor	Parameter	Min	Max	Condition
5	Height	10	40	Nugget 12
4	Height	10	40	Nugget 3
2	Width	10	16	
2	Range	10	40	Partial sill 40, Height 10
-0.5	Nugget	3	12	Width 16
-0.5	Partial sill	10	40	Nugget 3
-0.8	Partial sill	10	40	Nugget 12

Table 5: Approximate factors of change, per variable, on the largest Local Moran's I magnitude in the nine band distogram graphs of the 32 boundary cases.

A key finding in Table 5, of ranking the response of maximum Local Moran's I values to the variable parameters, relates two considerations already discussed. Increasing the outlier height majorly affects the Local Moran's I of the top class. That follows the consideration in the visual review about confidence interval widths for ordering the distograms in Figure 37. The finer distogram confidence intervals were associated with higher outliers. It also follows the observation of the largest magnitudes of each class,

that the tops had the most extreme positive and negative Local Moran's I values.

Across most of the 32 distograms of boundary parameter values there are generally stable linear patterns. This is partially due to the controlled shape of outliers and the design of controlling distogram bands relative to the scale of outliers. It is worth noting that a practitioner working with real data, subject to complexities of mixed signals and noise and without a clear set of outliers, will likely encounter difficulties in selecting appropriate bandwidths that match patterns of interest. An outlier of known scale and properties might guide an approximation for selecting bandwidths to search for other outliers of the same character.

With regard to the 32 distograms of the boundary cases in Figure 37, there are apparently two different linear patterns, in each of the two pairs of columns. The patterns differ based upon the outlier width parameter, suggesting a transition between the two patterns. Considering a possible transition reveals a potential of two main components of the linear patterns. The first component is relatively stable and evident in all of the distogram graphs. That linear pattern is an arch of moderately high Local Moran's I values across bands 1 to 6. The second component occurs in most of the distogram graphs, except for a few confusing graphs in the lower-left quadrant. It is a pattern that involves low Local Moran's I values and varies relative to outlier width. With the narrowest outliers, the low values form a dip only in band 7. With wider outliers, the low values of band 7 seem to extend the dip out to a broad trough into band 8. Other effects in band 9 appear to be related to this extension of the linear pattern with wider outliers. Those transitional variations in bands 7, 8, and 9 are examined in the next section, by addressing the more confusing cases in the lower-left quadrant of the distogram array.

5.3.6 Examples of distogram transitions

Although many of the distograms of the boundary cases exhibit generally stable linear patterns, there are several confusing cases in the lower-left quadrant of Figure 37. They are confusing in the sense that they have broad overlapping confidence intervals, which indicates the classes are not distinctive, and also in the sense that the linear patterns deviate from the stable versions. In order to investigate transitions between various linear patterns, four examples follow, three of which involve broad overlapping confidence intervals and unusual linear patterns. The fourth example, with fine confidence intervals, is included to contrast the three confusing cases with this clearest set of distinctive outlier part and background classes. The last set builds upon and extends a discussion about the transition between the two linear patterns generally evident in the distogram array, in each of the two pairs of columns.

Confusion highlighted by the examples is associated with both of the two general components of variability, the attribute value and the spatial extents. A lack of contrast between the attribute values of outlier and field locations produces a confusion factor expressed as overlapping confidence intervals in the first three examples (as opposed to the fourth clear case). All of the confusing cases involve a low outlier height and a field with high variability in both the partial sill and nugget values. However, the discussion of the examples is mainly related to the linear patterns of the distograms, which are influenced by parameters related to spatial extents. To that aim, the first of the examples addresses changes in the field range and the last three address changes in outlier widths. All of the four examples focus on bands 7, 8, and 9, beyond the radius of an outlier. Since it is difficult to visually identify low outliers in the highly variable fields, the outlier sets in Figure 32 are a helpful reference.

The first example concerns a transition between distogram graphs at (row, column) (6,1) and (8,1) in the Figure 37 distogram array (i.e. the third from the bottom and the bottom graphs in the left column). Figure 38 and Figure 39 show distograms and mixed grids with the same outliers and various field range parameter values. Of interest is the band 7 transition from negative values found in the first graph, and in the majority of results, to the positive values found in fields with larger range parameter values.

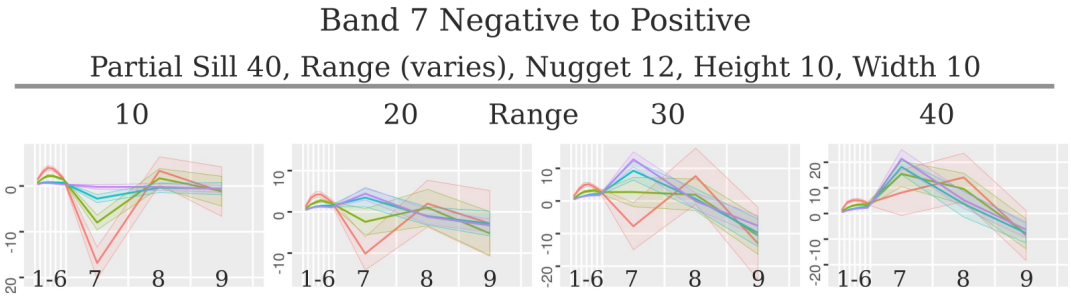


Figure 38: Distograms for the transition of Local Moran's I in band 7 from negative values to positive values. The top, side, base, and background classes are colored red, green, blue, and purple, respectively, each with a central average value and semi-transparent color out to the 95% confidence intervals.

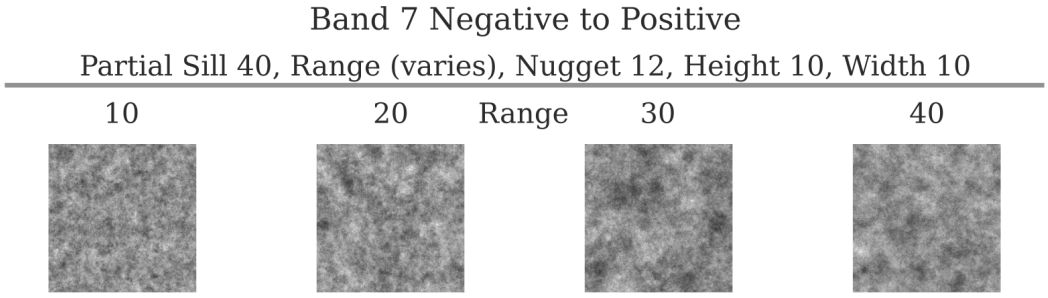


Figure 39: Mixed grids for the transition of Local Moran's I in band 7 from negative values to positive values. Note that the grayscale values vary for each mixed grid, but each ranges from lowest (white) to highest (black).

Figure 38 starts with a distogram that has a relatively common linear pattern for

small (low and narrow) outliers, the pattern transitions to a situation with nearly no difference between classes for the outer three bands. Only the top is visually different, but appears to trend toward a pattern matching the other three classes. The transition in this first example involves only an extension of the field range. With small range values, the classes have some distinction in band 7. The top class has the lowest Local Moran's I in its band 7, which corresponds to a neighborhood including both the background field in the imposed gap between outliers and the potential of origin points near the edge of the top class to possibly include neighboring outliers. The background class has the highest of the Local Moran's I values, but they are nearly neutral. This is likely due to the small field range creating fields that exhibit relatively mixed characteristics when values are combined across the relatively large bandwidth of band 7 (and the two other outer bands, 8 and 9). Across the transitional sequence, the field range becomes larger and the characteristics of higher spatial autocorrelation in band 7 of the field become apparent. This is likely due to the range becoming matched with the outer distogram bandwidths, as previously mentioned in 5.3.4 "Background class". As evidenced in Figure 39, the outliers are small and arguably imperceptible. Although the outliers are present, the effects of the field have the most influence on the linear pattern of the Local Moran's I distogram.

The second example concerns a transition between distogram graphs at (row, column) (8,1) and (8,2) in the Figure 37 distogram array. Figure 40 and Figure 41 show distograms and mixed grids of the same field with various outlier width parameter values. In band 8, the Local Moran's I values transition from positive to negative. In band 9 is the reverse, a transition from negative to positive.

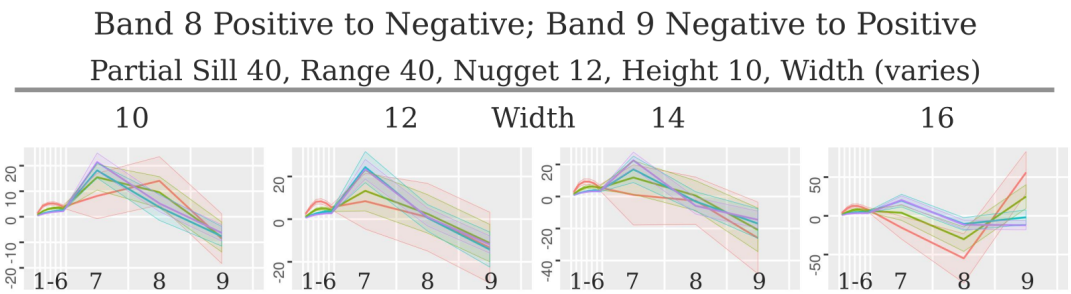


Figure 40: Distograms for the transition of Local Moran's I values in band 8 from positive to negative values and in band 9 from negative to positive values. The top, side, base, and background classes are colored red, green, blue, and purple, respectively, each with a central average value and semi-transparent color out to the 95% confidence intervals.

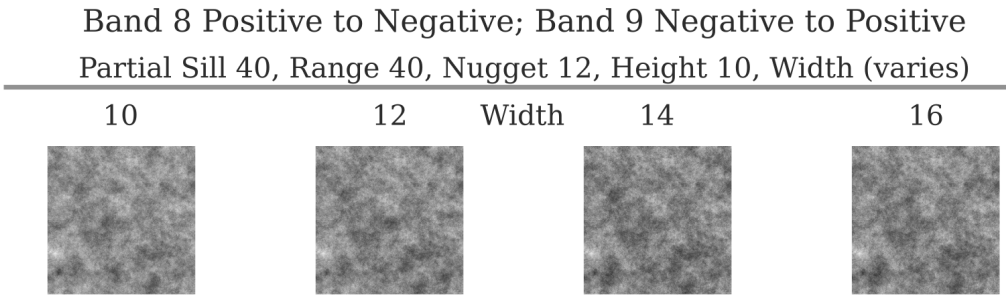


Figure 41: Mixed grids for the transition of Local Moran's I values in band 8 from positive to negative values and in band 9 from negative to positive values. Note that the grayscale values vary for each mixed grid, but each ranges from lowest (white) to highest (black).

Figure 40 starts with the last graph in the sequence shown in the first example, the most confused of the 32 distogram graphs in the array of boundary cases. It has the smallest (lowest and narrowest) of the outliers embedded in a field with the largest field range, partial sill, and nugget. The sequence varies only the outlier (standard deviation) width parameter across values in the set {10, 12, 14, 16}, which correspond to outlier radius values in the set {30, 36, 42, 48}. The top class begins with positive values in band 8. This is possibly related to the interpretation in the first example where, for the smallest outliers, any signal of the outlier part with the most extreme Local Moran's I values becomes overwhelmed by the high variability field, especially in such a field with a large

range, which generates large patches of similar values. As the outlier width approaches and exceeds the field range, the transitions in both bands 8 and 9 are consistently smooth and begin to reveal a pattern that is evident in most of the other boundary case scenarios with wide outliers. At the largest outlier width, there is evidently low local spatial autocorrelation in band 8, likely due to the large band capturing more of the effect of high variability of the background field, as opposed to a relatively minor contribution of the low and wide outliers. A similar effect is evident in band 9, where the first three of the sequence of graphs show all outlier part classes with negative Local Moran's I values. Only in the scenario with the widest of outliers does band 9 switch to positive values, with increasing values from the base to the top class. The final linear pattern of negative Local Moran's I values in band 8 and positive values in band 9 occurs in all scenarios involving wide outliers in the array of distogram boundary cases. Discussion of that pattern continues in the following examples with the aid of scenarios with more clarity of contrast between outliers and the background field.

The third example concerns a transition between distogram graphs at (row, column) (6,1) and (6,3) in the Figure 37 distogram array. Figure 42 and Figure 43 show distograms and mixed grids of the same field with various outlier width parameter values. In band 8, the Local Moran's I values transition from relatively neutral to negative. In band 9 is the reverse, a transition from neutral to positive values.

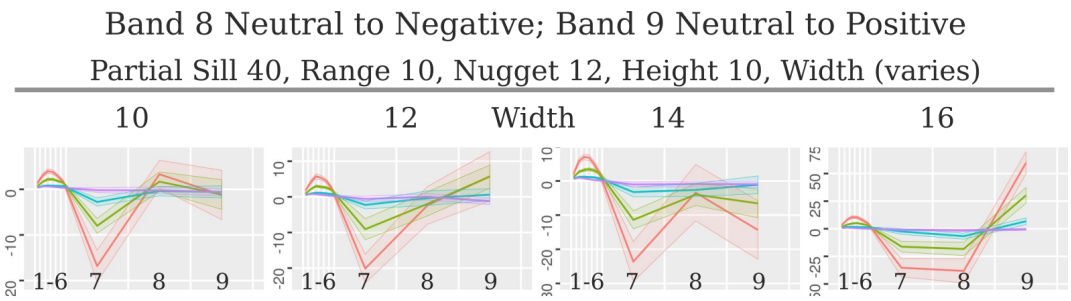


Figure 42: Distograms for the transition of Local Moran's I values in band 8 from neutral to negative values and in band 9 from neutral to positive values. The top, side, base, and background classes are colored red, green, blue, and purple, respectively, each with a central average value and semi-transparent color out to the 95% confidence intervals.

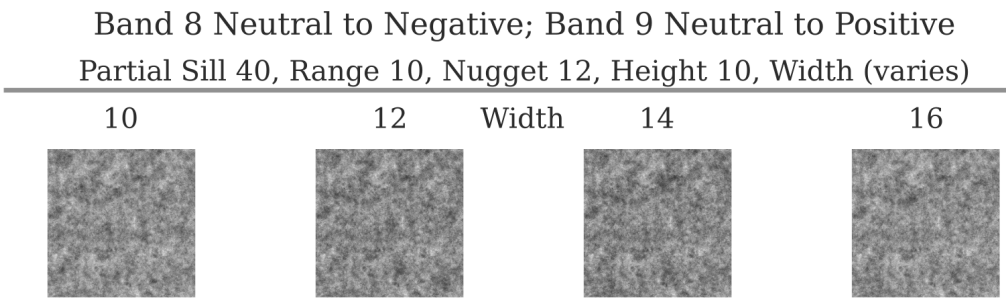


Figure 43: Mixed grids for the transition of Local Moran's I values in band 8 from neutral to negative values and in band 9 from neutral to positive values. Note that the grayscale values vary for each mixed grid, but each ranges from lowest (white) to highest (black).

Figure 42 starts with the first graph in the sequence shown in the first example. It has the smallest (lowest and narrowest) of the outliers embedded in a field with the largest partial sill and nugget, but the smallest range value, which creates a background field with smaller, noisier patches. As in the second example, the sequence varies only the outlier (standard deviation) width parameter across values in the set {10, 12, 14, 16}, which correspond to outlier radius values in the set {30, 36, 42, 48}. Across the sequence of graphs the confidence intervals of the outlier part classes do not overlap in bands 1 through 7 due to larger Local Moran's I magnitudes that are dependent upon distinctive neighborhood characteristics of each outlier part class and different from the noisiest

background field. On the other hand, in bands 8 and 9, all of the outlier part classes have Local Moran's I values approaching neutrality in the first graph of the sequence, representing the narrowest of the low outliers, which evidently have less influence in the larger outer bands. As the outliers become wider, band 8 slightly decreases until the widest outlier has extreme negative Local Moran's I values that are comparable to those in band 7. This appears to indicate that the band 7 characteristics begin to extend into band 8.

The extension of the linear pattern might also be related to the pattern of change for band 9. It appears that there is a corresponding movement of positive values from band 8 in the first graph to band 9 in the second graph. However, the top class temporarily flips band 9 Local Moran's I values to negative, in the third graph, and back to positive, in the fourth graph. Examining this transition uncovers perplexing results, but it is worth noting that this example includes scenarios that involve considerable confusion with nearly imperceptible outliers in a highly variable field, as shown in Figure 43. Regardless, there are three possible situations that might explain this transition. First, the third graph represents scenarios involving an outlier radius of 42, which is very similar to the field range value of 40. There is likely some interplay between the embedded outliers and patches of similarity in the fields as some patches in the field might be augmented by an embedded outlier around the same location. Second, if the outliers with the third largest radius are packed relatively tightly in the raster grid, the distogram of the top class might capture a nearby outlier in band 8, yielding high local spatial autocorrelation. Band 9 might capture either another imposed gap or a combination of outliers and background, yielding low local spatial autocorrelation. Finally, if this arises from a band sampling issue, an adjustment of distogram bands to different positions, or with finer bandwidths, might aid a deeper investigation.

The last graph eventually stabilizes, as does the first graph, into linear patterns consistent across outliers of the same widths in the other boundary cases. In the last graph the clear and common pattern, for the widest of outliers, is of negative to positive values in bands 8 to 9. It is reassuring that this third example of the confusing cases, in which low outliers of each width are visually imperceptible, the linear patterns of the Local Moran's I distograms begin to match those of the clearest cases involving the largest outlier height values.

The fourth example concerns a transition between distogram graphs at (row, column) (1,3) and (1,4) in the Figure 37 distogram array. Figure 44 and Figure 45 show distograms and mixed grids of the same field with various outlier width parameter values. This fourth example represents a clear case of distinctive outlier part and background classes. As in the third example, in band 8, the Local Moran's I values transition from relatively neutral to negative. In band 9 is the reverse, a transition from neutral to positive.

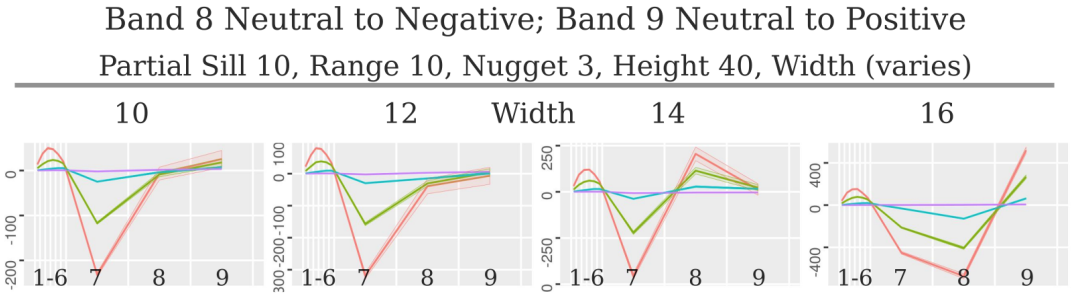


Figure 44: Distograms for the transition of Local Moran's I values in band 8 from neutral to negative values and in band 9 from neutral to positive values for a clear case of distinctive classes. The top, side, base, and background classes are colored red, green, blue, and purple, respectively, each with a central average value and semi-transparent color out to the 95% confidence intervals.

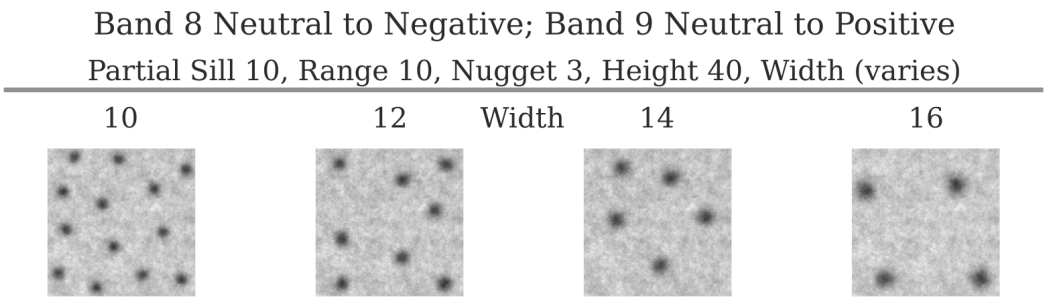


Figure 45: Mixed grids for the transition of Local Moran's I values in band 8 from neutral to negative values and in band 9 from neutral to positive values for a clear case of distinctive classes. Note that the grayscale values vary for each mixed grid, but each ranges from lowest (white) to highest (black).

In this fourth example, of a clear case, there is significant visual contrast between high outliers and low variability fields, as shown in Figure 45. Correspondingly, the minor variability previously observed in the third confusing case, with low outliers in a high variability field, are significantly reduced with this scenario of high outliers in a low variability field. This reduction alleviates issues found in both pairs of graphs of the narrower two and the wider two outliers.

In the third (confused) example with low outliers, the first and second graphs (of narrower outliers) each have slight positive values for band 8 and 9, respectively. These become less evident in this fourth (clear) example with high outliers. In the first and second graphs of this fourth example both bands 8 and 9 become relatively neutral. In part this visual change is relative to a considerable increase in the magnitudes of band 7, but the outer two bands have negligible signals of spatial autocorrelation in this clear case of high and narrow outliers.

Regarding the high and wide outliers, the clear case also alleviates the perplexing issues of Local Moran's I values switching sign, as previously discussed. For the third

and fourth graphs (of wider outliers), both the third and fourth examples have a similar linear pattern. The difference is that in the clear case the pattern becomes exaggerated and shifted to positive Local Moran's I values. More importantly, it appears that the pattern found with the second widest outlier expands for the widest outlier. Moving from the third to the fourth graph in this fourth example, the negative values in band 7 expand to band 8 and the positive values in band 8 move to band 9. This is the clearest case, but it also accords with the majority of other graphs found in the two pairs of columns in the distogram array of Figure 37.

The clear case begins to capture the high spatial autocorrelation in band 8 in the graph of the second widest outlier. That spatial autocorrelation extends to band 9 with the widest of outliers. This is likely due to the placement of wider outliers. Considering their larger size, the imposed gap between outliers and the constraint of the region extents become limitations on potential outlier placement locations. As opposed to the more random placement of narrow outliers, which exhibit visually perceptible clustering, wider outliers become more evenly distributed and move towards the corners of the region. They are all nearly the same distance from one another, separated by what appears to be about two to three outlier diameters. With the more even distribution, band 9 begins to more consistently capture the neighboring outliers. Also, especially for the wider outliers, the outlier part distogram origins are located near edges of the region. The edge effects decrease the area of the outer bands and increase the contribution of points within those bands. Compared to the outer bands of distograms with origins near the center of the region, the positive spatial autocorrelation of outlier tops is exaggerated with a calculation from fewer points in edge-affected outer bands.

The even outlier spacing and distogram edge effects are an artifact of the simulation

design. Larger raster grids might alleviate some constraints for investigating the same range of outlier width parameters. However, that would require exponentially increasing amounts of computational processing, particularly for determining the distance matrices for the points. Another possibility is to test whether an even distribution of neighboring outliers increases the Local Moran's I values in band 9. One way to achieve this for the scenarios with wider outliers is to remove all but one (or two) outliers in each scenario and compare the resulting distograms to those in Figure 37.

One last observation about this fourth example of a clear case is the comparable magnitude of Local Moran's I values associated with the imposed gap and the neighboring outliers. Although the contrasting relationship occurs in the compressed pattern of the third graph, it is particularly evident in the fourth graph of the widest outliers, which are most evenly dispersed and represented in band 9 of the distograms. Positive Local Moran's I values of neighboring outlier tops, in band 9, are comparable in magnitude to the negative values of the imposed gap, in band 8. Band 7 is similar to band 8, but appears dampened, likely due to the inclusion of points in outlier part regions. This dampening effect of aggregating outlier parts characteristics is the same as was previously discussed, in 5.3.5 "Outlier part classes", with respect to the difference in linear patterns of each outlier part class. In this circumstance, though, the dampening is with regard to nearby bands. This indicates the linear patterns of distograms, like any histogram, are dependent upon bandwidth selection. The effect is, therefore, evident in both of the two general components of variability, the attribute value and the spatial extents.

5.3.7 Distograms, seven bands

After considering the Local Moran's I distograms that extend far across the fields, a narrowed view of the Local Moran's I distograms of the 32 boundary cases is presented in Figure 46. That figure shows only bands 1 through 7. The narrowed distogram graphs provide a more straightforward perspective of the Local Moran's I values associated with outliers and a nearby surrounding field. Bands 1 through 6 cover the radius distance of an outlier. Beyond the boundary of the outlier, the breadth of band 7 covers the surrounding area across a distance equal to the diameter of the outlier. The linear patterns of Local Moran's I distograms in Figure 46 generally express a consistent pattern of local spatial autocorrelation related to the embedded spatial outliers.

First, it is noteworthy that of all the outlier part classes the top class has highest potential of representing the anomaly as a multi-scale signature of spatial dependence. The Local Moran's I distogram of the top class provides the most information about the outlier feature outward from its center. Compared to the other classes, the distograms of the top class clearly indicate a contrast of local spatial autocorrelation characteristics between the interior of the outlier and its surrounding area. This pattern of contrast is stable across the majority of scenarios, regardless of outlier and field parameters. The combination of two components, of positive Local Moran's I values at the scale of an outlier and of negative values just outside the outlier, is more informative than either component alone. It is, therefore, potentially the most informative pattern for discerning outliers. This is one of the key findings of the experiment.

The pattern only breaks down in the four distograms nearest to the lower-left corner of Figure 46. As presented previously, in 5.3.6 "Examples of distogram transitions", the

confusing cases involve outliers that are visually imperceptible. The scenarios represent low outliers, with a height of 10, embedded in high variability fields, with a partial sill of 40 and range of 40. The lowest outliers have minor neighborhood effects compared to the large patches of potentially high field values. In these four distograms band 7 has positive Local Moran's I values and the outlier part class confidence intervals overlap. The variability of surrounding region, overwhelms and confuses any potential classification based upon the Local Moran's I values in band 7 (as well as bands 8 and 9).

It is notable, though, that within the scale of the outlier feature, the effects of large field patches are dampened for the outlier part distograms. Evidently, the embedded outliers affect the distograms at that scale. Although the pattern of the surrounding area becomes confused in such scenarios, if the distogram matches the scale of the anomaly, the outlier part classes are distinct from each other.

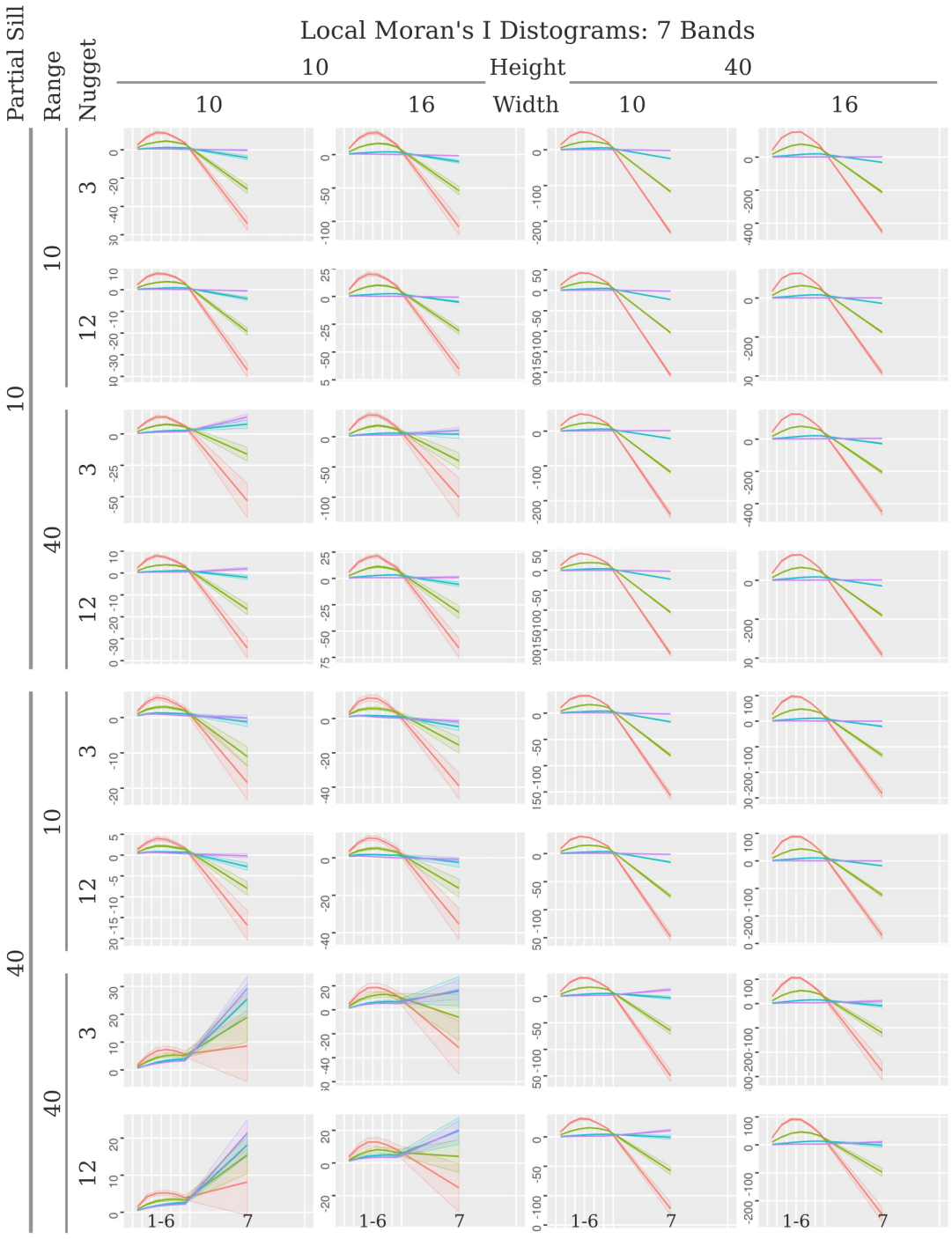


Figure 46: Local Moran's I distograms of boundary outlier and field parameter values. Showing seven bands, up to a distance that could include an outlier and the imposed gap. The top, side, base, and background classes are colored red, green, blue, and purple, respectively, each with a central average value and semi-transparent color out to the 95% confidence intervals.

5.3.8 Distograms, six bands

In order to more clearly illustrate that a Local Moran's I distogram, matched to the scale of a spatial outlier, can reveal distinctive characteristics of each outlier part class, another array of narrowed distograms is presented. Figure 47 shows the Local Moran's I distograms of the 32 boundary cases narrowed to bands 1 through 6. The six bands cover the radius of an outlier. For the top class, the bands roughly correspond to neighborhoods of outlier parts, with the bands 1 and 2 near the top, bands 3 and 4 near the side, and bands 5 and 6 near the base. The side, base, and background classes represent points of origin that are increasingly shifted away from and not centered on outliers. As such, those three classes provide decreasing information about the neighborhood effects directly related to the embedded outlier.

In general, across bands 1 through 6 and relative to the scale of the outlier, the linear pattern of the Local Moran's I distograms are similar. An arc of Local Moran's I values begins with a relatively small positive value in band 1, it increases to the largest height around bands 3 and 4, and decreases back toward neutrality in band 6. The small positive value for the small, circular neighborhood of band 1 was not expected. Three possibilities are suggested as causes. First, the value of the small neighborhood might characterize spatial autocorrelation as affected by the field nugget value. Second, each of the aggregated small neighborhoods in the class might only capture small sectors of curved variation at the top of the Gaussian form, potentially exhibiting low spatial autocorrelation and for which the relatively few data points would have strong influence on the Local Moran's I value.

Description of the third possible cause is more involved, and is related to the damp-

ening due to neighborhood characteristics of points near the distogram boundaries, as previously discussed with respect to the outlier classes. It is an aggregation effect common to histograms, but is also an effect of assuming isotropic two-dimensional spatial decay outward from the center of a point type of anomaly. There is a smaller proportion of distogram origin points near the center of the top region as compared to the proportion of origin points near the edge of the circular top. By design, points near the center of the top have higher local spatial autocorrelation, but they are relatively few in number. Points near the boundary between the top and side regions exhibit high variability, low spatial autocorrelation, and are more numerous. Aggregated to represent the top class, the few central points with high spatial autocorrelation have minor effect compared to the relatively abundant top origin points with low spatial autocorrelation toward the side.

The dampening effect of aggregating larger numbers of neighborhood points near the outer band boundaries makes some sense. However, the logic doesn't seem to hold for band 1, as its outer boundary is with band 2, which also represents the top region, arguably with more variation as its outer origin points approach the side class. Further, following the reasoning of the aggregation effects, band 2 should capture more variation, but instead returns a higher Local Moran's I. Therefore, some other influence appears to have affect. Placing less importance in the third reason, related to the proportions and contributions of points moving away from the center point of an outlier, and the second reason, related to the changing neighborhood characteristics, the first suggested reason remains. For every outlier class, the small neighborhoods of band 1 might be capturing the effect of random field variation at small distances, related to the nugget values.

Nonetheless, following the aggregation effects through the sequence of bands might aid an explanation of the rest of the arc of Local Moran's I values. Band 3, for example,

is about half of the size of the outlier. If the point of origin is at the center of the outlier, then it would not include any top points in the Local Moran's I calculation. That value is aggregated with other top points, which are not at the center of the outlier. Depending upon the distance that the point of origin is shifted away from the center, out toward the boundary of the circular top region, larger patches of top points will be included in the calculation of Local Moran's I, yielding higher values. Now, consider the low Local Moran's I values for band 6. Even when a point of origin in the top class has moved a distance of two bands, to the edge of the top region, the annulus of band 6 never intersects the top region. No top points are included in the Local Moran's I calculation of band 6, only points of other classes (i.e. the major transition of the side, the minor effect of the base, and the background variability), which together yield lower Local Moran's I values.

Considering the three examples of confusing cases previously discussed in 5.3.6 "Examples of distogram transitions", it is surprising that the multi-scale signatures for those cases are in any way distinctive, even at the scale of the outlier feature. It is nearly impossible to discern anomalies in the mixed grids of the lower-left quadrant of Figure 36. Moreover, bands 7 through 9 for those cases have overlapping confidence intervals indicating class confusion. Nonetheless, the classes are distinctive across nearly all of the bands 1 through 6 across the array of 32 boundary cases, even for these challenging cases. This suggests that, under circumstances where a distogram matches the outlier feature, the quantitative approach has the potential to outperform human perception with regard to a multi-scale signature of spatial dependence. The Local Moran's I distogram identifies distinctive characteristics of spatial outliers and their parts in highly variable fields.

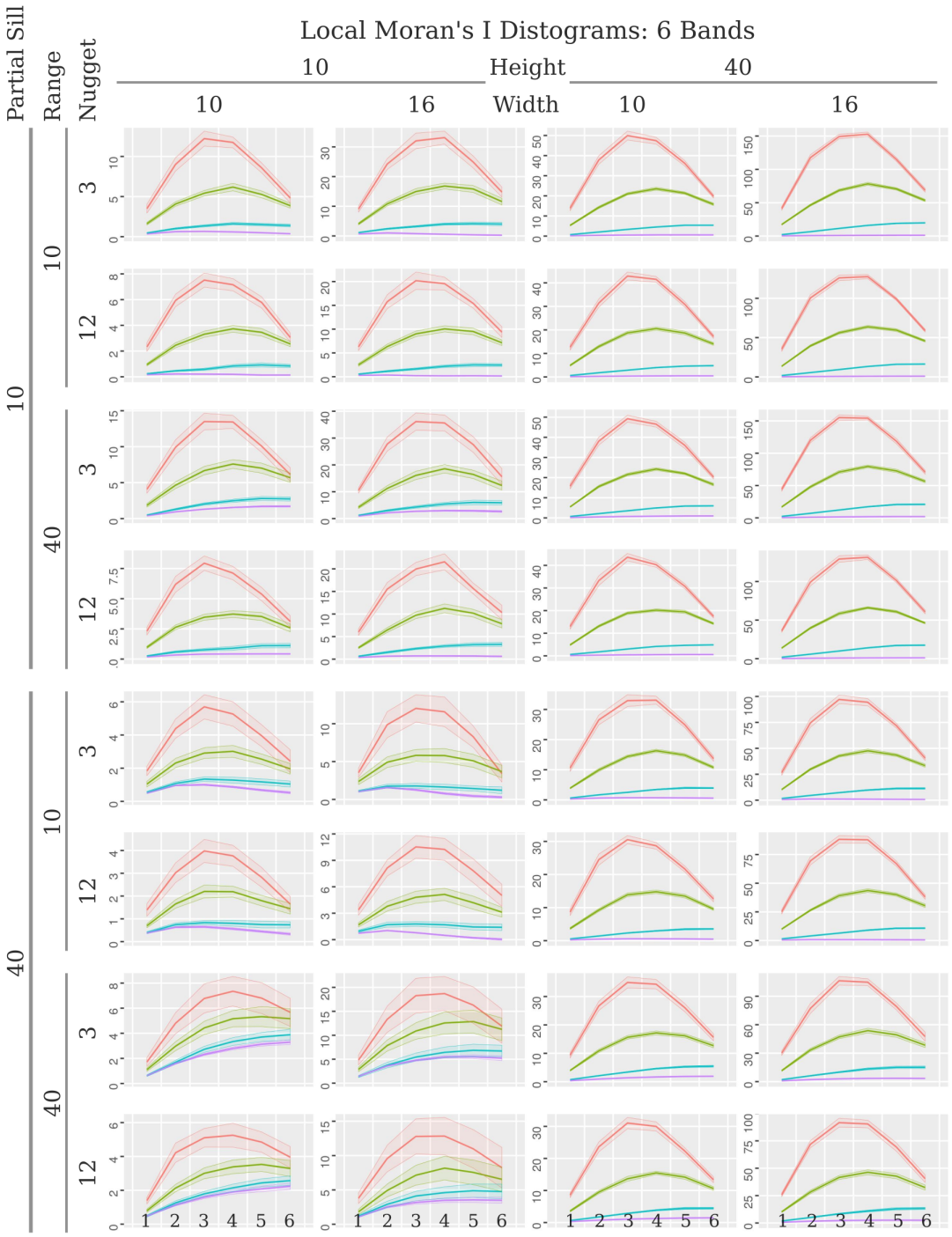


Figure 47: Local Moran's I distograms of boundary outlier and field parameter values. Showing six bands, up to a distance that could include an outlier. The top, side, base, and background classes are colored red, green, blue, and purple, respectively, each with a central average value and semi-transparent color out to the 95% confidence intervals.

5.3.9 Correlated variables

This section focuses on correlated proxy variables and their potential for providing information about distinctive outlier part and background classes. Up to this point only the results of the original grids have been discussed. Cases that gave the clearest results, scenarios with high outliers in low variability fields, and their stable distogram patterns have already been considered. Confusing cases were also discussed, where the original mixed grids produced outlier part class distograms that were not distinctive. The analytical perspective of this section generally begins with a clear original variable and progresses to noisier variables with decreasing correlation to the original. However, this part of the experiment also informs the reverse viewpoint. In practice, one might be interested in a variable with nearly indistinguishable features. Searches for spatial outliers in a fuzzy variable of interest are potentially guided by information of a correlated variable with clearly perceptible anomalies.

One of the aims of the experiment was to understand patterns of breakdown in the ability of the proposed approach to find distinctive multi-scale signatures of spatial dependence for outlier parts. The main concern of this section is the degradation of clear patterns, from the original grids, into fields overwhelmed by Gaussian noise, across a sequence of decreasing target correlation values for the correlated grids. In order to examine confusion across scales of the outlier and patches in the field, variability related to the property of spatial extents was the main focus. Variation related to attribute value was considered a straightforward finding, already addressed. Scenarios with high outliers in low variability fields produced the clearest outlier part distograms with small confidence intervals separated by large differences in Local Moran's I magnitudes. Because the previous arrays of 32 distogram graphs of boundary cases showed that the nugget

parameter had relatively little effect upon the distograms, it was not a factor covered in this analysis. Only data sets with the lowest nugget value of 3 were considered.

A small selection of four cases of related scenarios are presented as examples that approach and arguably cross into a breakdown of distinctive Local Moran's I distogram patterns. Three arrays of distograms are presented for nine, seven, and six bands in Figure 48, Figure 49, and Figure 50, respectively. As opposed to the extreme boundary cases, the examples generally involve intermediate parameter values. The partial sill and outlier height are held at 30 and 20, respectively. Two pairs of columns are grouped by outlier (standard deviation) width values of 14 and 16. The larger outlier widths were selected in relation to a sequence of three field range values of 20, 30, and 40. The middle range value is represented in the two central columns, once for each outlier width.

Shown above the distograms of correlated grids are the original mixed grids. The fields exhibit patterns of variability with moderately large patches to which the characteristics of the embedded low outliers are comparable. The outliers are somewhat lost in the field variability. Referring to Figure 32 aids with identification of outlier locations.

Nine bands of the Local Moran's I distograms are shown in Figure 48. The distogram covers the distance of an outlier, an imposed gap, a potential neighboring outlier, and the mixed field beyond. The discussion for this figure focuses on the two outer distogram bands. The outer distogram bands 8 and 9 of the left and right pairs of columns have different patterns arising from differences between the outlier sets. Regardless of variation in the spatial extents of patches in the fields, across all of the different field range values, the outliers in the left and right pairs of columns are self-similar and different from one another. The change based upon outlier sets is highlighted by the change in

outlier part distograms in the central two columns, where the field is the same. The left pair of columns have a relatively strong signal of local spatial autocorrelation in band 8. The set of five narrower outliers in the left pair of grids are more evenly spaced than the four wider outliers in the right two grids. Although the placement of the largest outliers was not the most evenly dispersed in this example, it is clear that a distogram band that includes larger numbers of large spatial outliers results in an increase of Local Moran's I values. The finding that a regular distribution of anomalies results in elevated Local Moran's I values suggests another potential application of multi-scale signatures of spatial dependence. It can provide an indicator of multi-scale texture, a different representation of quantifying the fractal dimension, with particular response to the degree of regularity of point-based perturbances across scales.

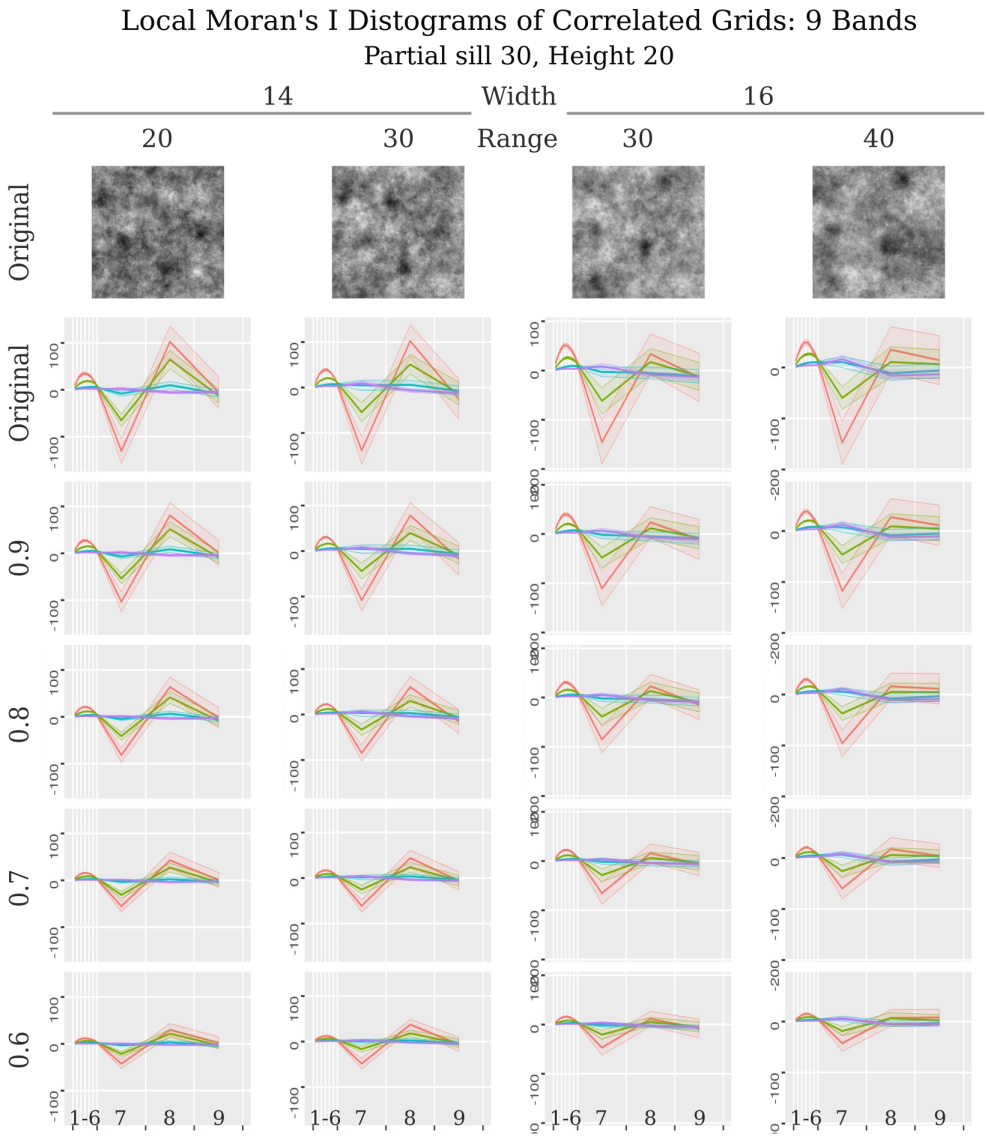


Figure 48: Local Moran's I distograms of selected correlated grids. Showing nine bands, up to a distance that could include an outlier, a gap, a neighboring outlier, and the background field. The top, side, base, and background classes are colored red, green, blue, and purple, respectively, each with a central average value and semi-transparent color out to the 95% confidence intervals. Note that the grayscale values vary for each mixed grid, but each ranges from lowest (white) to highest (black).

Since band 9 in all columns and band 8 in the right two columns exhibit confusion across all classes, band 8 in the left two columns is the most informative outer band re-

garding the properties of the correlated proxy variables. The following discussion refers to just band 8 in the left two columns. The distograms of the original data has minor overlap between the top and side class, but they are separate from the base and background. Although the linear patterns maintain relative shape and separation, as correlation decreases for the proxy variables the magnitudes of Local Moran's I values decrease. This is an artifact of the controlled simulation and the uniform application of Gaussian noise. It is notable, though, that even with noisy data, the Local Moran's I values are still rather large compared to the usually expected range of -1 to 1. Outliers that are just barely identifiable in the original data have distinctive distograms, which are maintained even in the presence of considerable noise.

Seven bands of the Local Moran's I distograms are shown in Figure 49. The distogram covers the scale of the outlier feature and the nearby neighborhood. The important pattern of contrasting Local Moran's I values, positive for inner bands and negative for band 7, is evident across distograms of all of the original and correlated grids. Although this set of examples involves arguably confusing regions with nearly imperceptible outliers, the estimated 95% confidence envelopes of the top and side distograms are separate in each of the distogram bands. It appears that the breadth of the confidence intervals are more responsive to the outlier width than to the changes in the field range parameter. Larger outliers result in larger confidence intervals. Interestingly, though, increasing the range values slightly increases separation between the top and side classes. Distogram magnitudes of the sides are slightly dampened and the top classes are slightly exaggerated. The confusion factor of larger patches in the background field apparently increases the local spatial autocorrelation of the nearby neighborhood, in band 7, of the side class. The nearby neighborhood of the top class responds to the larger field patches with a slight decrease in Local Moran's I values, perhaps indicating that the consistently

high local spatial autocorrelation of the outlier top region increasingly contrasts with variability in the field.

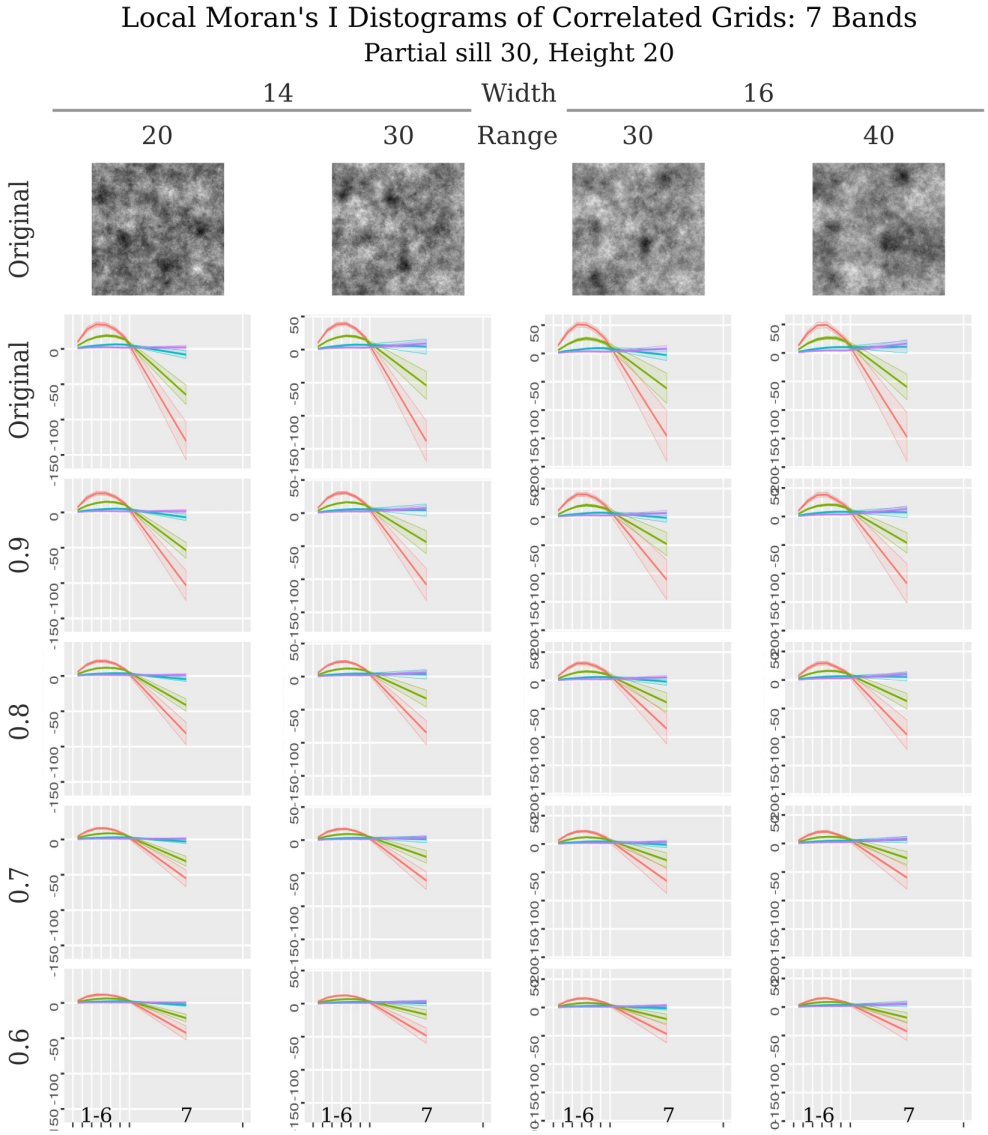


Figure 49: Local Moran's I distograms of selected correlated grids. Showing seven bands, up to a distance that could include an outlier and the imposed gap. The top, side, base, and background classes are colored red, green, blue, and purple, respectively, each with a central average value and semi-transparent color out to the 95% confidence intervals. Note that the grayscale values vary for each mixed grid, but each ranges from lowest (white) to highest (black).

Six bands of the Local Moran's I distograms are shown in Figure 50. The distogram matches the scale of the embedded outlier. The distograms of all classes are generally distinct across bands 1 though 6. The outlier base and background classes overlap in band 1 of the original grid. That overlap increases to perhaps the first three bands with the reduced Local Moran's I magnitudes related to decreasing correlation. There is also a slight overlap between the top and side classes in band 6. This occurs as they converge toward neutrality and switch from positive values in band 6 to negative values in band 7. It is notable that, across the scale of the outliers, the Local Moran's I distogram patterns are distinctive, especially considering that the low outliers are embedded in fields with patches of similar spatial extents and attribute variability. Even near the breakdown of distinction across large distances of bands 7, 8, and 9, there is quantified clarity between outlier part classes across the scale of the outliers.

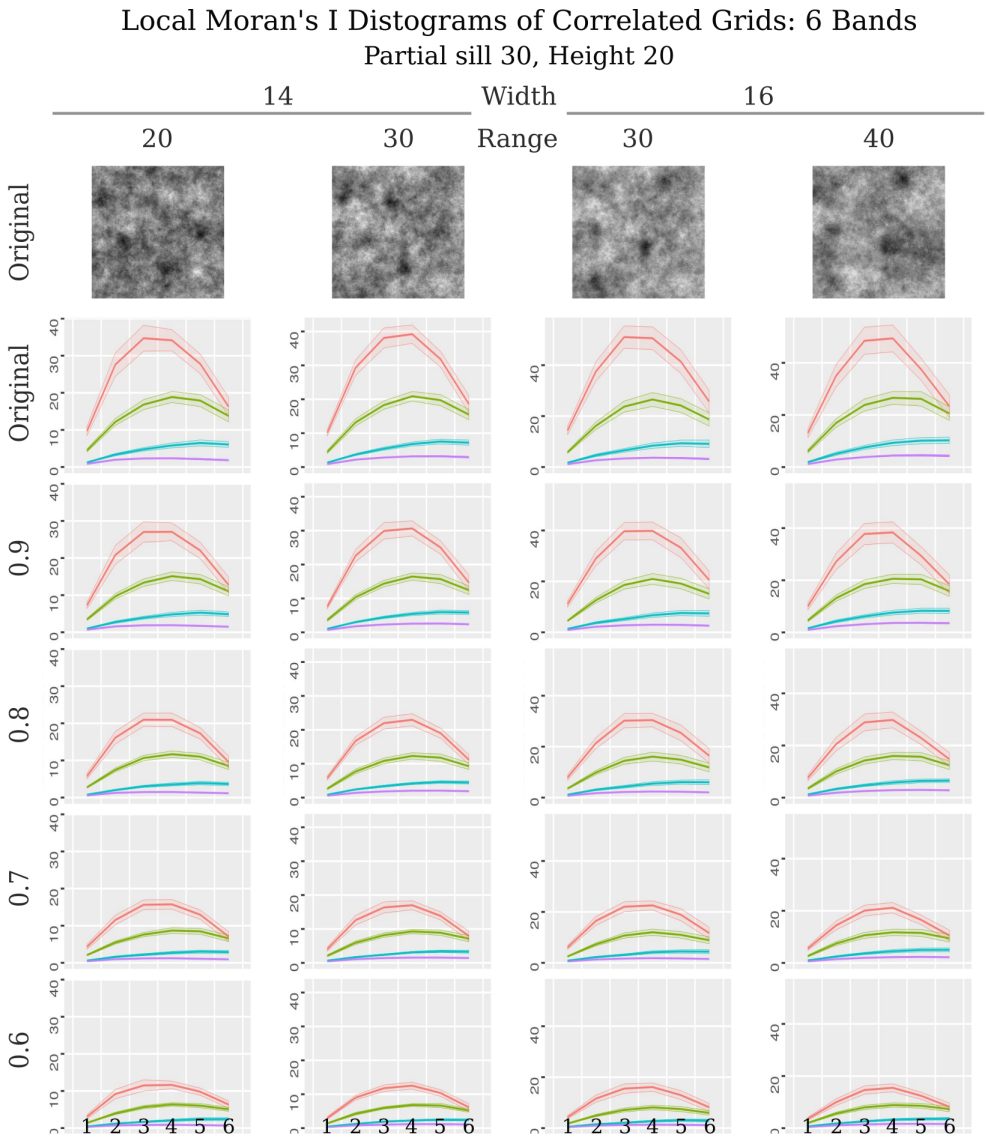


Figure 50: Local Moran's I distograms of selected correlated grids. Showing six bands, up to a distance that could include an outlier. The top, side, base, and background classes are colored red, green, blue, and purple, respectively, each with a central average value and semi-transparent color out to the 95% confidence intervals. Note that the grayscale values vary for each mixed grid, but each ranges from lowest (white) to highest (black).

5.3.10 Small samples

Throughout this experiment, only relatively small 1% random samples of the grids were employed. It is notable that, even with such limited information, distinctive multi-bandwidth Local Moran's I signatures for the outlier part classes were found. This even occurred with low target correlation proxy variables. Major patterns from the mixed grid might still be visible in a correlated grid, when all the values are viewed together. However, the obvious visible patterns of correlations might not be captured by a 1% random sample of points. The conditional Gaussian simulation introduces a salt-and-pepper noise effect into the correlated grids. In a small sample, noisy values have relatively large influence on any statistical summaries, compared to their minimal influence within the larger grid. This is especially an issue for discerning special characteristics of spatial outlier parts. The potential area sampled from each outlier part becomes increasingly smaller moving from the extensive background toward the key indicators of the outliers, the regions of the side and the top class, especially. It is, therefore, notable that the Local Moran's I distograms for the parts are distinctive with small samples through confusing scenarios with either the original or correlated data.

5.4 Conclusion

This chapter concerns an evaluation of characteristics of parts of wide spatial outliers. Similar to the third chapter, multiple simulations of sets of outliers embedded in a variable field were employed. In this case, the simulations enabled a controlled study of outlier parts. The top, side, and base parts of a wide outlier were defined by class boundaries at the first, second, and third standard deviations from the center of the Gaussian shape. For those parts, and extended distances into the background field, patterns of local spa-

tial autocorrelation were compared to ascertain whether each of the classes were distinct. The experiment also addressed two other considerations regarding a limited set of data for the spatial variable of interest: knowledge of only small samples and the potential of using correlated proxy variables.

A novel quantitative characteristic involving local spatial autocorrelation was calculated for each sampled location in every class for the variable of interest and for the proxies, separately. The characteristic, a multi-scale signature of spatial dependence, was obtained by assembling a series of Local Moran's I statistics calculated from sets of neighboring samples. The series of values, a Local Moran's I "distogram" was calculated from samples within one of several non-overlapping spatial bands defined by boundary distances from the point of analysis.

The results suggest that the top and side classes have distinctive patterns in the multi-scale signature of Local Moran's I values. The statistical characteristics of the base and background classes partially overlap. Boundary cases showing high contrast between the largest outliers and the background fields with the least variation give the best differentiation among the classes for obvious reasons. However, the inverse boundary cases with mixtures of small outliers in highly variable background fields maintain that differentiation. The differentiation is most evident in bands near to the point of analysis in the multi-bandwidth Local Moran's I signature, particularly those bands with distance boundaries about the scale, or size, of an outlier. All considered, the distinctive multi-scale patterns of spatial dependence are evident in challenging scenarios such as visually confusing data sets, low degrees of correlation, and with small samples.

6 Conclusion

This dissertation progressed through a sequence of investigations involving spatial data that represented a variable attribute. The main aim was to extract portions of data representing features, various types of local extrema. In the context of GIScience, and in order to reduce geographical uncertainty, this was an effort to address the impacts of spatial scale on the representation of features. The core of the problem is that any sampled data provide only limited information about real features. Nonetheless, it is from such limited sets of information that we describe, make inferences, and build our understanding about the geographical world.

Correspondence between real features and their representation in data is never exact. Conceptual models of features influence the way in which we interpret data. A conceptual model that involves continuous qualities, either smooth or textured, is an assumption about reality that can never be completely modeled by discrete spatially-referenced data values. The scale of analysis, or sampling interval, directly impacts the informative potential of spatial data. Leaving measurement errors aside, points of extrema are rarely sampled, aggregation methods change values, and the attribute variability across unsampled space is not represented.

Our conceptions of features, however, rely upon distinctive patterns of variability. Features have identifiable characteristics, such as the smoothness of a homogeneous region or a rapid transition near an anomalous point or along an edge. The best we can do with data is to identify subsets that represent a feature, to some degree, based upon attribute values and their relationships. This sub-setting, labeling, or extraction results in a simplified and inexact representation of real features. Regardless, representations

can provide a means of understanding how things in the natural world interact. For instance, flows across surfaces such as terrain have patterns of divergence and convergence related to local extrema of peaks, ridges, saddles, pits, and courses. Cartographic representations of those features provide approximate locations that can support predictions of flow directions and intensities.

In Chapter 2, however, representations of topographic features extracted from terrain data revealed issues with regard to the scale of analysis. Results were obtained separately by various algorithms and from different sets of data. The locations of features were dependent upon the extraction algorithms and each scale of analysis. Instead of arguing whether any method or scale was best, results of each method were combined into maps of multi-scale membership values for each feature class. An implication of the multi-scale fuzziness is that any location potentially exhibits characteristics of every feature class. The approach employed to resolve this issue was to compare multi-scale membership values, determine the dominant class, and compute the difference between the top two membership values as a measure of uncertainty. A more comprehensive solution could include all classes and a multi-method stage potentially extends this approach to combine semantics expressed in the conceptual models that are implemented in each of the extraction algorithms.

To isolate and perhaps clarify the effects of the scale of analysis, Chapter 3 presented a controlled study with synthetic data sets focused on only one type of the features, peaks, or spatial outliers. Performing spatial outlier detection methods on various combinations of outliers and scales of analysis was successful in finding the features in some cases but the results degraded in others. The unexpected yet informative results, clear upon visual review, were complex patterns of detected outliers, which rarely labeled the

centers of outliers. Scale-dependent changes in the representation of the spatial outlier apparently resulted in the presence of spatial dependence near the top of the otherwise anomalous feature. A mismatch between purpose of the detection algorithms and the representation of outlier with various sizes and with various raster resolutions provided empirical evidence of a known but important scale-based spatial outlier detection issue. As the scale of analysis transitions from too broad, to just right, and finally too fine, the spatial outlier is masked, detected, or it swamps across several data. The last scenario was considered most important to investigate. It is a common yet approachable problem in the sense that numerous spatial data sets contain variation that potentially represents features and parts of features that are discernibly different than their surroundings.

Since the previous tests involved only a single isolated outlier, Chapter 4 progressed to more complex scenarios of sets of various sized outliers embedded in variable fields. A benefit of the simulations was that the outlier shapes and locations were known, along with the statistical characteristics of the fields. This enabled tests of whether the characteristic of spatial dependence, relative to the variability of the field, occurs randomly in outliers. The experimental design compared results of various spatial sampling strategies: simple random, stratified random, regular, and a probabilistic method. A secondary finding was that, ranked in that order, the strategies returned increasing proportions of samples on outliers and unique outliers found. The probabilistic approach included information about locations with relatively high spatial dependence, as calculated with the Local Moran's I statistic. The results were clearly responsive to the inclusion of that quantifiable property. By applying the theory of spatial dependence to scenarios including anomalies broader than the scale of analysis, their locations were found. Visual review of intermediary results strongly suggested that the property of high spatial autocorrelation occurred in the central region of the outliers, near the top.

The final investigation, presented in Chapter 5, leveraged the theory of spatial dependence to model parts of spatial outliers. Supported by the previous evidence, it was hypothesized that a feature represented as a spatial outlier at one scale of analysis has, at finer scales, top, side, and base parts with identifiable patterns of local spatial autocorrelation. In one perspective, the locally extreme values of spatial outliers are anomalies. The property of spatial dependence, however, is a quantifiable means of characterizing the otherwise intuitive recognition of an anomaly as a feature, which is in turn made up of distinctive parts. The Local Moran's I "distogram" was proposed to represent a multi-scale signature of spatial dependence. The series of values in each bin of the distogram represented the Local Moran's I value calculated from neighboring data in adjacent, non-overlapping bands outward from each point of analysis.

The experiment included simulations of various outlier sizes and shapes embedded in variable fields. The variations of outlier and field parameters ranged from scenarios with clear anomalies to cases where the fields completely confused any visible identification of the embedded outlier. Although the model of the outlier involved a spatial decay function, the averaging effect of each distogram band potentially represents the change of value from an extreme point to the background as a transition. The approach, however, requires that the spatial scale of the distogram needs to match the scale of the feature of interest. In practice, to avoid a comprehensive calculation with many bands across large distances, a supervised approach that supplies a model feature of interest could aid the search for outliers of certain spatial and attribute characteristics.

Local Moran's I distograms of random points of analysis were labeled and aggregated relative to the outlier part or background region in which they were located. Regardless

of the amount of outlier to field confusion or the scale of analysis, distinctive patterns of spatial dependence were evident for parts of spatial outliers. Compared to the background field, the tops of outliers had the most pronounced differences followed, similarly, by the sides. Both of those two classes exhibited distinctively high local spatial autocorrelation across the distance of an outlier radius, which contrasted with low spatial autocorrelation of the nearby surroundings. The base, the outlier part with least influence of the tops, had the least difference and was often indistinct from the background. The results were sensitive to the spatial scale and the attribute variability of both the outliers and the fields.

An important aspect of the proposed approach is that it performed well in challenging scenarios. As opposed to the high information potential of a dense, regular grid, the results were clear even with a smaller set of random samples. The separation of Local Moran's I distograms for the top and side classes also remained distinctive across data sets with substantially degrading degrees of correlation. This is beneficial for a situation in which the variable of interest is highly confused or not available. A correlated ancillary variable has the potential of guiding a search for anomalies in the variable of interest. A related consideration is that the correspondence between two variables might be subject to a change in scale, as well. A peak in one variable might be represented as a wide spatial outlier in another, with a spatial decay function that extends farther into the distance.

The traditional conceptual model of features represented as spatial outliers is scale dependent. The assumption that there is a single extreme value compared to the neighborhood is a representation with a sampling interval that perfectly fits the size of the outlier. Based upon the findings of this dissertation, at least a top region and a side region are valid parts. An option would be to further specify the conceptual model to

include five parts (and a background, for comparison): a top point, a top region, a minor top-to-side transition, a major side transition, a minor side-to-background transition, and the background.

Several extensions to this research would confront issues with regard to spatial data types and the complex patterns of real features. With a goal of returning to the initial subject of real topographic data, a first step would be to continue the progression of developments with synthetic data. For example, the performance of the models of multi-scale signatures of spatial dependence could be tested. The accuracies of classifying random points as feature parts could be tested on features and fields of known properties. Shapes of outliers could be modified to include vertical profiles that are more pointed or flat topped. This could inform tests with data of real features, such as topographic peaks or more flat-topped mesas. Examination of horizontal, plan-view shapes, could also extend to involve various outlier sizes simultaneously and to include outliers exhibiting anisotropic decay, or irregular linear and perhaps braided patterns. Land cover or soil reflectivity are immediate candidates for extending the consideration of variables correlated with topography.

Considering spatial data types introduces further complexities, by which multi-scale aspects are not only with regard to the spatial scales, but also the scales of attribute values. An extension to include three-dimensional (3D) space is one that would serve the needs of numerous recent technological developments. Generalities of broad and fine scale structures in 3D space are as relevant as those in 2D space. However, the scales might need to be adjusted independently, perhaps by normalization. Such independent scaling of multidimensional values might also be addressed by employing Mahalanobis distances across any set of spatially referenced attribute variable values. Another concern

could be to address anomalous values represented by scales of measurement. Challenges exist with regard to the contrast between a feature and the background if variables are represented with decreasing quantifiable information. The interval and ordinal scales of measurement might still be possible with modifications for the level of information involved. It is unclear whether patterns would be as distinctive with categorical data.

The investigations and approaches developed in this dissertation engage with issues of uncertainty in geographical representation as influenced by scale. In essence, the aim was to address various aspects of limited geographical information. The true character of a spatial phenomena is not known with a limited set of spatial data. Where informed by a sample, only estimations are possible. The representation of geographical features, defined by patterns of variation, is impeded by a mismatch between the size of the feature and the scale of analysis. However, the theory of spatial dependence continues to show potential for alleviating such problems, even those involving extreme variation. Attribute characteristics at any location are enhanced by considering relationships across neighborhoods of varying extents. In this case, conflicting scale-dependent results were addressed by developing multi-scale representations. In Chapter 2, a comparison of multi-scale membership values for each surface network feature class revealed that the clearest features were found in high variability terrain. In Chapters 3 and 4, empirical evidence showed that wide outliers, with a nonrandom presence of local spatial autocorrelation, are not well detected by common methods, which particularly miss the tops. In Chapter 5, the Local Moran's I distogram is presented as a multi-scale signature of spatial dependence. This signature established distinctive multi-scale patterns of local spatial autocorrelation for the top and side parts of spatial outliers, even in confusing scenarios of background variability. Combining information from separate scales together, raises the level of information as a composite characteristic. Interpreting multi-scale representa-

tions is more challenging than those of a single scale. However, they are informative about the stability of attribute values around observations. Such representations of spatial data further specialize the identification and extraction of features. What are either missed or noise at some scales are revealed as features with special characteristics at other scales.

References

Achtert, E., Kriegel, H.-P., Schubert, E., & Zimek, A. (2013). Interactive data mining with 3D-parallel-coordinate-trees. In *Proceedings of the Association of Computing Machinery International Conference on Management of Data (SIGMOD)* (p. 1009-1012). New York City, NY.

Aggarwal, C. C. (2013). *Outlier analysis* (1st ed.). New York: Springer-Verlag. Retrieved from https://doi.org/10.1007/978-1-4614-6396-2

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical Analysis*, *27*(2), 93-115. Retrieved from https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Arundel, S., Phillips, L., Lowe, A., Bobinmyer, J., Mantey, K., Dunn, C., ... Usery, E. (2015). Preparing The National Map for the 3D elevation program - products, process and research. *Cartography and Geographic Information Science*, *42*(sup1), 40–53. Retrieved from https://doi.org/10.1080/15230406.2015.1057229

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester: Wiley.

Berry, B., & Baker, A. (1968). Geographic sampling. In B. Berry & D. Marble (Eds.), *Spatial analysis* (pp. 91–100). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Bone, C., Wulder, M. A., White, J. C., Robertson, C., & Nelson, T. A. (2013). A GIS-based risk rating of forest insect outbreaks using aerial overview surveys and the Local Moran's I statistic. *Applied Geography*, *40*, 161 - 170. Retrieved from https://doi.org/10.1016/j.apgeog.2013.02.011

Burrough, P., van Gaans, P., & Hootsmans, R. (1997). Continuous classification in soil survey: Spatial correlation, confusion and boundaries. *Geoderma*, *77*(2-4), 115-135. Retrieved from https://doi.org/10.1016/s0016-7061(97)00018-9

Cayley, A. (1859). On contour and slope lines. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *XVIII*, 264-268. Retrieved from https://doi.org/10.1017/cbo9780511703706.025

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: a survey. *Association for Computing Machinery, Computing Surveys*, *41*(3), 1-58. Retrieved from https://doi.org/10.1145/1541880.1541882

Chawla, S., & Sun, P. (2005). SLOM: a new measure for local spatial outliers. *Knowledge and Information Systems*, *9*(4), 412-429. Retrieved from https://doi.org/10.1007/s10115-005-0200-2

Cheng, T., Fisher, P. F., & Li, Z. (2004). Double vagueness: Uncertainty in multi-scale fuzzy assignment of duneness. *Geo-spatial Information Science*, *7*(1), 58-66. Retrieved from https://doi.org/10.1007/bf02826677

Clarke, K. C., & Romero, B. E. (2017). On the topology of topography: a review. *Cartography and Geographic Information Science*, *44*(3), 271-282. Retrieved from https://doi.org/10.1080/15230406.2016.1164625

Couclelis, H. (1992). People manipulate objects (but cultivate fields): Beyond the raster-vector debate in GIS. *Lecture Notes in Computer Science*, 65-77. Retrieved from

https://doi.org/10.1007/3-540-55966-3_3

Couclelis, H. (2010). Ontologies of geographic information. *International Journal of Geographical Information Science*, *24*(12), 1785-1809. Retrieved from https://doi.org/10.1080/13658816.2010.484392

Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data.* Hoboken, NJ: Wiley-Blackwell.

Drăguţ, L., & Blaschke, T. (2008). Terrain segmentation and classification using SRTM data. In Q. Zhou, B. Lees, & G.-a. Tang (Eds.), *Advances in digital terrain analysis* (pp. 141–158). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-540-77800-4_8

ESRI. (2012). *ArcGIS desktop, version 10.1.* Redlands, CA: Environmental Systems Research Institute. Retrieved from http://www.esri.com/software/arcgis (Accessed 28 June 2016)

Evans, I., Hengl, T., & Gorsevski, P. (2009). Applications in geomorphology. In T. Hengl & H. I. Reuter (Eds.), *Geomorphometry: Concepts, software, applications* (Vol. 33, p. 497 - 525). Amsterdam, The Netherlands: Elsevier. Retrieved from https://doi.org/10.1016/S0166-2481(08)00022-6

Evans, I. S. (2012). Geomorphometry and landform mapping: What is a landform? *Geomorphology*, *137*(1), 94-106. Retrieved from https://doi.org/10.1016/j.geomorph.2010.09.029

Fisher, P. F., Wood, J., & Cheng, T. (2004). Where is Helvellyn? Fuzziness of multi-scale landscape morphometry. *Transactions of the Institute of British Geographers*, *29*(1), 106-128. Retrieved from https://doi.org/10.1111/j.0020-2754.2004.00117.x

Fowler, R. J., & Little, J. J. (1979). Automatic extraction of irregular network digital terrain models. *ACM SIGGRAPH Computer Graphics*, *13*(2), 199-207. Retrieved from https://doi.org/10.1145/965103.807444

GDAL. (2014). *GDAL - Geospatial Data Abstraction Library, version 1.11.* Retrieved from http://www.gdal.org (Accessed 28 June 2016)

Gerçek, D., Toprak, V., & Strobl, J. (2011). Object-based classification of landforms based on their local geometry and geomorphometric context. *International Journal of Geographical Information Science*, *25*(6), 1011-1023. Retrieved from https://doi.org/10.1080/13658816.2011.558845

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, *24*(3), 189-206. Retrieved from https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

Gómez-Hernández, J. J., & Journel, A. G. (1993). Joint sequential simulation of multigaussian fields. In A. Soares (Ed.), *Geostatistics troia '92* (Vol. 1, p. 85-94). Springer, Dordrecht. Retrieved from https://doi.org/10.1007/978-94-011-1739-5_8

Goodchild, M. F. (1992). Geographical information science. *International Journal of Geographical Information Systems*, *6*(1), 31-45. Retrieved from https://doi.org/

10.1080/02693799208901893

GRASS Development Team. (2012). *Geographic Resources Analysis Support System (GRASS), software, version 6.4.3.* Open Source Geospatial Foundation Project. Retrieved from http://grass.osgeo.org (Accessed 28 June 2016)

Kedron, P. (2016). Identifying the geographic extent of environmental inequalities: a comparison of pattern detection methods. *The Canadian Geographer / Le Géographe Canadien*, *60*(4), 479-492. Retrieved from https://doi.org/10.1111/cag.12297

Lu, C.-T., Chen, D., & Kou, Y. (2004). Multivariate spatial outlier detection. *International Journal on Artificial Intelligence Tools*, *13*(04), 801-811. Retrieved from https://doi.org/10.1142/s021821300400182x

MacMillan, R. A., Pettapiece, W. W., Nolan, S. C., & Goddard, T. W. (2000). A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. *Fuzzy Sets and Systems*, *113*(1), 81-109. Retrieved from https://doi.org/10.1016/S0165-0114(99)00014-7

MapWindow Open Source Team. (2008). *Mapwindow GIS open source project, version 4.8.6.* Retrieved from http://www.mapwindow.org (Accessed 28 June 2016)

Mark, D. (1977). *Topological randomness of geographic surfaces of geomorphic surfaces* (Technical Report No. 15). Office of Naval Research. (Contract N00014-75-C0886)

Mark, D., & Sinha, G. (2012). Surface networks and the ontology of topography? In *GeoVocamp 2012 / 5th annual SOCoP workshop.* Reston, Virginia, USA.

Mark, D. M., & Turk, A. G. (2003). *Ethnophysiography.* (Paper presented at Workshop on Spatial and Geographic Ontologies (prior to COSIT03))

Mark, D. M., & Turk, A. G. (2017). Ethnophysiography. In D. Richardson, N. Castree, M. Goodchild, A. Kobayashi, W. Liu, & R. Marston (Eds.), *International encyclopedia of geography: People, the earth, environment and technology.* Hoboken, NJ: John Wiley & Sons, Ltd. Retrieved from https://doi.org/10.1002/9781118786352.wbieg0349

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, *58*(8), 1246-1266. Retrieved from https://doi.org/10.2113/gsecongeo.58.8.1246

Maxwell, J. (1870). On contour lines and measurements of heights. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, *40*, 421-427.

McMaster, R. B., & Usery, E. L. (Eds.). (2005). *A research agenda for geographic information science.* Boca Raton, FL: CRC Press.

Miliaresis, G. C. (2008). Quantification of terrain processes. In Q. Zhou, B. Lees, & G.-a. Tang (Eds.), *Advances in digital terrain analysis* (pp. 13–28). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-540-77800-4_2

Minár, J., & Evans, I. S. (2008). Elementary forms for land surface segmentation: the theoretical basis of terrain analysis and geomorphological mapping. *Geomorphology*, *95*(3-4), 236-259. Retrieved from https://doi.org/10.1016/j.geomorph.2007.06.003

Montello, D. R. (2005). Navigation. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (p. 257-294). New York, NY: Cambridge University Press.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1-2), 17-23. Retrieved from https://doi.org/10.1093/biomet/37.1-2.17

Morse, M. (1925). Relations between the critical points of a real function of n independent variables. *Transactions of the American Mathematical Society*, *27*(3), 345-396. Retrieved from https://doi.org/10.2307/1989110

National Research Council. (2006). *Beyond mapping: Meeting national needs through enhanced geographic information science.* Washington, DC: The National Academies Press. Retrieved from https://doi.org/10.17226/11687

National Research Council. (2010). *New research directions for the national geospatial-intelligence agency: Workshop report.* Washington, DC: The National Academies Press. Retrieved from https://doi.org/10.17226/12964

NetworkX Developers. (2014). *NetworkX, a Python language software package, version 1.9.1.* Retrieved from http://networkx.github.io (Accessed 28 June 2016)

Nyquist, H. (1928). Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, *47*(2), 617-644. Retrieved from https://doi.org/10.1109/t-aiee.1928.5055024

O'Callaghan, J. F., & Mark, D. M. (1984). The extraction of drainage networks from digital elevation data. *Computer Vision, Graphics, and Image Processing*, *27*(2), 247. Retrieved from https://doi.org/10.1016/s0734-189x(84)80047-x

OpenTopography. (2010). *2010 channel islands lidar collection.* Dewberry & Davis and Terrapoint Aerial Services for the USGS under the Channel Islands ARRA LiDAR Task Order. Retrieved from https://doi.org/10.5069/G95D8PS7 (OpenTopography Collection ID: OT.082012.26911.1; Accessed 28 April 2017)

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, *30*(7), 683-691. Retrieved from https://doi.org/10.1016/j.cageo.2004.03.012

Peuquet, D. J. (1988). Representations of geographic space: Toward a conceptual synthesis. *Annals of the Association of American Geographers*, *78*(3), 375-394. Retrieved from https://doi.org/10.1111/j.1467-8306.1988.tb00214.x

Pfaltz, J. L. (1976). Surface networks. *Geographical Analysis*, *8*(1), 77-93. Retrieved from https://doi.org/10.1111/j.1538-4632.1976.tb00530.x

R Development Core Team. (2013). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org (Accessed 28 June 2016)

Rana, S. (Ed.). (2004). *Topological data structures for surfaces.* West Sussex, England: John Wiley & Sons, Ltd. Retrieved from https://doi.org/10.1002/0470020288

Reeb, G. (1946). Sur les points singuliers d'une forme de pfaff completement intergrableou d'une fonction numerique. *Comptes Rendus de l'Académie des Sciences*, *222*, 847-

849.

Reech, M. (1858). Demonstration d'une propriété générale des surfaces fermées. *Journal de l'Ecole Polytechnique*, *37*, 169-178.

Robinson, A. H., Morrison, J. L., Muehrcke, P. C., Kimerling, A. J., & Guptill, S. C. (1995). *Elements of cartography* (6th ed.). New York, NY: John Wiley and Sons.

Rücker, G., & Schwarzer, G. (2014). Presenting simulation results in a nested loop plot. *BMC Medical Research Methodology*, *14*(1). Retrieved from `https://doi.org/10.1186/1471-2288-14-129`

SAGA Development Team and User Group Association, Bock, B. J., M., Conrad, O., Koethe, R., & Ringeler, A. (2008). *System for automated geoscientific analyses, version 2.0.8.* Retrieved from `http://www.saga-gis.org` (Accessed 28 June 2016)

Schmidt, J., & Hewitt, A. (2004). Fuzzy land element classification from DTMs based on geometry and terrain position. *Geoderma*, *121*(3-4), 243-256. Retrieved from `https://doi.org/10.1016/j.geoderma.2003.10.008`

Shannon, C. (1949). Communication in the presence of noise. *Proceedings of the IRE*, *37*(1), 10-21. Retrieved from `https://doi.org/10.1109/jrproc.1949.232969`

Shekhar, S., Lu, C.-T., & Zhang, P. (2003). A unified approach to detecting spatial outliers. *GeoInformatica*, *7*(2), 139-166. Retrieved from `https://doi.org/10.1023/A:1023455925009`

Shi, X., Zhu, A.-X., & Wang, R. (2005). Fuzzy representation of special terrain features using a similarity-based approach. In F. E. Petry, V. B. Robinson, & M. A. Cobb (Eds.), *Fuzzy modeling with spatial information for geographic problems* (pp. 233–251). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from `https://doi.org/10.1007/3-540-26886-3_11`

Sinha, G., Kolas, D., Mark, D., Romero, B., Usery, G., L.E.Berg-Cross, & Padmanabhan, A. (n.d.). Surface network ontology design patterns for linked topographic data. *Semantic Web*. Retrieved from `http://www.semantic-web-journal.net/system/files/swj675.pdf` (Under revision for resubmission)

Sokal, R. R., Oden, N. L., & Thomson, B. A. (2008). Local Spatial Autocorrelation in Biological Variables. *Biological Journal of the Linnean Society*, *65*(1), 41-62. Retrieved from `https://doi.org/10.1111/j.1095-8312.1998.tb00350.x`

Stepinski, T., & Jasiewicz, J. (2011). Geomorphons - a new approach to classification of landforms. In E. I. W. J. Hengl T. & M. Gould (Eds.), *Proceedings of geomorphometry 2011* (p. 109-112). Redlands, CA, USA.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680. Retrieved from `https://doi.org/10.1126/science.103.2684.677`

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(2), 234. Retrieved from `https://doi.org/10.2307/143141`

Usery, E. (1996). Membership functions for fuzzy set representation of geographic features. In *Proceedings, International Society for Photogrammetry and Remote*

Sensing International Congress, International Archives of Photogrammetry (Vols. XXXI, Part B4, p. 881-883). Vienna, Austria.

USGS, The National Map. (2016). *3DEP products and services: The National Map, 3D Elevation Program.* Retrieved from `http://nationalmap.gov/3DEP/3dep_prodserv.html` (Accessed 28 April 2017)

van Rossum, e. a., G. (2010). *The Python Language Reference, Python Software Foundation, version 2.7.* Retrieved from `https://docs.python.org/2.7/reference/index.html` (Accessed 28 June 2016)

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer-Verlag. Retrieved from `https://doi.org/10.1007/978-0-387-21706-2`

Warntz, W. (1966). The topology of a socio-economic terrain and spatial flows. *Papers in Regional Science*, *17*(1), 47-61. Retrieved from `https://doi.org/10.1111/j.1435-5597.1966.tb01341.x`

Warntz, W., & Woldenberg, M. (1967). *Concepts and applications: Spatial order* (Technical Report No. 1). Cambridge, Mass.: Harvard University. (Project 389-147)

Wood, J. (1996). *The geomorphological characterisation of digital elevation models* (Ph.D. Thesis). University of Leicester, UK.

Wood, J. (2007). *LandSerf, version 2.3.* Retrieved from `http://www.landserf.org` (Accessed 28 June 2016)

Yuan, Y., Cave, M., & Zhang, C. (2018). Using Local Moran's I to identify contamination hotspots of rare earth elements in urban soils of london. *Applied Geochemistry*, *88*, 167-178. Retrieved from `https://doi.org/10.1016/j.apgeochem.2017.07.011` (SI: ISEG 2016)

Zhang, C., Luo, L., Xu, W., & Ledwith, V. (2008). Use of Local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. *Science of The Total Environment*, *398*(1), 212 - 221. Retrieved from `https://doi.org/10.1016/j.scitotenv.2008.03.011`

Zhang, C., & McGrath, D. (2004). Geostatistical and GIS analyses on soil organic carbon concentrations in grassland of southeastern ireland from two different periods. *Geoderma*, *119*(3), 261 - 275. Retrieved from `https://doi.org/10.1016/j.geoderma.2003.08.004`

Zhang, J., & Goodchild, M. F. (2002). *Uncertainty in geographical information.* Taylor & Francis. Retrieved from `https://doi.org/10.4324/9780203471326`