PARCEL LEVEL AGRICULTURAL LAND COVER PREDICTION

by

Jonathan Lisic
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Computational Science and Informatics

Committee:

_____	Dr. James Gentle, Dissertation Director
_____	Dr. Edward Wegman, Committee Member
_____	Dr. Thomas Wanner, Committee Member
_____	Dr. Carol Crawford, Committee Member
_____	Dr. Kevin Curtin, Acting Department Chair
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science

Date: _____ Fall Semester 2015
George Mason University
Fairfax, VA

Parcel Level Agricultural Land Cover Prediction

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Jonathan Lisic
Master of Science
University of Akron, 2005
Bachelor of Arts
Western Washington University, 2003

Director: Dr. James Gentle, Professor
Department of Computational and Data Sciences

Fall Semester 2015
George Mason University
Fairfax, VA

Dedication

I dedicate this dissertation to my mother, Wendy Lisic (1949-2014).

Acknowledgments

I would first like to thank my advisor, Dr. James Gentle, who provided guidance and focus to this dissertation. This dissertation would figuratively and literally not have been possible without his help. I greatly enjoyed the arguments we have had over this work, and the pleasant, usually off topic, discussion about computing. I would also like to thank his friendship and mentorship throughout the eight years of my time at George Mason University.

I would like to thank the members of my committee: Dr. Edward Wegman, who's expert knowledge and comments of the dissertation proposal helped guide this research; Dr. Thomas Wanner, who's excellent presentation of numeric analysis in prior courses has helped invaluably through this research; Dr. Kirk Borne, who's good humor, energy, and dependability have made both the dissertation and prior courses enjoyable; Dr. Carol Crawford, who's expert knowledge, humor, and energy has helped me greatly in both understanding the material in this dissertation, and provided a number of excellent ideas on this research.

I would like to thank my family: My wife, Ling, and daughter Elizabeth who have graciously allowed me to have time at night and during the weekend to work on my dissertation, and who's smiles have provided motivation even in the most challenging of times; My father, James, and mother, Wendy, who have supported my education throughout the years, and have also shown an interest in my pursuits; My brother and sisters, Sebastian, Kathryn, and Amanda, that have encouraged me to finish my dissertation.

I would like to thank my colleagues at the U.S. Department of Agriculture's National Agricultural Statistical Service, Research and Development Division (RDD). In particular I would like to thank: Wendy Barboza, who has been a great friend and provided me the opportunity to work as a research statistician; the RDD Director Dr. Linda Young, who has provided both mentorship and guidance as a statistician; the late Dr. Joseph Kosler, who was a great friend and motivation in the early stages of this research; my supervisor Kay Turner, who has allowed me to have a flexible working arrangement to balance work, life, and dissertation; Dr. Matt Williams, who has provided numerous insights and suggestions to this research; Dr. Nathan Cruze, who has been a great friend and made a number of useful suggestions; Darcy Miller, who has helped me edit the final draft of this dissertation, and has been a great friend; Jason Bell, who has been a great friend and supporter of my work; Claire Boryan, who has provided me both motivation and assistance in areas of GIS and geography; Dr. Zhengwei Yang, who cheerfully made many suggestions, and fixed the cropscape CDL server after my excessive use.

I would also like to thank my prior Colleagues at the Bureau of Labor Statistics; notably, Dee Zamora, Jackson Crockett, Adam Issan, and Brad Rhein. I would also like to thank all that have given feedback on this research, including Dr. Clifford Spiegelman (Texas A&M) and Dr. Balgobin Nandram (WPI), that both provided input into the Joint Statistical Meetings 2015 proceedings and presentation of this material.

Table of Contents

List of Tables

List of Figures

Abstract

PARCEL LEVEL AGRICULTURAL LAND COVER PREDICTION

Jonathan Lisic, PhD

George Mason University, 2015

Dissertation Director: Dr. James Gentle

The purpose of this research is to develop and study the methodology and the underlying theory for prediction of agricultural land cover for a set of commercial crops at the parcel level. The observational unit will be called a *land cover unit* (LCU). Each LCU will have an associated *land cover sequence*. A land cover sequence is an ordered set of known categories indexed by a set of fixed consecutive years. An LCU is the maximally contiguous section of land with respect to a single land cover sequence, not transected by public transportation arteries or permanent hydrographic boundaries.

LCUs are not observed, instead they are estimated through an image segmentation algorithm known as mean shift, applied to high resolution imagery products. The predictors of land cover are constructed under the assumption of temporal stationarity; this assumption limits the length of the land cover sequence that can be used to aid in the prediction. Land cover sequences of the LCUs are estimated from classified pixel level data. Prediction of future land cover is performed through a Bayesian hierarchical multinomial probit model accounting for spatially correlated crop rotation preferences.

Application to major commercial crops in La Porte County, Indiana, is provided using high resolution imagery and thematic maps from the United States Department of Agriculture. The theory and methods are applicable to prediction of agricultural crops in other

areas with a relatively stable pattern of agricultural land cover.

Chapter 1: Introduction

The purpose of this research is to develop and study the methodology and the underlying theory for prediction of agricultural land cover for a set of commercial crops at the parcel level. The term "parcel level" in Geospatial Information Systems (GIS) literature and this dissertation, is the geographic scale where individual commercial agricultural fields can be identified. Prediction of land cover at the parcel, or agricultural field level is useful for many purposes including: agricultural production; natural resource management (see Thenail et al., 2009); survey development (see Zimmer et al., 2012); predicting and measuring changes in the local environment (see Castellazzi et al., 2007); residential land conversion; and land cover response to local market changes such as the establishment of ethanol plants (see Livingston et al., 2008 and Livingston et al., 2012). In this dissertation, an approach to predicting future land cover through an areal spatial-temporal autoregressive probit model is presented. This approach relies on image segmentation to construct parcel-like units from high-resolution imagery, and an agricultural production process known as *crop rotations*.

Given a set of parcels the prediction of future land cover can be facilitated by the agricultural practice of crop rotations. Crop rotations are a repeating sequence of crops used to promote yield while retaining soil quality and mitigating pests and disease. Common rotations in the United States include corn-to-soybeans-to-corn which has been studied by Livingston et al. (2012) and others. In the context of parcel-based models, crop rotations are especially useful since they are a prevalent locally observable phenomenon. The prevalence of these rotations was studied by Sahajpal et al. (2014) where it was noted that the U.S. cropping patterns could be represented with 90% accuracy by just 82 distinct three-year sequences.

Modeling agriculture via crop rotations is certainly not novel, and has seen employment in simulations and agricultural econometric models. Simulation models have been used by

Schönhart et al. (2011), Castellazzi et al. (2008), and Detlefsen (2004) utilizing stochastic matrices to model crop rotations. These models are primarily used for simulating land cover, for long term prediction, based on assumed crop rotations. Estimation of transition parameters in these papers is lacking or non-existent; due to the simple structure of the stochastic matrices it is not possible to directly admit covariates.

The framework of agricultural economic models of crop rotations has been explored at great length by Hennessy (2006). Application to optimal crop rotations of corn and soybeans has been explored by Livingston et al. (2012), and a more general framework has been considered by Cai et al. (2013). Simplifications to Hennessy (2006) has been considered by Ji and Rabotyagov (2015). This framework is based on expected return, which cannot be practically applied at a parcel level due to return being a function of yield which is highly volatile even at the county level for many crops. Since economic models cannot be practically applied at the parcel level, they will not be considered in this dissertation.

The most similar approach to mine in the literature is the graphical model based approach by Osman et al. (2015). In this approach, however, the spatial units are already provided through the French Registre Parcellaire Graphique Land Parcel Identification System (RPG LPIS). Since parcels provided by RPG LPIS may include multiple land uses, the authors chose to restrict the modeling units to the subset of parcels that contain only a single crop and field, potentially introducing bias. Prediction in this model is through a graphical modeling approach known as Markov logic network (MLN) (see Richardson and Domingos, 2006). A Markov logic network is similar to other approaches to learning networks such as a Bayesian network (BN), with the exception that the graphs are not directional. In the MLN each node is a random variable with the Markov property, conditionally independent of other nodes in the graph given its neighbors. This model is non-spatial, does not readily accept continuous exogenous information, and is dependent on the existence of classified parcels. The model I developed does not have these shortcomings.

The approach in this dissertation resolves the issues found in Osman et al. (2015). First, parcel-like land cover units (LCUs) are constructed using high-resolution imagery. These

units contain a single land cover avoiding the exclusion of units due to multiple land uses in Osman et al. (2015). A spatial-temporal multinomial probit model is used to model crop rotations. This model provides a straightforward way to add covariates, incorporate the spatial dependence of LCUs, and produce statistically meaningful parameter estimates. This approach breaks the land cover prediction into three tasks:

1. Image Segmentation - to estimate LCU boundaries from high resolution imagery through the mean shift algorithm.

2. Classification - to classify LCU estimates based on coarse pre-classified pixels and other exogenous data sources.

3. Prediction - to predict LCU estimates' land cover contents using crop rotations through a spatially auto-regressive process.

Through this dissertation each of these three tasks is addressed separately.

Application to the agricultural land in La Porte County, Indiana, is provided in this dissertation. The methods presented in part or full can be employed anywhere high resolution remotely sensed images and associated land cover classification results are available, not necessarily sharing the same resolution. These data requirements are less onerous year-by-year given the prevalence of high resolution images from companies such as Google, and freely available global Landsat data (see Yan and Roy, 2014), with a variety of methods to use this data to classify land cover (see Boryan et al., 2011).

1.1 Image Segmentation

Image segmentation is the process of defining a mapping between a set of pixels in a computer-generated image to a set of indexed partitions. This problem exists in many fields, particularly in remote sensing and computer vision. The earliest published methods date back to 1965 (see Zhang et al., 2008). The number of algorithms to perform this mapping has significantly grown since 1965 with over 3000 published journals articles and

books on the subject as of 2010, many being application specific (see Dey et al., 2010). It is noted by numerous authors that there is no general solution to image segmentation, and that it is application and data specific.

The specific domain of the problem being addressed in this dissertation is simultaneous segmentation of multiple images of different resolutions and type (categorical or continuous). In computer vision literature, the analysis of categorical images is completely absent. This absence is to be expected, given that computer vision focuses on digital images that are represented as either binary or continuous data sets. Computer vision literature does cover multiple images through multi-resolution and multi-view image segmentation, primarily for modeling 3-D objects using multiple images or to identify objects moving in images (video).

In agricultural and land cover applications, image segmentation often coincides with classification, where classification is a method of assigning a meaningful class to a pixel (see Wilkinson, 1999). Pixel classification is outside of the scope of this research; instead, it is assumed that a set of classified pixels with the desired classes already exist. Classified pixels for the 48 contiguous states in the United States are available through the United States Department of Agriculture's Cropland Data Layer (CDL) (see Boryan et al., 2011). The CDL is used in application of the methods in this paper. Classification of segments is within the scope of this research and is covered in the classification section of this dissertation.

In agriculture, photographic and satellite imagery such as NAIP (National Agricultural Imagery Program) and Landsat provide spectral intensity over multiple bands, gradient and texture properties. Segmentation of these images has been entertained numerous times in the literature to form parcel-like units. The most relevant paper to the image segmentation portion of this dissertation is Yan and Roy (2014). In Yan and Roy (2014), remotely sensed images are used to perform extraction (segmentation and identification) of crop fields. Crop fields are regions of land used to grow crops bounded by roads or other land cover; this spatial unit assumes that crops are reasonably static but further details are not provided. Other notable work includes Zhang (2009), where wavelet transforms were used with watershed segmentation to identify agricultural images. Both of these works rely on

the use of watershed segmentation (see Beucher and Lantuéjoul, 1979), but differ on data sources. Yan and Roy (2014) use a function of spectral bands over hundreds of daily $30m^2$ images, and Zhang (2009) uses a single high resolution $2.4m^2$ image.

There are a number of problems with these approaches. First, is that these approaches are non-stochastic, and little care is given to the data generating mechanism. In both papers, no distribution assumptions are made, and there is a heavy reliance on heuristics. In this dissertation, a similar but more statistically well defined approach is taken through a kernel density estimation based classification method known as mean shift.

Mean shift has been used before in remote sensing, for example by Huang and Zhang (2008), Büschenfeld and Ostermann (2012), and Friedman et al. (2013). The application to large scale areas or high resolution imagery is significantly impacted by the computational cost of the algorithm. The mean shift algorithm and other clustering algorithms also have issues with high variance structures such as trees and similar vegetation prevalent in agriculture. In this dissertation, six contributions are made to improve the quality of the segmentation and to decrease the computation cost of the mean shift algorithm:

1. A novel well defined spatial-temporal land cover unit;

2. A novel approach to separating edge detection from high variance structures through local variances;

3. A novel combination of the mean shift fixed point iterator and Newton's method fixed point iterator under a Gaussian kernel;

4. A novel implementation of the mean shift algorithm using the observed property that the mean shift sequences tend to merge into a few distinct paths approaching a stationary point;

5. A novel image stratification and sampling method for mean shift segmentation of spatial images;

6. An improvement to the existing local pivotal method used for spatially balanced

sampling (see Grafström et al., 2012), reducing the computational complexity from $\mathcal{O}\left(n^2\right)$ to $\mathcal{O}\left(\log(n)n\right)$.

An application of the accelerated mean shift algorithm for the purposes of image segmentation is provided. I use NAIP imagery of La Porte County, Indiana, in the United States, and compare the results to hand-drawn land cover units. NAIP (National Agriculture Imagery Program) aerial imagery is an orthorectified 1-2 m^2 imagery product that cover the entire lower 48 states every 3-5 years (freely available through the USDA/NRCS geospatial data gateway).

1.2 Classification

Land cover sequences for the LCUs are estimated through classified pixel level data. Unfortunately, land cover sequences cannot be directly applied in most circumstances, due to differences in resolutions between the classified pixels and the boundaries formed from high resolution imagery.

The problem of determining LCU estimate classification from pixels is a *polygon overlay problem*. The polygon overlay problem is a geographic problem where a set of target units are desired but the properties of these units are not available. Instead, a set of source units are available with known properties, but do not share the same areal units. In this context, classified pixels are considered the source units, and the LCU estimates are the target units. A set of approaches to this problem can be found in Gotway and Young (2002). In this dissertation, a simple initial approach within the class of pixel aggregation and areal weighting is chosen, namely mapping classified pixels that are interior to an LCU's boundary to the LCU. Interior pixels are used to reduce the measurement error from pixel classifications between LCUs and misalignment of the geospatial data sources used for classification (see Dean and Smith, 2003 and Boryan et al., 2011). LCUs that do not have sufficient dimensions to have an interior classified pixel, are unlikely to have agricultural land cover and are removed.

The boundaries suggested by the mean shift or other segmentation algorithms are not perfect. LCU estimates may contain portions of other LCUs or may be a proper subset of another LCU. These errors cause loss of information with respect to the LCU's land cover sequence, make interpreting relationships between LCUs more difficult. To reduce these issues with the LCU estimators, three post-segmentation steps are performed after the assignment of interior classified pixels to LCU estimates:

1. Identification and remediation of under-segmentation;
2. Identification of agricultural LCUs, and removal of non-agricultural LCUs;
3. Identification and remediation of over-segmentation.

The sequence of these steps is important to avoid the merging of non-agricultural LCUs with agricultural LCUs. Methods to perform these three steps are presented in this dissertation.

Evaluation is performed using the misclassification rates of classified LCUs and the adjusted Rand index (see Hubert and Arabie, 1985). The adjusted Rand index is based on the Rand index (see Rand, 1971), but is adjusted to account for random clustering. The Rand index, and likewise the adjusted Rand index are statistics, based on the overlap between two segmentations of the same population. The statistic uses the pairwise counts of elements, in this dissertation pixels, that are in the same segment in both segmentations or different segments in both segmentations. For the case of two segmentations, $B = \{B_1, \cdots, B_K\}$ and $C = \{C_1, \cdots, C_J\}$, the value of $n_{i,j}$ indicates the number of pixels that are in segment i from segmentation B, and segment j in segmentation C. A two-way table

of the pixel counts is provided below,

$$\begin{array}{c|cccc|c} & C_1 & C_2 & & C_J & \\ \hline B_1 & n_{1,1} & n_{1,2} & \cdots & n_{1,J} & n_{1,\cdot} \\ B_2 & n_{2,1} & n_{2,2} & & n_{2,J} & n_{2,\cdot} \\ \vdots & & & \ddots & & \vdots \\ B_K & n_{K,1} & n_{K,2} & & n_{K,J} & n_{K,\cdot} \\ \hline & n_{\cdot,1} & n_{\cdot,2} & & n_{\cdot,J} & n \end{array} . \tag{1.1}$$

This approach is useful since there is no requirement for the indexes to be matched between the sets of segments being compared. The adjusted Rand index r has the following form

$$r = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J}\binom{n_{k,j}}{2} - a}{\frac{1}{2}\left(\sum_{k=1}^{K}\binom{n_{k,\cdot}}{2} + \sum_{j=1}^{J}\binom{n_{j,\cdot}}{2}\right) - a} \tag{1.2}$$

where

$$a = \binom{n}{2}^{-1}\left(\sum_{k=1}^{K}\binom{n_{k,\cdot}}{2}\sum_{j=1}^{J}\binom{n_{j,\cdot}}{2}\right). \tag{1.3}$$

Each of the three steps are compared against USDA's Farm Service Agency (FSA) common land units (CLUs) from 2011 or hand-drawn LCUs. FSA CLUs are a set of spatial units (agricultural field boundaries) provided to the Farm Service Agency by farmers that participate in particular federal programs, details are discussed in Section 2.2. FSA CLUs are not required to perform any of the methods in this section, but do provide a large amount of "ground truth" within the area studied.

1.3 Prediction

A spatial-temporal multinomial probit model is used to predict agricultural land cover through crop rotations. The choice of this approach was motivated by initial investigations into the spatial correlation of the number of popular crop rotations that occur at each LCU, namely corn-to-corn and corn-to-soy rotations in a subset of La Porte County, Indiana, Figure 1.1. Where each LCU, y, is indexed by $\xi \in \Delta$ with Δ being the index for the LCUs in a subset of La Porte County, Indiana. Weights between spatial units $\xi_1 \in \Delta$ and $\xi_2 \in \Delta$, $w(\xi_1, \xi_2)$ are binary valued, where a weight of one implies the existence of adjacent edges, otherwise zero. A buffer of 50 meters around each LCU was used in determining adjacency, to allow for neighbors across roads and other boundaries.

In the literature the use of adjacent edges is called a "rook" based approach, while using adjacent edges and vertices is considered a "queen" based approach, in reference to the game of chess. The presence of spatial association was assessed through Moran's I (1.4), and simple plots of the prevalence of particular rotations, seen in Figure 1.1. Both the Moran's I results, Table 1.1, and plots seemed to suggest the presence of some spatial dependence.

$$I = \frac{n}{\sum_{\xi_1 \in \Delta} \sum_{\xi_2 \in \Delta} w(\xi_1, \xi_2)} \frac{\sum_{(\xi_1 \in \Delta)} \sum_{(\xi_1 \in \Delta)} (y(\xi_1) - \bar{y}) (y(\xi_2) - \bar{y})}{\sum_{\xi_1 \in \Delta} (y(\xi_1) - \bar{y})^2} \tag{1.4}$$

A multinomial probit model (MNP) was chosen due to its ability to link categorical response, major commercial crops, to a multivariate normal linear model. This multivariate normal form is easy to extend as a hierarchical model, and has simpler error structure than multinomial logits and related approaches. In this section, a brief introduction to multinomial probit models and spatial autoregressive models is presented followed by the crop rotation model.

Multinomial probit models provide a link function between a categorical response and a $J = C - 1$ dimensional linear model with a multivariate normal error structure, where C

Table 1.1: Measures of spatial association for specific rotation counts at each LCU under a "rook" based neighborhood.

Rotation	Moran's I	p-value
Corn-to-Corn	0.25	0.005
Corn-to-Soy	0.24	0.005

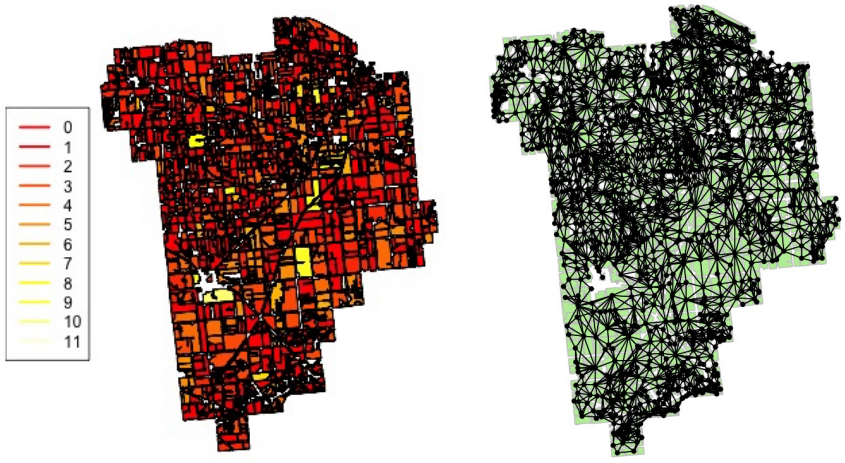

Figure 1.1: Counts of Corn-to-Corn Rotations 2001-2013 in a subset of La Porte County, Indiana (Left), initial spatial neighborhood using Rook based association (Right).

is the number of categorical classes. This linking is performed by partitioning the support of the linear model Figure 1.2,

$$\mathbf{y}(\xi) = \begin{cases} c = 1 & \max\left(\mathbf{z}_j(\xi)\right) \leq 0 \qquad\qquad\qquad\qquad \forall j \in \{1, \cdots, J\} \\ c = j+1 & \text{argmax}_{j \in \{1,\cdots,J\}} \left\{ \mathbf{z}_j(\xi) \mathbb{I}_{\{\mathbf{z}_j(\xi) > 0\}} \right\}) \end{cases} \tag{1.5}$$

where the observation of category $c \in \{1, \cdots, C\}$ is determined by $z_j(\xi)$, the j^{th} element of the latent vector $z(\xi)$ with $j \in \{1, \ldots, J\}$.

Inference in multinomial probit models is done primarily through the data augmentation within the Bayesian paradigm, an extension of the univariate approach suggested by Albert and Chib (1993). Issues of parameter identifiability are handled in McCulloch et al. (2000),

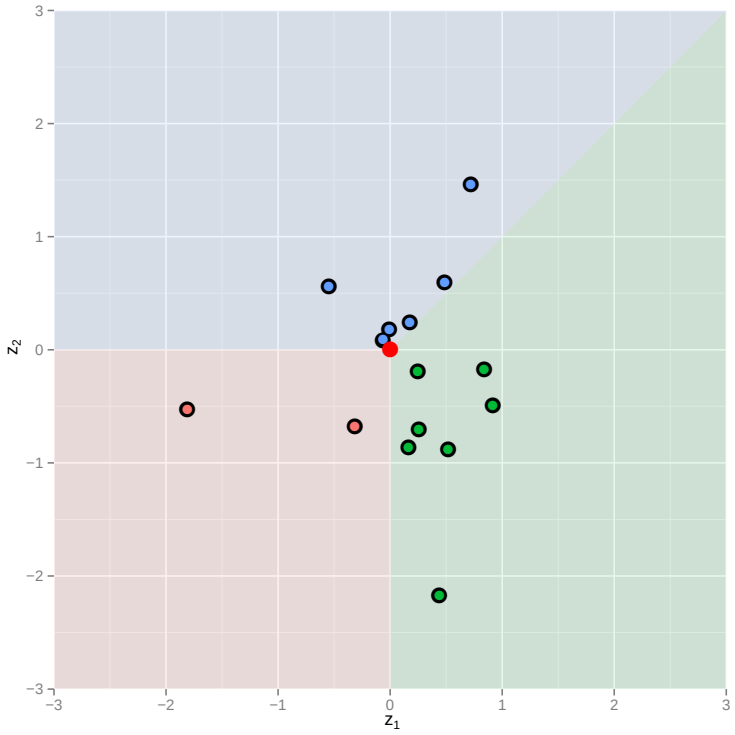

Figure 1.2: An example of data augmentation in MNP, intercept in red and other points represent deviates generated from the latent variable.

and advanced in Imai and van Dyk (2005) and Burgette and Nordheim (2012). The multivariate extension of the univariate approach of Albert and Chib (1993) was first suggested by McCulloch and Rossi (1994) with latent variable form

$$\mathbf{Z}^*|Y = U^*\beta^* + \boldsymbol{\epsilon}, \tag{1.6}$$

where

$\mathbf{Z}^*|Y$ = $(\mathbf{z}_{1,1}, \cdots, \mathbf{z}_{1,J}, \cdots, \mathbf{z}_{n,J})$, a latent response vector of length nJ conditioned on the observed state Y,

β^* = a vector of coefficients of length mJ,

$\boldsymbol{\epsilon}$ = is the error distributed as $\mathcal{TN}(0, I_n \otimes \Sigma, A, B)$ a truncated normal distribution where A and B are vectors of lower and upper bounds respectfully defined by the observed states Y,

Σ = a $J \times J$ covariance matrix, and

U^* = $U \otimes I_J$ an $nJ \times mJ$ matrix with

U = an $n \times m$ matrix of covariates and

I_J = an $J \times J$ identity matrix.

$\otimes$ is the Kronecker product.

There are two common spatial autoregressive Gaussian models, simultaneous autoregressive (SAR) (see Whittle, 1954) and conditional autoregressive (CAR) (see Besag, 1974). Of these two forms, the SAR model is more popular in MNP models (see LeSage and Pace, 2009 and Wang et al., 2012). The standard SAR model has the following form (called the spatial error model (SEM) by LeSage and Pace (2009))

$$\mathbf{Z} = B\mathbf{Z} + (I - B)U\beta + \boldsymbol{\epsilon}, \tag{1.7}$$

where

$\mathbf{Z}$ = is a vector of length n,

β = is a length m vector of location parameters,

$\boldsymbol{\epsilon}$ = is a random vector composed of length n with distribution $(0, \Sigma)$,

Σ = an $n \times n$ covariance matrix, and

B = is an $n \times n$ matrix of spatial weights.

Under this approach Z is distributed as

$$\mathbf{Z} \sim \mathcal{N}\left(U\beta, (I-B)^{-1}\,\Sigma\left(I-B^{T}\right)^{-1}\right) \tag{1.8}$$

In practice, the $n \times n$ weight matrix B is usually simplified to ρW, where ρ is a scalar and W is an $n \times n$ matrix of fixed spatial weights. The matrix W is sometimes referred to in the literature as a spatial adjacency matrix. In the simplest of cases, pairwise neighbors in this matrix are set to one while all other values are set to zero, and to provide a more natural weighting of observations each row is scaled to sum to one. The scalar ρ is sometimes called an autocorrelation parameter, but as noted by Wall (2004), the value does not necessarily indicate any particular amount of autocorrelation. In the case of the row-scaled W, ρ is necessarily bounded on the interval $(1/\lambda_{\min}, 1)$, where $\lambda_{\min}$ is the minimum eigenvalue from W.

Spatial autoregressive MNP models or SAR MNP models have been explored by LeSage and Pace (2009), with spatial-temporal approaches developed by Wang et al. (2012). Although these models are called SAR models they differ from (1.7),

$$\mathbf{Z}^{*} = B\mathbf{Z}^{*} + U\beta + \boldsymbol{\epsilon}, \tag{1.9}$$

where

$$\begin{aligned} B &= \rho W \otimes I_J, \\ W &= n \times n \text{ matrix of spatial weights (“rook” in this application), and} \\ \rho &= \text{scalar parameter for } W. \end{aligned}$$

The type of model is sometimes called “mixed regressive and spatial autoregressive model” or “lagged response model“ in ecology.

The MNP and SAR MNP models described above provide a way to link categorical response to a linear model with multivariate normal error structure. What remains to be done is to link these models with the crop rotation phenomena used for prediction. To do

this, it is assumed that the categorical response is temporally stationary and conditionally independent given a prior state, sequence of prior crops, on the same LCU. Therefore, by specifying a set of prior states (rotations) P, it is then possible to create a design matrix to include the prior state information for each LCU.

This simple MNP model using prior crop rotations states works well, in that it can admit other exogenous variables when available such as transportation cost, or soil moisture data, but provides no way to model potential missing variables that may have spatial relationships such as land ownership or the social relationship between farmers. A novel hierarchical model is presented in this paper that models spatially correlated crop rotations, e.g. corn-to-soy rotations are close to corn-to-soy rotations. The autocorrelated crop rotation model approach has the form

$$\mathbf{Z}^* = U^{**}\boldsymbol{\beta}^{**} + X\gamma + \boldsymbol{\epsilon}, \tag{1.10}$$

$$\boldsymbol{\beta}^{**} = B\boldsymbol{\beta}^{**} + (I - B)\beta_0 + \boldsymbol{\upsilon}, \tag{1.11}$$

where

U^{**}	=	$(T-l)nJ \times nPJ$ design matrix,
T	=	the number of observed years,
l	=	the lag, number of prior years,
$\boldsymbol{\beta}^{**}$	=	nPJ vector of covariates,
X	=	other covariates,
γ	=	other covariates coefficients,
β_0	=	hyper parameter for β_0 and
$\boldsymbol{\upsilon}$	=	random vector of length n with distribution $\mathcal{N}(0, \Sigma_\upsilon)$.

This model differs from prior SAR MNP models of LeSage and Pace (2009) and Wang et al. (2012), where the response is considered auto-correlated, e.g. corn grows next to corn. A problem with this model is the large number of parameters introduced; however, for larger values of ρ, the number of effective parameters is actually much lower than nPJ

avoiding excessive over-fitting.

This model exhibits several computational challenges in the traditional gibbs sampling framework, largely due to the sparse nature of the design matrix U^{**}. Therefore, extensive use of sparse matrix libraries was needed, this prevented direct use of any pre-existing software to handle this model. To provide results and to evaluate this model, a software package for the R language (see R Core Team, 2015) was created (see Lisic, 2015a). This software package for R was created based on the *MNP* package (see Imai and Van Dyk, 2005).

In summary the following contributions to crop rotation modeling are provided:

1. A novel spatial-temporal model for crop rotations;

2. Computational methods to handle the sparse structure of the proposed model;

3. An evaluation of the proposed model against a real data set.

Chapter 2: Literature Review and Relevant Theory

In this dissertation, the topic being studied is the prediction of agriculture at the parcel level, through the use of well defined spatial-temporal units. In this chapter, the well defined spatial-temporal unit is introduced, and the three steps to construct and perform prediction on these well defined units are presented. No existing literature deals with all four of these issues simultaneously even when relaxing the constraints on algorithms and models; therefore, the literature review will approach each issue independently.

2.1 Image Segmentation

In Section 2.1, a well defined land cover unit will be defined, and the means shift algorithm will be applied to estimate these units from high resolution imagery. The mean shift algorithm is used for classification and estimation of local maxima of a kernel density estimate. The properties of the mean shift sequence will not be explored in this dissertation. Appendix A provides a review of both the mean shift algorithm and kernel density estimators. To provide sufficient background on both prior theory and relevant works, an overview of the following topics is provided; filtered estimators, spatial sampling, topics associated with increasing the performance of the mean shift iterator, and application of image segmentation in agriculture.

2.1.1 Land Cover Unit

A parcel could define many possible spatial units such as a city block, a subdivision, or an agricultural field. In this dissertation, the desired spatial unit for land cover modeling is one that has a maximal boundary surrounding a single land cover sequence. This spatial unit is called a *land cover unit* (LCU). This spatial unit is a result of favoring unique land use

sequences, over having ideal boundaries for a specific year. The unique land use sequence requirement creates spatial units that are the collection of intersections of yearly spatial units. These spatial units are also a function of the length of the land cover sequence, where a land cover sequence of length one may produce spatial units significantly different from spatial units formed from longer land cover sequences. The change in spatial units as a function of sequence length, also affects the relationships between adjacent spatial units.

An alternative approach is to assume that the boundaries are constant over some window of time. In this case, there is a potential compromise on both relationships between adjacent spatial units and the land cover sequence under the condition of boundary changes within the temporal window. The constant boundary approach can be seen in Long et al. (2014) through a study of changes in field-level cropping sequences in North-East Montana. Long et al. (2014) use the Montana cadastral framework from the Montana State Library's Geographic Clearinghouse for June 2012. These boundaries are assumed constant over a temporal window from 2001-2012. This set of land cover units were considered to generally follow agricultural field boundaries, but no measure of the accuracy over that period was provided in the document. Other authors, such as Yan and Roy (2014), also make the strong assumption of little change in field boundaries between years in their spatial units constructed from remotely sensed images.

A third approach is to simply evict units in a data frame that fail to meet a particular definition. Osman et al. (2015) evicts all units from a set of spatial units that do not have a distinct land cover sequence, as does Boryan et al. (2011) in the selection of training data for classification. Such an approach may introduce bias into estimates of parameters, and may distort true spatial relationships.

2.1.2 Filtered Estimators

Kernels and filters are two words used in different ways by different disciplines describing similar functions. In image processing, a filter and kernel sometimes refers to the convolution of a computer image with a spatially weighted function. In statistics and also in image

processing, kernel may refer to a function with a set of properties defined in Appendix A. In this dissertation the term kernel, and filter when dictated by convention, will refer to the latter case. The term *filtered estimate* will refer to the convolution of an estimate g with the kernel k, a generalization of the former case.

Kernel density estimates, are therefore filtered estimates where $g(x) = 1$. Literature for filtered estimators in general is more limited than that of kernel density estimators, and tends to follow two general paths in the statistical literature:

- local application of statistical methods;
- robust estimators for noisy surfaces.

The first group of filtered estimates covers topics including local polynomial regression such as the popular Nadaraya-Watson Estimator, local likelihood estimation for generalized linear models (see McCullagh and Nelder, 1989), and localized EM for fitting mixture models (see Kauermann, 2001). The second group may use similar methods, but the focus is on robust estimators of surfaces (see Lee, 1983 and Chu et al., 1998). Applications of the second group are common in medical imaging (see Chu et al., 1998 and Godtliebsen and Spjsotvoll, 1991) and remote sensing (see Lee, 1980). This dissertation will focus on methods in the second group.

Filtered estimates are used in image segmentation to enhance edges and to reduce noise. In signal processing, this is known as increasing the signal-to-noise ratio (SNR). Filtered estimates in applications to image segmentation operate on pixels that are addressed by indexes s in an image D, where s is a Cartesian coordinate of form $s = (s_1, s_2)$. The most common approach, is the application of a Gaussian filtered estimate. The Gaussian filtered estimate is a Nadaraya-Watson estimator,

$$f_{NW}(y) = \frac{\sum_{i=1}^{n} x_i \kappa\left(\frac{y_i - y}{h}\right)}{\sum_{i=1}^{n} \kappa\left(\frac{y_i - y}{h}\right)}. \tag{2.1}$$

The Nadaraya-Watson estimator is a conditional filtered estimate, where instead of applying

the convolution to x, the convolution is applied to $x|y$, and $g(x|y) = x|y$. By setting $x_i = v(s)$ and $y = s$ the Nadaraya-Watson becomes the Gaussian filtered estimate,

$$f(s) = \sum_{s^* \in D} v(s^*) w(s^*, s) \tag{2.2}$$

where

$$w(s^*, s) \propto \phi\left(||s - s^*||_H^2\right), \tag{2.3}$$

$s, s^* \in D$, H is a bandwidth parameter, $||\cdot||_H^2$ is the Mahalanobis distance, and $\sum_{s \in D} w(s, s^*) = 1$.

A secondary filtered estimate can be used to identify regions of an image where the magnitude of the gradient with respect to s is large (see Canny, 1986). In computer vision these methods are called edge-detection methods. Edge-detection filtered estimates are finite distance approximations to the magnitude; the most popular are the Sobel filtered estimate that provides directional gradients, and the Lapalacian that provides non-directional gradients Bradski and Kaehler (2008).

These methods can be combined to form edge-preserving filtered estimates. The most common edge-preserving filtered estimate is the bilateral filtered estimate. The name bilateral filtered estimate or "bilateral filter" was coined by Tomasi and Manduchi (1998), although the original work by Lee (1983) and extension by Chu and Marron (1991) used the term "sigma filter". The bilateral filtered estimate has a form similar to (2.2) with the weight w replaced by

$$w^*(s^*, s) \propto \phi\left(||s - s^*||_H^2\right) \phi\left(||v(s) - v(s^*)||_{H_v}^2\right) \tag{2.4}$$

where H_v is the scalar bandwidth for the kernel involving the pixel value. In this approach, the weight is a function of the value and spatial distance. This enhances edges by decreasing the weight of neighboring pixels that are spatially close, but have values significantly

different from the pixel being filtered.

Both using a Gaussian filtered estimate with edge detection and bilateral filtering work well at reducing variance and preserving edges in images with moderate or constant variation, but work poorly with high variation or large differences in variation within the same image. These situations are common in high resolution imagery due to trees and developed structures such as buildings. If the trees and structures provide regular patterns, wavelet transforms can be used with other filtered estimates to assist in segmentation (see Trias-Sanz et al., 2008 and Aksoy et al., 2012).

An alternative to using wavelets is to use the local variance. The local variance as described by Lee (1980) is the sample variance applied to a spatial neighborhood;

$$\mathbf{S}_h^2(s,t) = \sum_{s^* \in D} \sum_{s^{**} \in D} w_h(s,s^*) w_h(s,s^{**}) \frac{(\mathbf{x}(s^*) - \mathbf{x}(s^{**}))^2}{2n(n-1)} \tag{2.5}$$

Local variances have been employed in dynamic or adaptive bandwidth for kernel density estimation and filtering. However, these filtered estimates simply increase the bandwidth of the kernel in areas of high variance, or decrease the bandwidth areas of low bandwidth, but do not filter using the local variance. This is the goal of the local variance filter proposed in this dissertation.

The *log variance filter* (LVF) can be considered a filtered estimate, or a filter applied to the log of the local variance (3.3). The log transformation helps easily identify areas of high and low local variation. An example of this log variance can be seen in Figure 2.1. The LVF is used in conjunction with an intensity value for estimating LCU boundaries.

In the computer vision and graphics literature little attention is given to the pointwise and global properties of filter estimates. Instead, more heuristic approaches are taken by calculating the IMSE over a set of images such as the popular "lena.jpg" or collections of images categorized by image attributes. This may be understandable, given that the asymptotic results are not generally useful in direct application of filtered estimates to

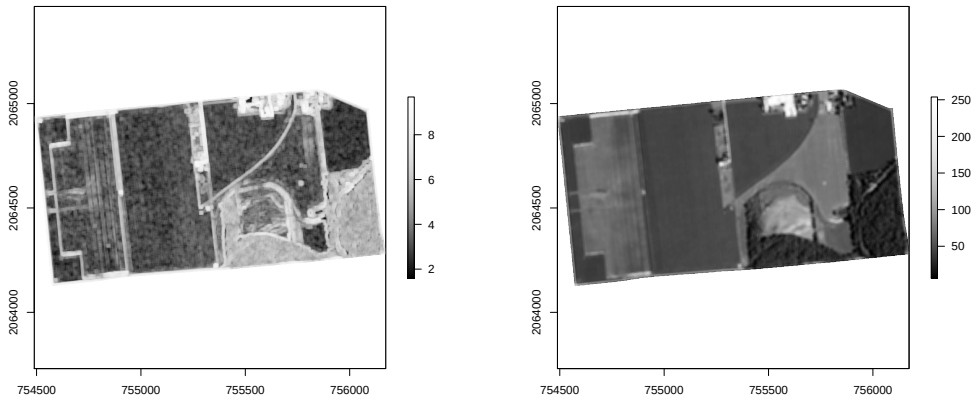

Figure 2.1: Local log-variance (Left), and NAIP imagery (Right).

images where larger bandwidths may be needed.

The LVF is approximately normally distributed and is proportional to Fisher's Z statistic. This implies that the LVF is a spatial average of pdf values of a local test statistic for the variance. The theoretical properties of the LVF are related to the work by Panaretos et al. (2005) on ratios of correlated gamma random variables, and Shoemaker (2003) on robust alternatives to the F-test.

The LVF is used for two tasks in this dissertation. (1) the LVF is indirectly used as a component of, the kernel density based, mean shift segmentation algorithm. This application applies a Gaussian kernel to the output of a filtered estimates and the log of the local variance. (2) the LVF is used to help with image stratification to accelerate the mean shift algorithm. To minimize variance between samples and to ensure useful landmarks are retained, a spatially stratified design is used. The strata in this design are determined through a threshold applied to large and small values of the LVF, where large LVF values indicate areas of constant variance, and small values indicate areas of non-constant variance, such as LCU boundaries.

2.1.3 Spatial Sampling

Sampling is often an effective method of data reduction. Sampling as a form of data reduction in mean shift can be found in Xiao and Liu (2010) and Freedman and Kisilev (2009) to reduce computational burden. However, the "sampling" in Freedman and Kisilev (2009) is confusing in its deviation from statistical theory. In Freedman and Kisilev (2009), error is added to each element of the sample to simulate noise in the image. This is both a computationally and statistically inefficient way to sample from an image.

An alternative method is to follow the well developed literature in survey sampling. In survey sampling, sampling is performed to reduce the cost of complete enumeration of a population. In this dissertation, cost implies computational burden, however in sample survey methodology cost is usually financial.

If the population can be associated with a spatial index, then a spatial sample can be drawn. The advantage of sampling through a spatial index depends on the degree of spatial correlation between elements in the population. If the elements in the population have a positive spatial correlation, then by sampling one unit, some information on the surrounding units will also be collected. Therefore, by "balancing" the survey, sampling such that the units are well distributed in space, a spatial sample can be more efficient with respect to variance reduction than non-spatial approaches.

The idea of spatial sampling has a long history; early works include Quenouille (1949). Spatial sampling can be broken down into aligned and unaligned designs (see Quenouille, 1949), where aligned designs follow a systematic pattern such as a grid. In the more general world of survey sampling the grid method would be akin to systematic sampling. Much like systematic sampling, a disadvantage of aligned designs is that the sampled units may be coincident with a reoccurring phenomena such as roads.

Breidt (1995) described a generalization of grid sampling, which allows the user to control how aligned the vertexes in a grid are allowed to be. In the most restrictive case the grid is aligned, in the less restrictive case the grid is unaligned. In this generalization each vertex in the grid is associated with an index $(i, j) : i = 1, \ldots, I$ and $j = 1, \ldots, J$. Each

Table 2.1: Design type given parameter ϕ_k in Breidt (1995).

ϕ_k	Design Description
1	Systematic sample.
0	One point per stratum sample (stratum = grid cells).
-1	Balanced systematic sampling.

vertex location is determined through a Markov chain of the form

$$\mathbf{x}(i,j) = s_1 + a(j-1) + a\Phi(\mathbf{v}_1(i,j)) \tag{2.6}$$

$$\mathbf{x}(i,j) = s_2 + b(j-1) + b\Phi(\mathbf{v}_2(i,j)) \tag{2.7}$$

$$\mathbf{v}_1(i,j) = \begin{cases} \mathbf{z}_1(i,j) & j = 1 \\ \phi_1 \mathbf{v}_1(i,j-1) + z_1(i,j)\sqrt{1-\phi_1^2} & j \in \{2,\cdots,J\} \end{cases} \tag{2.8}$$

$$\mathbf{v}_2(i,j) = \begin{cases} \mathbf{z}_2(i,j) & j = 1 \\ \phi_2 \mathbf{v}_2(i-1,j) + z_2(i,j)\sqrt{1-\phi_2^2} & j \in \{2,\cdots,J\} \end{cases} \tag{2.9}$$

where s_1 is a spatial location (East-West), s_2 is a separate spatial location (North-South), and $\mathbf{Z} \sim \mathcal{N}(0, I)$ is the vector of all $\mathbf{z}(i,j)$. In this design, the parameter ϕ_k, $k \in \{1,2\}$, determines the structure of the sample Table 2.1. The exact probability that a particular location will be in the sample, the inclusion probability, is difficult to calculate under this approach; therefore, Breidt (1995) provides an approximation through simulation.

The local pivotal method (LPM) of Grafström et al. (2012), which is an extension of the pivotal method of Deville and Tille (1998), departs from the prior generalization of sampling on a grid. This method creates *spatially balanced samples*. Stevens and Olsen (2004) defines the term spatially balanced through a Voronoi tessellation for each sample.

Under this definition, the statistic

$$B = \sum_{i=1}^{n} \left(1 - \sum_{s \in V_i} \pi(s) \right)^2 n^{-1} \tag{2.10}$$

is used to determine how spatially balanced a sample of size n is, where V_i is a polygon from the Voronoi tessellation of the sample, and $\pi(s)$ is the probability that a sampling unit at location s is included a sample. Note, $\sum_{s \in D} \pi(s) = n$ and $\pi(s) > 0 \ \forall s \in D$, where D is a spatial index with N elements. If s is on the boundary between multiple polygons, then the sampling weight is divided equally between the polygons. If B is close to zero for a sample, then the sample is considered spatially balanced. This occurs when each polygon contains approximately the same total inclusion probability.

Algorithm 1. $S \leftarrow$ Random Sequence of $\{1, \ldots, N\}$

for all $s \in S$ **do**

$m \leftarrow 0$

$R \leftarrow \emptyset$

{Get neighbors of equal distance.}

for all $r \in \{q : q \in S, q \neq s\}$ **do**

if $d(s, r) < m$ **then**

$R \leftarrow \{r\}$

$m \leftarrow d(s, r)$

else if $d(s, r) = m$ **then**

$R \leftarrow R \cup \{r\}$

end if

end for

{Handle ties.}

$r \leftarrow$ SAMPLE(R) {A function that returns a random sample from a set.}

{Check if s is the closest neighbor of r.}

for all $t \in \{q : q \in S, q \notin \{r, s\}\}$ **do**

if $d(t, r) < m$ **then**

$R \leftarrow \{t\}$

$m \leftarrow d(s, t)$

else if $d(t, r) = m$ **then**

$R \leftarrow R \cup \{t\}$

end if

end for

if $s \in R$ **then**

$(\pi'(s), \pi'(r)) \leftarrow g(\pi(s), \pi(r))$

end if

end for

The LPM algorithm is provided as Algorithm 1 where $\pi(s) + \pi(r) < 1$ and

$$g(x, y) = \begin{cases} (0, x + y) & \text{with probability} \frac{y}{x+y} \\ (x + y, 0) & \text{otherwise} \end{cases} \tag{2.11}$$

otherwise

$$g(x, y) = \begin{cases} (1, x + y - 1) & \text{with probability} \frac{1-y}{2-x-y} \\ (x + y - 1, 1) & \text{otherwise.} \end{cases} \tag{2.12}$$

This algorithm has since been updated by Grafström et al. (2012) with the LPM2 algorithm, providing substantial speed up over LPM by removing the second neighbor check. The worse case running time for LPM in a naïve implementation is $\mathcal{O}(N^3)$, where N is the population size, while the LPM2 algorithm has a worse case running time of $\mathcal{O}(N^2)$. The worst case running time under LPM2 comes from the N^2 distances calculated for each sampling unit.

The computational burden of LPM2 for large sample sizes is noted in Grafström et al. (2012). In a later paper, Grafström et al. (2014), LPM2 is approximated through the use of a restricted spatial neighborhood to decrease the computational burden. An alternative approach is provided in this dissertation, but without approximating the LPM2 result. This approach replaces the linear searches with a k-d tree data structure to obtain an average computational complexity of order $\mathcal{O}(N \log(N))$.

2.1.4 Mean Shift and Newton's Method

Mean shift is a mode searching algorithm similar to the typical Newton's Method for finding roots of a smooth function but applied to a kernel density estimator. A short overview of the algorithm and its properties used in this dissertation are provided in Appendix A. For the purposes of the literature review the mean shift will be quickly reviewed.

The mean shift algorithm is used to classify a set of *query* points, identified by the set Q to a set of local maxima in the kernel density estimate from another set of points R, known as the *reference* set. Each point v in R or Q is assumed to be an iid observation of the random variable $\mathbf{x}$ with pdf f and support in the d-dimensional space $\mathbb{R}^d$. Q and R are assumed to both have the same support, and frequently $Q = R$. Because this is a root searching method, there is no dependency on a fixed number of clusters, instead the choice of both kernel and bandwidth parameters are determine the number of local maxima in the kernel density estimate (KDE). For convenience, the cardinality of the sets Q and R will be denoted by N_Q and N_R respectfully, and the KDE will be rewritten as

$$\hat{f}_{N_R,H}(v) = \sum_{j=1}^{N_R} |H|^{-1} n^{-1} \kappa\left(H^{-1}\left(v - x_j\right)\right) = \sum_{j=1}^{N_R} c_{j,H} k\left(||v - x_j||_H^2\right) \tag{2.13}$$

where $c_{j,H}$ is a constant ensuring that $k\left(||v^i - x_j||_H^2\right) c_{j,H} = \kappa\left(||v^i - x_j||_H^2\right)$. In the context of mean shift, the kernel κ is known as the *shadow kernel*, k is known as the *shadow profile*,

and $g = -k'$. The form of the mean shift iterator is

$$v^{i+1} = \frac{\sum_{j=1}^{N_R} x_j c_{j,H} g\left(||v^i - x_j||_H^2\right)}{\sum_{j=1}^{N_R} c_{j,H} g\left(||v^i - x_j||_H^2\right)}. \tag{2.14}$$

The mean shift algorithm is a steepest ascent method used to find critical points of $\hat{f}_n$, the kernel density estimate calculated using the points in R. Other methods to find critical points include the application of root finding methods to $\nabla \hat{f}_n$, the gradient of the kernel density estimate $\hat{f}_n$. A popular root finding method is Newton's method (2.15).

$$v_n^{(i+1)} = v_n^{(i)} - \lambda_n'\left(v_n^{(i)}\right)^{-1} \lambda_n\left(v_n^{(i)}\right) \tag{2.15}$$

where first and second derivatives of the kernel estimate are

$$\lambda_n = \sum_{j=1}^{N_R} \pi_j c_{j,H} g\left(||(v^{(i)} - x_j)||_H^2\right)\left(v^{(i)} - x_j\right) \tag{2.16}$$

and

$$\begin{aligned} \lambda_n' &= \sum_{j=1}^{N_R} \pi_j c_{j,H} \left(g\left(||(v^{(i)} - x_j)||_H^2\right) I_d \right. \\ &+ \left. g'\left(||(v^{(i)} - x_j)||_H^2\right) |H|^{-1/2}\left(v^{(i)} - x_j\right)\left(v^{(i)} - x_j\right)^T |H|^{-1/2}\right). \end{aligned} \tag{2.17}$$

Under the condition that both Newton's method and mean shift converge to the same stationary point, Newton's method has the distinct advantage of a quadratic convergence rate, Table 2.2.

The relationship between Newton's method or improvements to mean shift through Newton's method have been explored by numerous authors for example Yang et al. (2003a), Fashing and Tomasi (2005), and Chiu et al. (2008). Direct comparisons of the method have

Table 2.2: Comparison of Newton's method and Mean Shift convergence to a stationary point at -0.14.

Iteration	Mean Shift	Newton
1	1.00	1.00
2	0.51	-0.78
3	0.25	-0.12
4	0.09	-0.14
5	-0.00	-0.14
6	-0.06	-0.14
7	-0.09	-0.14
8	-0.11	-0.14
9	-0.13	-0.14
10	-0.13	-0.14
11	-0.14	-0.14

been discussed in Yang et al. (2003a) where the gradients were compared. Further work done by Fashing and Tomasi (2005) shows that the Newton step (2.15) is equivalent to mean shift when g is a piecewise constant function. This does not hold for other kernels such as the Gaussian.

Approximation of mean shift through the use of other iterators or transformations has been explored by numerous authors. An early approximation of mean shift in the literature occurs in Yang et al. (2003a), where a quasi-Newton method known as BFGS (Broyden, Fletcher, Goldfarb, and Shanno) is used to approximate a mean shift method with a Gaussian kernel, where a quasi-Newton method is an approximation to Newton's method obtained by approximating the Hessian. Similarly, Chiu et al. (2008) used a quasi-Newton's method, but delayed employment of this method until a threshold on $||\hat{f}_n(y^{(i+1)}) - \hat{f}_n(y^{(i)})||_2^2$ is met. The same authors of Yang et al. (2003a) also proposed another approximation of mean shift in Yang et al. (2003b) through the use of an efficient expansion of the sum of Gaussian kernels known as as a "fast Gaussian transform". This method suffers in higher dimensions and is considered out of scope for this research.

Convergence of Newton's method to a unique root under a set of regularity conditions and a variety of modes has been explored by a number of authors. The most relevant

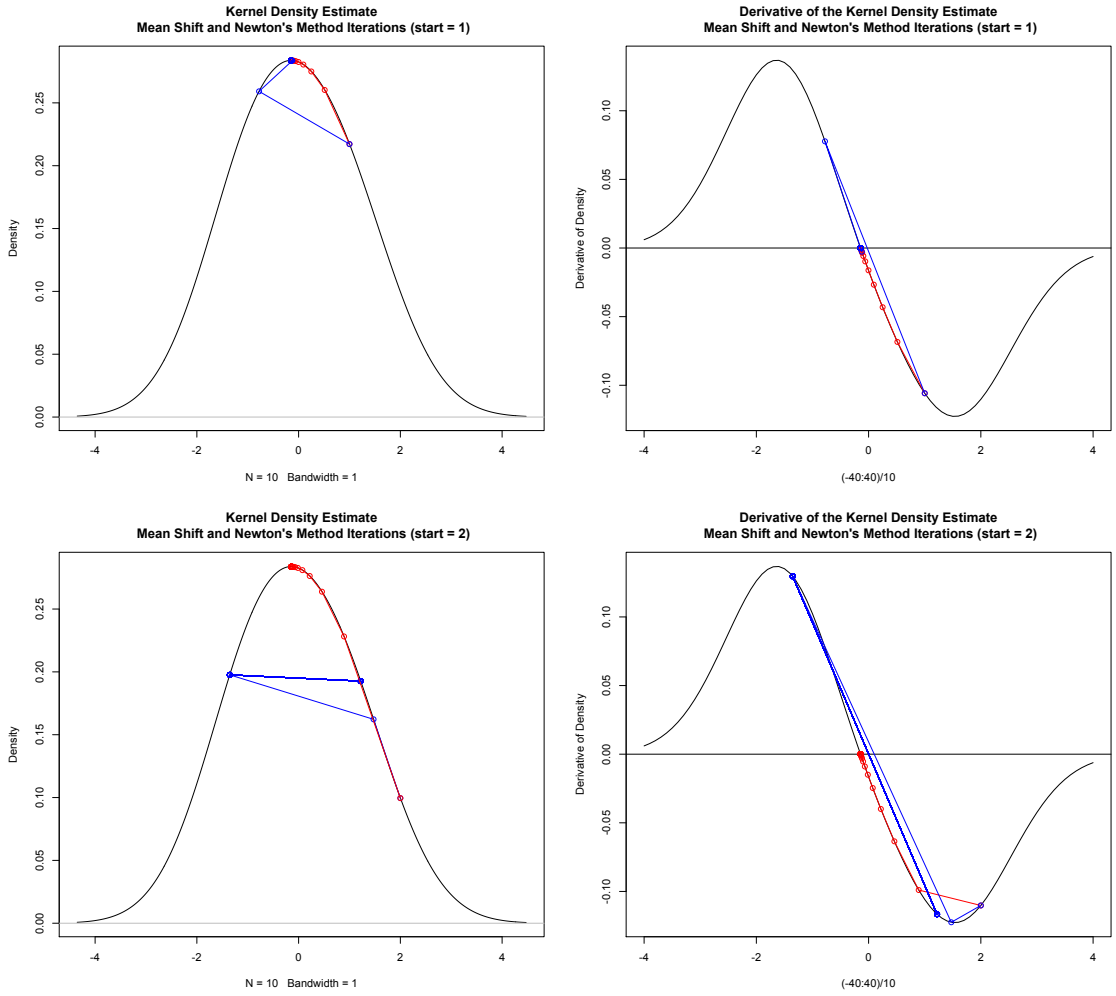

Figure 2.2: Comparison of Newton's method and Mean Shift

article on this topic is Clarke and Futschik (2007), which provides conditions for uniform convergence of an estimator of the local maxima of a KDE to the true local maxima of the density the KDE is estimating. Similar to Chen et al. (2014a), the uniform convergence convergence in Clarke and Futschik (2007) result is based on the strong uniform convergence of multivariate KDEs in Giné and Guillou (2002).

Clarke and Futschik (2007) provides a proof for the uniform convergence of $\hat{\theta}_{n,0} = \theta_0$ as n goes to zero where $\hat{\theta}_{n,0} = \theta : \nabla \hat{f}_n(\theta) = 0$ and $\theta_0 = \theta : \nabla f(\theta) = 0$ under a set of regularity conditions:

A1 Both λ and λ_n are continuously differentiable.

A2 All components and partial derivatives of λ_n converge uniformly to those of λ.

A3 The Jacobian of λ' corresponding to the derivative of λ is continuous and $|\lambda| \neq 0$.

N1 For x and y in a ball B_0 around a unique root θ_0 for a sufficiently large n λ_n' is Lipschitz continuous under the L2 norm.

N2 $||(\lambda')^{-1}(x)|| \leq \beta$ and $||(\lambda')^{-1}(x)\lambda(x)|| \leq \eta$ for constants β and η satisfying $\alpha = \beta\eta\gamma^{(e)} < \frac{1}{2}$ on some ball $B_\delta \subset B_0$ satisfying $r(\beta_\delta) + t* < r(B_0)$,

where $r(\cdot)$ is the radius function for a ball argument, A1–A3 provide unique roots, and N1–N2 provide sufficient conditions for convergence to those unique roots. N1 is a typical Lipschitz continuity requirement placed on the complete metric space of B_δ as seen in many fixed-point theorems, and N2 is a condition specified in Giné and Guillou (2002). This is met for a Gaussian kernel over a sufficiently smooth f with MSE optimal bandwidth.

2.1.5 Mean Shift Implementations

One major issue with mean shift is that in a naïve implementation, each iteration would require calculating the kernel, including the distance, for all observations in the data set. This gives the procedure a computational complexity of order $N_R N_Q m$ where m is the number of iterations. This polynomial running time would make the algorithm unusable for large data sets. Therefore, binning the observations in the feature space for some reduction in precision can be performed to gain a massive speed up in computation, a common technique for kernel density estimators (seee Wand and Jones, 1994). Unfortunately, binning is not useful in higher dimensions, since the number of bins required to retain a fixed level of precision increases exponentially as the number of dimensions increase, due to the curse of dimensionality.

An alternative approach would be to only calculate the kernel for points sufficiently close to the point being shifted, or a fixed number of close points (k nearest neighbors). Casting

mean shift as a nearest neighbors problem opens up the implementation to advances in the well developed nearest neighbors literature. Because not all points are used to calculate the shift, it should be noted that this approach does sacrifices some precision for a potential large increase in speed.

Initial work in increasing the speed of the algorithm has been through an associated problem to kernel density estimation, k nearest neighbor estimation (KNN). For KNN in high dimensions, the most computationally difficult part of the problem is finding neighbors. Instead of searching all N_R points, for a fixed initial cost, an alternative data structure can be used to store all N_R points, and achieve sub linear search times.

The first published work on implementation of fast nearest neighbor algorithms in mean shift can be found with Georgescu et al. (2003), who used an approach called locality sensitive hashes (LSH). LSH has fallen out-of-favor in the current literature to other data structures such as k-d trees, that are more computationally efficient. k-d trees have an average search time of $\mathcal{O}(\log(N_r))$ and tree construction time of $\mathcal{O}(N_R \log(N_R))$ Wang et al. (2007) Xiao and Liu (2010). Approximate nearest neighbor algorithms can further reduce the computational burden by returning the closest neighbors of an observation that can be found in a fixed time, or fixed search radius.

The dual k-d tree approach to mean shifting by Wang et al. (2007) is interesting in that it both speeds up the search time for neighbors and computational burden, by operating on branches of the k-d tree instead of leaves. This acceleration is achieved by forming a static k-d tree for the reference data and a separate query tree that is rebuilt each iteration. In this scenario, the reference tree returns an approximate estimate of the mean shift contribution for each observation within a branch or leaf of the k-d tree with a fixed tolerance level.

2.1.6 Image Segmentation of Remotely Sensed Images

Segmentation of remotely sensed images for the purpose of identifying crop boundaries, or "crop field extraction" has been explored by numerous authors. The most notable examples can be found in Zhang (2009) and Yan and Roy (2014). An application of mean shift to

Table 2.3: WELD Landsat bands, Yan and Roy (2014)

Band	Description
1	Blue
2	Green
3	Red
4	Near Infrared
5	Short-wave Infrared
6	Thermal Infrared
7	Short-wave Infrared

remotely sensed images explicitly for crop field extraction, has not been found in the literature. However, this algorithm has found use in the more general case of image segmentation of remotely sensed images.

Yan and Roy (2014) recently proposed an automatic method for crop field extraction. This method is considered the most comparable to the method presented here due to its ability to produce field extractions over large areas.

The extraction method uses web enabled Landsat data (WELD). WELD tiles are 150km^2 in size, and cover five years in increments of one week. Each pixel within these tiles is 30m^2 in size, and include a number of bands, Table 2.3. The bands of greatest interest are red and near-infrared, these are used to form the NDVI (normalized difference vegetation index), a common indicator of plant activity.

A function of the NDVI is used to determine areas that are likely to be agricultural fields. To remove clouds, aircraft and other undesirable features, a time series of the maximum weekly NDVI measurements over five years is formed for each pixel. Medians are calculated over a d week moving window of this time series, and are denoted as $NDVI(t, s)$ where t is an index for the d week window, and s is a pixel index. The value d is determined based on the growth season for the crop being targeted. The maximum NDVI for each of these medians is recorded as $NDVI^*(s)$.

The ratio of $NDVI^*(s)$ and the 95% quantile of $NDVI(s)$ for all s in the WELD tile are used to create a heuristic to determine the likelihood of an agricultural field. If the ratio,

$P^*(s)$, is greater than one then the pixel is considered to be part of an agricultural field, and if the ratio is below a threshold it may be a field. The paper uses the term "probability" for $P^*(s)$ but it is unclear what underlying random process is generating this probability, and if this probability actually reflects the probability that a given pixel belongs to an agricultural field.

Another function, $P(s)$, is label the probability that the pixel at location s is an edge. This function provides some degree of edge detection through the function $f(x, s_0)$,

$$f(x, s_0) = \frac{\sum_{s \in B} ||x(s_0) - x(s)||_2 ||s_0 - s||_2}{\sum_{s \in B} ||s_0 - s||_2} \tag{2.18}$$

where B is the set of pixels that are adjacent to s_0 (including corners). $P(s)$ is the average over all weeks of

$$P(s) = \sum_{k=1}^{52} \frac{NDVI(t, s) f_2(\rho(t, s), s) f(NDVI(t, s), s)}{52} \tag{2.19}$$

where $\rho(t, s)$ is calculated in a similar method to $NDVI(t, s)$, but uses bands 2,3,4,5 and 7 of the WELD data.

Segmentation in this method uses variational region based geometric active contour (VRGAC). VRGAC uses thresholding and an iterative procedure defined in Caselles et al. (1993). In Yan and Roy (2014), a fixed value threshold of 0.85 is applied to each $P^*(s)$ to form initial agricultural fields, where values below the threshold are discarded. It is stated that the threshold value is fairly insensitive over the range 0.5 to 0.9. VRGAC is then applied to the binary image obtained through the thresholding, where VRGAC applies smoothness constraints to the binary edges iteratively.

More post processing is done to handle under-segmentation. The post processing includes watershed segmentation on the distances obtained through VRGAC. A "skeleton" is then formed by identifying minima within each watershed segment and ridges between

these minima, then a heuristic is used to determine if the ridges were likely between two distinct fields or within the same field.

The most interesting part of this method occurs through the use of thresholds to determine agricultural fields that use center pivot irrigation (circular fields). The algorithm uses the fit statistic

$$\text{fit}(i,j,l) = \frac{\sum_{r=1}^{k} n_r^0 + \sum_{r=1}^{k} n_r^c}{\sum_{r=1}^{k} n_r + \sum_{r=1}^{k} n_r^0} \tag{2.20}$$

where

- s is a point within the initial field,
- l is the length of the rays,
- n_r is number of pixels along ray r within the field,
- n_r^0 is number of pixels along ray r that are not within the field and
- n_r^c is difference between n_r and the $\max_{r\in\{1,\ldots,k\}} n_r$.

The fit statistic is minimized over a set of lengths and all points sufficiently within the interior of the initial field. If the fit statistic is below a pre-determined threshold then the rays are used to form a set of "pie slice" shaped fields. These slices are merged through another fit statistic

$$\text{fit} = \frac{\sum_{r=1}^{k} n_r^c}{\sum_{r=1}^{k} n_r + \sum_{r=1}^{k} n_r^0} \tag{2.21}$$

and another threshold. The final result uses a two pixel dilation and a one pixel erosion to smooth the final boundaries.

Zhang (2009) used wavelet transforms with watershed segmentation to identify agricultural images. This work differs from Yan and Roy (2014) in that it does not use NDVI, and instead it uses high resolution imagery (2.4m^2). Since high resolution imagery is used as opposed to NDVI a finite difference approximation of the magnitude of the gradient of a grayscale converted image is used. This finite difference approximation is used with

watershed segmentation and a number of post-processing steps to produce a segmentation.

Application of the mean shift algorithm to remotely sensed data include Huang and Zhang (2008), Büschenfeld and Ostermann (2012) and Friedman et al. (2013). An interesting approach of using image segmentation and classification to jointly minimize over-segmentation and improve the classification was presented by Büschenfeld and Ostermann (2012). In this application mean shift was used for an initial segmentation, then pixels classifications were assigned to pixels within each segment from a separate SVM (support vector machines) classification process. A majority vote then determined the classification of the interior pixels. This approach is somewhat similar to the approach taken in the classification section of this dissertation. However, in this dissertation a test is performed to determine if disagreement between pixels within each segment are due to miss-classification, or miss-segmentation. Friedman et al. (2013) performed a short exploration of the mean shift algorithm from Georgescu et al. (2003), applying it to remotely sensed data in a distributed environment.

2.2 Classification

Classification methods for parcel-level spatial units associate labels with specific land cover units. These labels are from GIS sources such as remotely sensed imagery or digitized thematic maps, where thematic maps are geographic maps with categorical values. Assuming that the spatial unit boundaries are reasonably accurate and larger than the pixels used for classification, the classification of spatial units may improve the classification rate of the original image (see Gao et al., 2007). This increase in classification rate is due to more information being available per unit area relative to pixel based classifications. The classified pixels may also be useful in identifying potential problems with the boundaries, identifying areas where land cover units should be split or merged.

2.2.1 Classification of Land Cover Units

Literature on the classification of imagery to known boundaries can be found in the European Union and in the United States. The European Union's Land Parcel Identification System (LPIS), is a geospatial database that retains data on every farm that participates in crop payments programs (see Taşdemir and Wirnhardt, 2012). The LPIS database is updated via the farmer, through a web application (see Taşdemir et al., 2012), and each member state is required to maintain quality standards. Oesterle and Wildmann (2004) classified this land parcel system via "fuzzy methods", namely identification through a hierarchical object based system. These fuzzy methods, assign a probability of a given land use type.

In the United States, mandatory agricultural reporting is done via the Agricultural Census every seven years, where no geospatial data is obtained; geospatial data with regards to parcel-level spatial units is collected by the Farm Service Agency (FSA), but participation is voluntary leading to under-coverage. Despite the incompleteness of the segmentation, FSA spatial units known as common land units (CLU) have been used in the literature. One example is in California where FSA CLUs in conjunction with CDL data and other multi-source products were used to classify agricultural content with up to 84% accuracy (see Falkowski and Manning, 2010).

Long et al. (2014) and Kipka et al. (2016) also applied CDL data to parcel-level spatial units. The spatial units used in Long et al. (2014) were from Montana State Library's Geographic Clearinghouse from June 2012, and Kipka et al. (2016) used individual farm with mapped GIS boundaries. For Long et al. (2014) the most dominant crop was used to determine land cover of the spatial unit. Kipka et al. (2016) determines dominant crop by multiplying the acres classified within a spatial unit times the classification rate. The crop with largest value is then selected as the land cover for the parcel.

In the United Kingdom, (Woodwalton Fen, Cambridgeshire) Dean and Smith (2003) used a multivariate normal likelihood function to classify land-uses against known class

parameters. An entropy based measure was further used to measure the quality of the estimate. Another land parcel system was commissioned by the States of Jersey in the United Kingdom for the Institute of Terrestrial Ecology (see Smith and Fuller, 2001). This system integrated multi-source geospatial imagery with additional data, and used a maximum a posteriori probability (MAP) estimate of land cover (see Fuller et al., 1994). Details of the implementation of this method unfortunately are not cited in the literature.

2.2.2 Under-Segmentation

Splitting of under-segmented spatial units is addressed in a small number of papers in remote sensing. The approaches to identification of under-segmentation in these papers are varied. Yan and Roy (2014) and Turker and Kok (2013) take the form of deterministic ad-hoc approaches, while Johnson and Xie (2011) uses a measure of intra- and inter-pixel homogeneity within each segment to determine under-segmentation.

A method of joint under- and over-segmentation is addressed in Yan and Roy (2014), where watershed segmentation is assisted through thresholding. Watershed segmentation works by identifying convex subsets of a smoothed grayscale image. In Yan and Roy (2014), instead of grayscale images a function of the NDVI is used. The thresholding is based on the largest value between local minima along the "spine" or trough in the image. Only adjacent watershed segments with maximum values below this threshold are merged.

Turker and Kok (2013) uses a rule based approach to identification of under-segmentation through "perceptual grouping." This method uses an edge detection method known as "Canny edge detection" (see Canny, 1986) to identify edge pixels, then uses rule matching to form pixels within a spatial unit into "natural" edges.

Johnson and Xie (2011) used the statistic

$$H = \frac{n\tilde{S}^2 - \tilde{I}}{n\tilde{S}^2 + \tilde{I}} \tag{2.22}$$

where $\tilde{S}^2$ and $\tilde{I}$ are the variance and Moran's I normalized under each spectral band. This statistic measures inter-pixel homogeneity through Moran's I, and intra-pixel homogeneity through the variance estimate. Values with high variance and high Moran's I were identified with segments containing more than one land cover. A formal test was not specified, instead a set of thresholds on H were set, and re-segmentation was performed on segments above this threshold.

In statistical literature the problem of under-segmentation, is in the scope of cluster detection. Waller and Gotway (2004) notes that the existing global indices do not have sufficient power for determining the existence of a single cluster. Identification of a small number of clusters, instead falls under visual exploratory analysis through local indicators of spatial association (LISA). The most popular LISA is the local version of Moran's I. It can be found in practice in numerous papers in ecology, health, and population growth. The properties of this statistic can be found in most standard spatial statistics texts such as Schabenberger and Gotway (2004), Cressie (1993), and Cressie and Wikle (2011).

The statement in Waller and Gotway (2004) that existing global indices do not have sufficient power for determining the existence of a single cluster holds in the general case, but may not hold in specific cases. The specific case of interest is with binary response with reasonably low error rates, in this case global tests may provide reasonable power. In this dissertation a global test applied to the pixels within each LCU will be used.

The binary case admits two approaches, the first approach follows from spatial point patterns, the second from areal models. Spatial point patterns and areal models are both classes of statistical models defined by Noel Cressie in the seminal text Cressie (1993). Spatial point patterns are spatial models based on the distribution of points in the spatial support. Areal models, sometimes called lattice models, are spatial models based on a partitioning of the spatial support into spatial units.

Popular spatial point tests include quadrat methods, count based approaches to test for spatial uniformity. The basic approach is to divide the spatial support into a set of quadrants and compare counts between regions. The usual test statistic for these tests is

Pearson's χ^2 statistic

$$\boldsymbol{\chi}^2 = \sum_{i=1}^{n} \frac{(\mathbf{x}_i - \mathbb{E}[\mathbf{x_i}])^2}{\mathbb{E}[\mathbf{x_i}]} \tag{2.23}$$

where under the null hypothesis each $\mathbf{x}_i$, the number of points in quadrant i, is an iid random variable from a uniform distribution.

A similar approach in the class of areal models is the popular global test Moran's I (2.24). Moran's I can be viewed as a moving average approach to the Pearson's test, and the relationship between Moran's I and Pearson's $\boldsymbol{\chi}^2$ has been investigated by Rogerson (1999).

$$I = \frac{n}{\sum_{s_1 \in R} \sum_{s_2 \in R} w(s_1, s_2)} \frac{\sum_{(s_1 \in R)} \sum_{(s_1 \in R)} w(s_1, s_2) \left(x(s_1) - \bar{x}\right) \left(x(s_2) - \bar{x}\right)}{\sum_{s_1 \in R} \left(x(s_1) - \bar{x}\right)^2} \tag{2.24}$$

The key component of Moran's I is the interior term in the numerator,

$$\sum_{(s_1 \in R)} w(s_1, s_2) \left(x(s_1) - \bar{x}\right) \left(x(s_2) - \bar{x}\right). \tag{2.25}$$

This term is large under positive spatial autocorrelation, and small under negative spatial correlation.

Moran's I can be used for binary data, where it is quite similar to the Black-Black join count statistic Cliff and Ord (1981),

$$J_{BB} = \frac{1}{2} \sum_{(s_1 \in R)} \sum_{(s_1 \in R)} x(s_1) x(s_2) w(s_1, s_2). \tag{2.26}$$

The Black-Black join count statistic, is used under either assumptions of binary error, or under a fixed number of positive outcomes. Therefore, the statistic is well suited for testing spatial dispersion. Only the Black-Black joint count statistic is considered in this

dissertation due to its direct applicability to binary areal data (binary valued pixels).

2.2.3 Identification of Non-Agricultural Land Cover Units

Classification of agriculture is a topic addressed in great depth in a large number of papers such as Boryan et al. (2011) for the CDL. The approach in these papers is based on reflectance of different spectral wavelengths over time. However, the pixels provided by the CDL and Landsat derived classifications provide pixels of $30m^2$, too large to identify roads, creeks, and other small or narrow geographic features.

Identification of roads through remotely sensed imagery is a fairly well established area of research. Zhang and Couloigner (2006) provides a good overview and classification of methods used to identify roads from remotely sensed images. The quality of the methods vary, with the SVM method of Song and Civco (2004) providing fairly reasonable results and the Canny edge detection based method of Sharma et al. (2013) providing mixed results.

In the United States and most industrial nations, GIS information about roads and hydrography has been identified and made available through GIS databases. Therefore, the focus of this dissertation is on handling smaller features such as paths between fields. No literature has been found explicitly dealing with the identification of small non-agricultural land uses.

2.2.4 Over-Segmentation

Unlike identification of under-segmentation, identification of over-segmentation is a popular topic in image segmentation literature. This popularity is strongly associated with the popularity of the watershed segmentation method. An example of merging in watershed segmentation can be found in Yan and Roy (2014), as described in Section 2.2.2, and Zhang (2009).

Merging of U.S. Census tracts or "regionalization" was performed by Spielman et al. (2014), using optimization techniques to minimize the coefficient of variation (CV). The CV is a unit-less measure equal to the standard deviation of an estimate divided by the

estimate. The method employed, although not stated, is an exchange based method where U.S. Census tracts above a CV threshold are randomly sampled. A sampled tract and its adjacent tracts are merged "one-by-one" until either the merged units meet the CV threshold for a particular item of interest, or all the tracts adjacent to the sampled tract have been merged.

Johnson and Xie (2011) used the H statistic described in 2.2.2 to identify the least homogeneous regions. From least-to-most homogeneous below a threshold on H, attempts to maximize H by merging using mean intensity difference over spectral bands.

In this dissertation, merging over-segmented pixels is accomplished through the use of exogenous pre-classified pixels, such as CDL data. Unfortunately, no paper in the GIS or image segmentation literature was found on this topic.

2.3 Prediction

In this dissertation, a spatial-temporal multinomial probit model is used to predict agricultural land cover through crop rotations. Estimation of parameters, and prediction of land use is done under a hierarchical Bayesian framework. In this section, the necessary theory and relevant literature are introduced, including an overview of the multinomial probit model, the simultaneous autoregressive model, and spatial-temporal simultaneous autoregressive model. The focus of this theory will be largely on the problem of unidentified parameters, and the conditional distributions required for gibbs sampling.

2.3.1 Multinomial Probit Model

As shown in Section 1.3, multinomial probit models provide a link function between a categorical response and a $J = C - 1$ dimensional linear model with a multivariate normal error structure, where C is the number of classes being modeled. This linking is performed

by partitioning the support of the linear model Figure 1.2,

$$\mathbf{y}_i = \begin{cases} c = 1 & \max\left(\mathbf{z}_{i,j}\right) \leq 0 \qquad\qquad\qquad\qquad \forall j \in \{1, \cdots J\} \\ c = j+1 & \text{argmax}_{j \in \{1,\cdots,J\}} \left\{ \mathbf{z}_{i,j} \mathbb{I}_{\{\mathbf{z}_{i,j} > 0\}} \right\}) \end{cases} \tag{2.27}$$

where the observation of category $c \in \{1, \cdots, C\}$ is determined by $z_{i,j}$ is the j^{th} element of the latent vector z_i with $j \in \{1, \ldots, J\}$ (see McCulloch and Rossi, 1994). $j = 1$ is referred to as the base class or the default choice; in MNP modeling covariates based on choice are defined relative to the base class. Inference in multinomial probit models is done primarily under a Bayesian paradigm using data augmentation through gibbs sampling.

Data augmentation is a method to simulate and model unobserved phenomena through a latent variable. Tanner and Wong (1987) popularized data augmentation under the Bayesian paradigm, with its application in calculating posterior distributions. Albert and Chib (1993) applied this methodology to probit models, in this application the latent variable $\mathbf{z}$ is simulated under a truncated normal distribution, with a truncation point determined by the observed state of $\mathbf{y}$. Multivariate extensions of this approach can be found in McCulloch and Rossi (1994), Nobile (1998), and McCulloch et al. (2000).

There are some notable issues with approach in MNP models, namely the linear model parameters are not identifiable,

$$\Pr\left(\mathbf{z}_{i,j} > a\right) = \Pr\left(\alpha \mathbf{z}_{i,j} > \alpha a\right), \, \alpha \in \mathbb{R}. \tag{2.28}$$

An approach to fix this issue is offered in McCulloch et al. (2000), where under gibbs sampling, draws from $\mathbf{\Sigma}$ are done under an inverse Wishart distribution with the constraint that the fist element, $\sigma_{1,1} = 1$. This restriction fixes the issue of identifiability, but at the same time creates two other issues. The first issue is that the generation of deviates is more computationally difficult due to this restriction, and second this restriction increases the correlation between subsequent draws under gibbs sampling. The amount of correlation

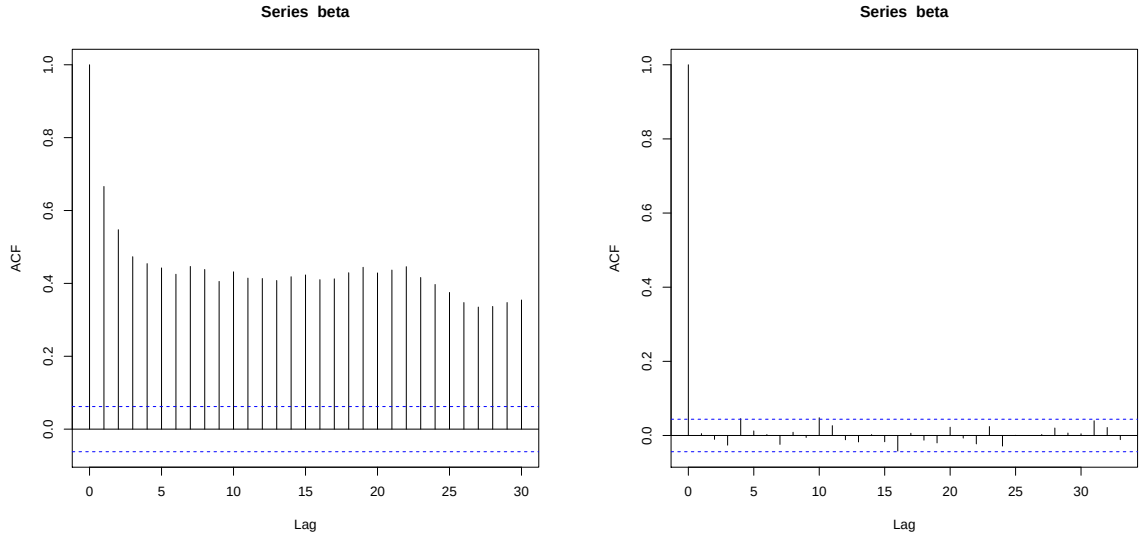

Figure 2.3: An example of extreme autocorrelation of a location parameter under McCulloch et al. (2000) (Left), compared to the autocorrelation for the same parameter under Imai and van Dyk (2005) (Right).

can be quite extreme (Figure 2.3). Burgette and Nordheim (2012) provides an alternative trace based restriction to provide identifiability, but uses an alternative data augmentation approach (see Imai and van Dyk, 2005).

An attempt to fix both issues introduced by McCulloch et al. (2000) is provided in Imai and van Dyk (2005), under the term marginal data augmentation. Marginal data augmentation, (see Meng and Van Dyk, 1999 and Van Dyk and Meng, 2001), improves the convergence speed of the data augmentation algorithm through the use of *working parameters*. The working parameters are unidentified parameters that are used to help generate less correlated draws under gibbs sampling, and can be easily integrated out (2.29).

$$\ell(\theta|y) \propto \int f(y,z|\theta)dz = \int \left(\int g(y,z|\theta,\alpha)h(\alpha|\theta)d\alpha \right) dz \tag{2.29}$$

The marginal data augmentation approach creates a random variable $\boldsymbol{\alpha}$ and places a distribution over it. The random variable $\boldsymbol{\alpha}$ is scalar valued, and is used to scale $\boldsymbol{\Sigma}$ such that

$\sigma_{1,1} = 1$ without explicitly generating deviates under the constraint. The term conditional data augmentation is used to describe the approach of McCulloch et al. (2000), where the constraint on $\mathbf{\Sigma}$ can be viewed as a conditioning the random variable on α.

Gibbs Sampling

The gibbs sampling approach in Imai and van Dyk (2005) and McCulloch and Rossi (1994) follows from the normal-inverse-Wishart (NIW) prior, common in Bayesian hierarchical models. In this section the conditional distributions under McCulloch and Rossi (1994) and the modifications to these distributions by Imai and van Dyk (2005) are presented.

The gibbs sampling schemes in MNP models are similar to those for Bayesian hierarchical models using multivariate normal distributions. The notable exception is that the response, $\mathbf{Z}^*$ is not-observed. $\mathbf{Z}^*$ is instead a latent variable and conditioned on the observed categorical response Y, (2.27). Gibbs sampling requires that the conditional distributions are known, e.g. $f(\beta|\Theta_{-\beta})$, where Θ_{-c} is the set of parameters being estimated, excluding the parameter c. The set of parameters of interest under McCulloch and Rossi (1994) are Z, β, and Σ.

The NIW prior is a prior on the joint distribution of $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$, and can be rewritten as a prior on $\beta|\Sigma$ and Σ,

$$\boldsymbol{\beta}|\Sigma \sim \mathcal{N}(\beta_0, \Sigma_0) \tag{2.30}$$

and

$$\mathbf{\Sigma} \sim \mathcal{W}^{-1}(\nu, S) \tag{2.31}$$

where Σ_0, ν, and S are all hyper parameters. Under the NIW conjugate prior the likelihood

follows,

$$\begin{aligned} P(\mathbf{\Theta}, \mathbf{Z}^*|Y, A, B) &\propto \exp\Big((Z^* - U^*\beta)^T I_n \otimes \Sigma^{-1} (Z^* - U^*\beta) + (\beta - \beta_0)\, \Sigma_0\, (\beta - \beta_0) \\ &+ \log|\Sigma|(\nu + J + 1) + \mathrm{tr}\left(S\Sigma^{-1}\right)\Big). \end{aligned} \tag{2.32}$$

The conditional distributions can then be easily calculated through the terms in the exponent:

$$\begin{aligned} & (Z^* - U^*\beta)^T I_n \otimes \Sigma^{-1} (Z^* - U^*\beta) + (\beta - \beta_0)\, \Sigma_0\, (\beta - \beta_0) \\ + \;& \log|\Sigma|(\nu + J + 1) + \mathrm{tr}\left(S\Sigma^{-1}\right) \\ = \;& \beta^T \left((U^*)^T I_n \otimes \Sigma^{-1} U^* + \Sigma_0\right)\beta - 2\beta^T \left(I_n \otimes \Sigma^{-1}(U^*)^T + \Sigma_0\beta_0\right) \\ + \;& \left((U^*)^T + \Sigma_0\beta_0\right)^T \left((U^*)^T I_n \otimes \Sigma^{-1} U^* + \Sigma_0\right)^{-1} \left((U^*)^T + \Sigma_0\beta_0\right) \\ + \;& (Z^*)^T I_n \otimes \Sigma^{-1} Z^* + \beta_0^T \Sigma_0 \beta_0 \\ - \;& \left((U^*)^T + \Sigma_0\beta_0\right)^T \left((U^*)^T I_n \otimes \Sigma^{-1} U^* + \Sigma_0\right)^{-1} \left((U^*)^T + \Sigma_0\beta_0\right) \\ + \;& \log|\Sigma|(\nu + J + 1) + \mathrm{tr}\left(S\Sigma^{-1}\right) \end{aligned} \tag{2.33}$$

Terms in blue are directly marginalized out through a normal distribution, and terms in red are marginalized out through the inverse Wishart. The remaining terms form a multivariate normal distribution,

$$\begin{aligned} \boldsymbol{\beta}|\Theta_{-\beta}, Y, A, B &\sim \mathcal{N}\Big(\left((U^*)^T I_n \otimes \Sigma^{-1} U^* + \Sigma_0\right)^{-1} \left((U^*)^T + \Sigma_0\beta_0\right), \\ &\qquad \left((U^*)^T I_n \otimes \Sigma^{-1} U^* + \Sigma_0\right)^{-1} \Big). \end{aligned} \tag{2.34}$$

The conditional distribution of $\mathbf{\Sigma}^{-1}$ is obtained in a similar way, with blue terms marginalized out,

$$(Z^* - U^*\beta)^T I_n \otimes \Sigma^{-1} (Z^* - U^*\beta) + (\beta - \beta_0) \Sigma_0 (\beta - \beta_0) + \log|\Sigma|(\nu + J + 1) + \text{tr}\left(S\Sigma^{-1}\right) \tag{2.35}$$

with distribution

$$\mathbf{\Sigma}^{-1}|\Theta_{-\Sigma^{-1}}, Y, A, B \sim \mathcal{W}^{-1}\left(\nu + J + 1, S + (Z^* - U^*\beta)^T (Z^* - U^*\beta)\right). \tag{2.36}$$

$\mathbf{Z}^*|\Theta_{-Z^*}$ has the conditional distribution equal to a truncated normal distribution,

$$\mathbf{Z}^*|\Theta_{-Z^*}, Y, A, B \sim \mathcal{TN}(\beta, \Sigma, A, B). \tag{2.37}$$

Deviates for Z^* are generated under a truncated normal distribution using a scheme detailed in McCulloch and Rossi (1994).

Under marginal data augmentation (1.6) is rewritten as

$$\boldsymbol{\alpha}\mathbf{Z}^*|Y = U^*\boldsymbol{\alpha}\beta^* + \boldsymbol{\alpha}\boldsymbol{\epsilon}; \tag{2.38}$$

following the notation of Imai and van Dyk (2005), $\boldsymbol{\alpha}\mathbf{Z}^*|Y = \tilde{\mathbf{Z}}$, $\boldsymbol{\alpha}^2\mathbf{\Sigma} = \tilde{\mathbf{\Sigma}}$, and $\boldsymbol{\alpha}\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$. The set of parameters without the tilde, are unidentified, and the parameters with a tilde are identified. The later is due to $\boldsymbol{\alpha}$ being used to constrain $\mathbf{\Sigma}$. Prior specifications are made to the unidentified parameters. $\boldsymbol{\beta}$ retains the same prior as before in McCulloch and Rossi (1994).

$$\beta|\Sigma\ \mathcal{N}(\beta_0, \Sigma_0). \tag{2.39}$$

A joint prior is specified on $(\boldsymbol{\alpha}, \mathbf{\Sigma})$,

$$f(\alpha, \Sigma) \propto |\Sigma|^{-(v+p)/2}(\alpha^2)^{-\nu(J-1)/2+1}\text{exp}\left(\frac{\alpha_0^2}{2\alpha^2}\text{tr}(S\Sigma^{-1})\right) \tag{2.40}$$

with hyper parameters α_0^2,ν, and S. The distribution of $(\boldsymbol{\alpha}|\Sigma)/(\sigma_0^2 \text{tr}(S\Sigma^{-1})$ follows an inverse χ^2 distribution, with $\nu(J-1)$ degrees of freedom, and

$$\boldsymbol{\Sigma} \sim \mathcal{W}^{-\infty}(\nu - J, S). \tag{2.41}$$

Imai and van Dyk (2005) provides two different sets of conditional distributions for gibbs sampling. Each set is used for different sampling schemes; in the first scheme the working parameter is completely marginalized out at each iteration. In the second their working parameter is not marginalized out, but instead is updated at each iteration. In both approaches the conditional distributions for gibbs sampling are given in terms of the identified parameters.

2.3.2 Spatial Autoregressive Models

Cressie (1993) defines three types of spatial models by domain, geostatistical, lattice, and point processes. Within this dissertation the domain, the set of ALCUs, is of the second type, lattice. Lattice is also referred to areal, and the term areal is used in this dissertation. This domain consists of a finite, and disjoint set of indexed ALCUs. Relationships between these ALCUs can be modeled via graphs, where ξ indexes an ALCU, and edges between ALCUs are used to identify relationships for spatial stochastic processes.

The spatial autoregressive Gaussian model used in this dissertation is the, simultaneous autoregressive (SAR) model. This model was introduced in Section 1.3, and its form will be repeated here for convenience.

$$\mathbf{Z} = B\mathbf{Z} + (I - B)U\beta + \boldsymbol{\epsilon}. \tag{2.42}$$

Parameter estimation in SAR models can be performed under least squares, maximum likelihood, or Bayesian-based approaches. There are a number of issues with parameter estimation under all these approaches, but for the purposes of this dissertation only the Bayesian approach is reviewed.

The likelihood of the conditional distribution is

$$\begin{aligned} P(\beta, \Sigma, \rho|Z) &\propto \exp\Big(((Z - U\beta)^T \left((I - B)^{-1}\Sigma^{-1}(I - B^T)^{-1}\right)^{-1} ((Z - U\beta) \\ &+ (\beta - \beta_0)\,\Sigma_0\,(\beta - \beta_0) \\ &+ \log|\Sigma|(\nu + n + 1) + \mathrm{tr}\left(S\Sigma^{-1}\right)\Big) \\ &\cdot \mathbb{I}_{\{\rho\in(a,b)\}} \end{aligned} \tag{2.43}$$

using the standard NIW priors with a uniform prior on ρ over the interval (a, b). In practice, the interval (a, b) is defined such that any ρ in the interval ensures that $B = I - \rho W$ is positive definite. LeSage and Pace (2009) suggests the interval $(0, 1)$ when a positive autocorrelation is assumed and W is row stochastic. Where row stochastic implies that each element in W is greater than or equal to zero, and each row sums to one.

The most computationally burdensome part of the Bayesian approach is in the drawing from the conditional distribution of ρ given β, Σ. Since the distribution of ρ does not follow any standard form, deviates are generated either by interpolation Metropolis-Hastings, or through interpolation and inversion of the CDF (see LeSage and Pace, 2009). In either approach, the calculation of the determinant of $I - \rho W$ contributes the most to this computational burden. In SAR models with sparse W this calculation can be accelerated using sparse matrix handling techniques (see Bivand, 2015).

2.3.3 Spatial Autoregressive Multinomial Probit Models

Spatial autoregressive MNP models or SAR MNP models were introduced in Section 1.3, and the latent model specification is repeated here for convenience.

$$\mathbf{Z}^* = B\mathbf{Z}^* + U\beta + \boldsymbol{\epsilon}, \tag{2.44}$$

where

$$
\begin{aligned}
B &= \rho W \otimes I_J, \\
W &= n \times n \text{ matrix of spatial weights (“rook” based in this application), and} \\
\rho &= \text{scalar parameter for } W.
\end{aligned}
$$

Spatial-temporal SAR multinomial probit models have been studied by Wang and Kockelman (2009) and Wang et al. (2012). These models, abbreviated STAR MNP are less prominent in the literature, and are largely derivatives of LeSage and Pace (2009). The general form of Wang et al. (2012) follows from the likelihood,

$$
\begin{aligned}
P(\beta, \Sigma, \rho, \lambda, Z^{**}|Y^{**}) \propto\ & \exp\big(\\
& ((Z^{**} - U^{*}\beta)^{T} \left((I-B)^{-1}(\Sigma \otimes \Omega)(I-B^{T})^{-1}\right)^{-1} ((Z^{**} - U^{*}\beta) \\
& + (\beta - \beta_0)\,\Sigma_0\,(\beta - \beta_0) \\
& + \log|\Sigma|(\nu + J + 1) + \operatorname{tr}\left(S\Sigma^{-1}\right)\big) \\
& \cdot \mathbb{I}_{\{\rho \in (a_1, b_1)\}} \\
& \cdot \mathbb{I}_{\{\lambda \in (-1,1)\}}
\end{aligned}
\tag{2.45}
$$

where Z^{**} is similar to the vector Z^{*} except it is stacked for each year in the model, likewise Y^{**} has the observed classes for each year in the model, and Ω is a temporal autoregressive covariance matrix. The parameter λ is the coefficient for the autoregressive model. Similar to $\boldsymbol{\rho}$, $\boldsymbol{\lambda}$ also has a uniform prior.

The model in Wang and Kockelman (2009) and Wang et al. (2012), follow the identified MNP specification of Nobile (2000). The MNP specification of Nobile (2000) is very similar to McCulloch et al. (2000), and both use conditional data augmentation and retain the same conditional distributions.

Wang et al. (2012) noted a large number of computational and theoretical issues with their model. The first issue is that with just 100 spatial units on a simulated data set, using four classes simulated over four years, the run time for a single iteration was 5 seconds on

a 2.66GHz PC. As noted by the authors, increasing the number of observations, tends to cause an exponential increase in run time (see LeSage and Pace, 2009), making this model unsuitable for even moderate population sizes.

The second, and most pressing issue is that the parameters in the gibbs sampling diverged. It is unclear in Wang et al. (2012), if the results were due to the model, the nature of the simulated data set, or an error in programming.

Chapter 3: Segmentation

This chapter will cover four themes: the first theme is the establishment of a well defined spatial-temporal land cover unit; the second theme provides details and analysis of the log variance filter; the third theme is the acceleration of the mean shift algorithm; the fourth and final theme is the application of the mean shift algorithm to high resolution imagery. The results of this chapter provide the land cover unit boundaries for the next chapter on land cover classification.

3.1 Land Cover Unit

The observational unit will be called a *land cover unit* (LCU). Each LCU is the maximally contiguous section of land with respect to a single land cover sequence not transected by public transportation arteries or permanent hydrographic boundaries, where a land cover sequence is an ordered set of known categories indexed by a set of fixed consecutive years. Under this description and later formal definition, an LCU is dependent on both the temporal window and the categories specified, consider the land cover in Figure 3.1.

LCUs are identified by two components a spatial index $\xi \in \Delta$, where Δ is the set of LCU indexes in region $R \subset \mathbb{R}^2$, and $y(\xi)$ a land cover sequence (vector valued) for the indexed LCU. Similar to a pixel, each ξ identifies a subset of $\mathbb{R}^2$, however the LCUs are not

Table 3.1: Land cover sequence of LCUs over 12 years in Iowa through CDL pixels (Left-to-Right, Top-to-Bottom, from Figure 3.1). In this sequence C = Corn and S = Soybeans.

LCU	Land Cover Sequence
Upper	CSC SCS CSC SCS
Lower	SCS CSC CSC SCS

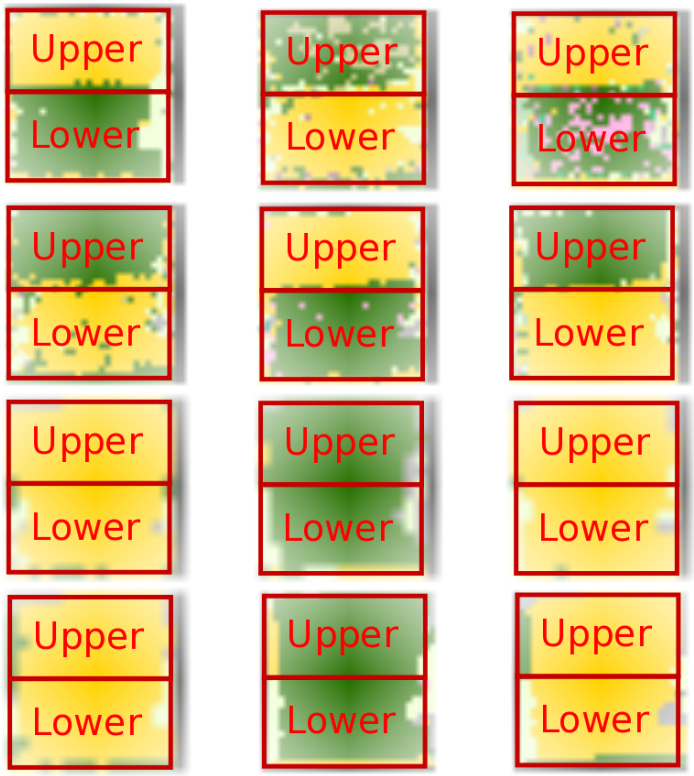

Figure 3.1: An example of two LCUs over 12 years in Iowa through CDL pixels (green soybeans, yellow corn).

necessarily uniform in shape. The land cover sequences will be identified by $c \in \{1, \cdots, C\}$, a set of land cover categories, and a sequence of years $\{1, \cdots, T\}$. As an example, an LCU $y(\xi)$ could rotate between land cover classes for five years producing the sequence $(1, 2, 1, 2, 1)$ where ξ indexes a subset of the product space $R \times \{1, \cdots, C\}^T$.

To ensure that land cover units are well defined the following assumptions are made:

1. Open sets will be defined in $\mathbb{R}^2$ by the typical ϵ-neighborhood definition.

2. The measure space $(\mathbb{R}^2, \mathcal{B}^2, \lambda)$ will be used, where $\mathcal{B}^2$ is the sigma algebra of all open subsets of $\mathbb{R}^2$, and λ is the Lebesgue measure.

3. $\nu(a)$ where a is an index for an LCU or a pixel will be the function that returns the subset of $\mathbb{R}^2$ for the indexed unit.

4. $\lambda(a)$ where a is an index for an LCU or a pixel will be used as shorthand for the Lebesgue measure on the subset of $\mathbb{R}^2$ covered by the indexed unit.

5. Each LCU is assumed to be aligned to a set of n pixels, x indexed by s, each with a land cover sequence $x(s) \in C$, and can be represented in a binary matrix G of dimension $n \times \#\Delta$, where $\#\Delta$ is the cardinality of Δ and 1 represents the assignment of a pixel to an LCU. This approach is reasonable given the application and the 1m^2 resolution of the imagery.

Given these assumptions any LCU, $y(\xi)$, must also be a member of a maximally geographically connected subset of R.

Definition 3.1 (Geographically Connected Subset of R)**.** A geographically connected subset of R is any union of finite connected closed sets in R with the same set of land cover sequences. The set of all geographically connected subsets or R is

$$\mathcal{G} = \left\{ A : A = \bigcup_{s \in B} \nu(s) \; B = \left\{ s : x(s) = x(s'), |s_1 - s_1'| \leq 1, |s_2 - s_2'| \leq 1 \right\} \right\}. \tag{3.1}$$

Definition 3.2 (Maximally Geographically Connected Subset of R)**.** A maximal geographically connected subset or R, is the finite union of closed sets in R that has the greatest measure under λ. The set of all maximal geographical connected subsets of R is

$$\mathcal{G}_{\max} = \{A \in \mathcal{G} : \lambda(A) \geq \lambda(A') > 0, \lambda(A \cap A') > 0 \; \forall A' \in \mathcal{G}\}. \tag{3.2}$$

3.2 Log Variance Filter

Variation in the intensity of the remote sensing image pixels can be problematic for estimating LCU boundaries. In particular, trees and other land cover exhibit pronounced but potentially irregular texture (see Figure 3.11). In this dissertation, an approach using the log of the local sample variance is presented as an alternative to more complicated filtering

approaches. This approach takes the form

$$\hat{f}_h(s) = \sum_{s^* \in D} \frac{1}{nh^*} \kappa_{h^*} \left((h^*)^{-1} \left(\log \left(\mathbf{S}_h^2 \right)(s) - \log \left(\mathbf{S}_h^2 \right)(s^*) \right)^2 \right), \; s, s^* \in D \tag{3.3}$$

where $\mathbf{S}_h^2(s,t)$ is as defined in (2.5), repeated here for convenience,

$$\mathbf{S}_h^2(s,t) = \sum_{s^* \in D} \sum_{s^{**} \in D} w_h(s,s^*) w_h(s,s^{**}) \frac{\left(\mathbf{x}(s^*) - \mathbf{x}(s^{**}) \right)^2}{2n(n-1)}. \tag{3.4}$$

κ_h is a standard kernel, such as a normal density, with bandwidth h, and $w_h(a,b)$ is a spatial weight between two points a and b from a kernel density estimate with bandwidth h. For isometric kernels, $w_h(a,b)$ can be rewritten as $w_h(\delta)$ where $\delta = |a-b|$. The filtered estimate will be referred to as the log variance filter (LVF). The two bandwidth values h^* and h are respectfully associated with the kernel density estimator applied to the log of the local variance, and the local sample variance. When s and s^* are in the same LCU, the LVF should be close to $1/h^*$. If s and s^* are in different LCUs or along the border the LVF should decrease in value.

3.2.1 Properties of the Log Variance Filter

In this application the LVF is used on regions of constant variance and mean, LCUs, and on steps between regions, LCU edges. As can be seen in (3.3), the LVF is a kernel applied to the pairwise difference of the logs of the local variances. Linear functions of the LVF, such as the expectation, can be examined through the properties of each pairwise difference,

$$\log \left(\mathbf{S}_h^2 \right)(s) - \log \left(\mathbf{S}_h^2 \right)(s^*) \tag{3.5}$$

In this dissertation, only the expected value of the LVF will be examined. It is shown, that even the expectation does not have an easy to obtain closed form, and under certain

conditions numerical methods may need to be used.

Each pairwise difference consists of a pair of points, s and s^*. Each point is the center of a spatial neighborhood. The spatial neighborhood of the local variance at point s is the set of all spatial locations, $s^* \in D$, such that $w_h(s, s^*) > 0$. In this dissertation the spatial neighborhoods will be square in shape, containing h^2 pixels centered around s.

Each pair, has local neighborhoods that can occur either within the same LCU, two different LCUs, or between different LCUs. Each pair can also overlap to some degree, creating correlation.

The pairwise distance consists of the log of two local variances. The local variance for the spatial neighborhood centered around s includes a set of h^2 pixels. Each pixel, follows a normal distribution with mean $\mu(s)$ and variance $\sigma^2(s)$. It is assumed that each pixel within an LCU has the same mean and variance. This creates four different cases:

1. Equal correlated local sample variances from iid samples.
2. Unequal uncorrelated local sample variances from iid samples.
3. Unequal correlated local sample variances from independent samples.
4. Unequal uncorrelated local sample variances from independent samples.

The first case includes the case where the correlation may be 0 between samples.

I will now examine these four cases and provide a closed-form expression where possible. The expectation of these pairwise difference can then be used to determine the expectation of the LVF for location s.

Case 1: Equal Correlated Local Sample Variances from iid Samples.

Based on the assumption that each of the $\mathbf{x}(s)$ are iid with a normal distribution then it follows that $\bar{\mathbf{x}}(s)$ and $\bar{\mathbf{x}}(s+\delta)$ have correlation $\rho = \max\left\{\frac{h-|\delta|}{h}, 0\right\}$ for $\mathbb{R}$, and for $\mathbb{R}^2$ under a square support the correlation changes to $\rho = \max\left\{\frac{h^2-\delta^2}{h^2}, 0\right\}$.

Panaretos et al. (2005) examined the distribution of the ratio of two dependent gamma distributions formed from random samples of iid standard normal random variables. This distribution known as the correlated gamma ratio distribution (CGR), has the following pdf

$$f(y) = \frac{(1-\rho^2)^k}{\beta(k,k)} y^{k-1}(1+y)^{-2k}\left(1-\left(\frac{2\rho}{y+1}\right)^2 y\right)^{-\frac{2k+1}{2}}, \tag{3.6}$$

mean

$$\text{mean}(\mathbf{y}) = \frac{\left(1-\rho^2\right)+k-1}{k-1}, \tag{3.7}$$

and variance

$$\text{var}(\mathbf{y}) = \frac{(5k-4)\left(1-\rho^2\right)+(k-1)(k-2)\left(1-\rho^2\right)}{(k-1)^2(k-2)} \tag{3.8}$$

where $k = h^2$ under a square support. When $\rho = 0$ this distribution returns to a standard F distribution.

The log transformed CGR has the distribution

$$g(z) = \frac{\left(1-\rho^2\right)^k}{\beta(k,k)}\left(\frac{e^z}{(1+e^z)^2}\right)^k\left(1-4\rho^2\frac{e^z}{(1+e^z)^2}\right)^{-\frac{2k+1}{2}}. \tag{3.9}$$

The form of this distribution is unsurprisingly similar to that of Fisher's Z-distribution, that is proportional to the log of an F statistic. Plots of this distribution are remarkably close to a normal distribution for all sizes of k and ρ. Given the similarity between the distributions it is trivial to find the centered normal distribution that closest fits this distribution, Figure 3.2.

Since (3.9) is a bit unwieldy, a normal approximation can be used to simplify this expression. To do this a relationship between the log-CGR distribution and the normal distribution is needed. One approach is to use the delta method, and adjust for the number of independent observations. The delta method can be used to estimate the mean and

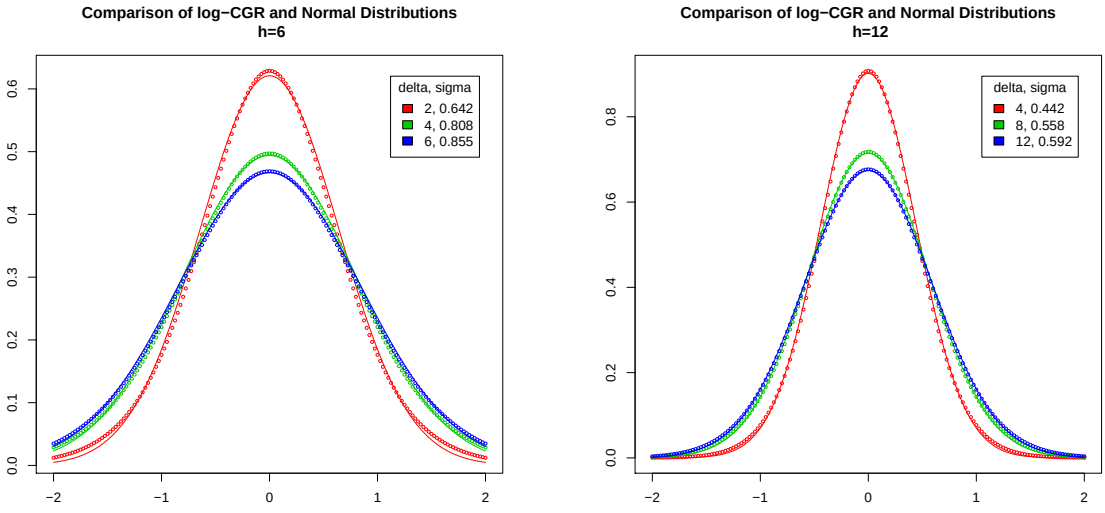

Figure 3.2: Plots of normal densities (lines) and log-CGR density at various bandwidth (h) and distances between points (δ).

variance of $\mathbf{y}$ using the identity

$$S^2 = (2n(n-1))^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2. \tag{3.10}$$

It is possible to calculate the mean and covariance of $\mathbf{y}$ with a bivariate distribution of two correlated χ^2 distributions, formed by two samples of h iid random variables with distribution $\mathcal{N}(0, \sigma^2)$, and η shared observations between the samples.

$$\mathbb{E}[y] = \begin{bmatrix} \sigma^2 \\ \sigma^2 \end{bmatrix} \tag{3.11}$$

$$\text{var}[y] = \begin{bmatrix} 1 & \frac{\eta^2(\eta-1)+(h-\eta)^2\eta+2\eta(h-\eta)(\eta-1)}{h^2(h-1)} \\ \frac{\eta^2(\eta-1)+(h-\eta)^2\eta+2\eta(h-\eta)(\eta-1)}{h^2(h-1)} & 1 \end{bmatrix} \frac{2\sigma^4}{h-1} = \Sigma \tag{3.12}$$

Central limit theorem can be used by growing the sample size h for a fixed $\omega = \eta/h$

$$\sqrt{2h(1-\omega)}\mathbf{y} \xrightarrow{d} \mathcal{N}(0, \Sigma_0) \tag{3.13}$$

where

$$\Sigma_0 = \begin{bmatrix} 1 & \omega \\ \omega & 1 \end{bmatrix} 4(1-\omega)\sigma^4. \tag{3.14}$$

The covariance term is a direct result of the relationship

$$\begin{array}{rcl} \frac{\eta^2(\eta-1)}{h^2(h-1)} + \frac{(h-\eta)^2\eta}{h^2(h-1)} + 2\frac{\eta(h-\eta)(\eta-1)}{h^2(h-1)} & = & \omega^3 + (1-\omega)^2\omega + 2\omega^2(1-\omega) + o\left(\frac{1}{h}\right) \\ & = & \omega + o\left(\frac{1}{h}\right) \end{array} \tag{3.15}$$

This provides a remarkably simple approximation via the delta method of

$$\mathbf{z} = \log(y_1) - \log(y_2) \sim \mathcal{N}\left(0, \frac{4(1+\omega)}{h}\right), \tag{3.16}$$

or for the square neighborhoods of interest

$$\mathbf{z} = \log(y_1) - \log(y_2) \sim \mathcal{N}\left(0, \frac{4(1+\omega^*)}{h^2}\right), \tag{3.17}$$

with $\omega^* = \eta/h^2$. The pairwise expectation of $\kappa_h(z)$ can then be used for the expectation of the LVF.

Case 2: Unequal Uncorrelated Local Sample Variances from iid Samples.

If $\mathbf{x}(s_1)$ is iid normal with mean μ_1 and variance σ_1^2 and $\mathbf{x}(s_2)$ is iid with mean μ_2 and variance σ_2^2, and $|s_1 - s_2| > h$, then by (3.16) the distribution of the log ratio of the local

sample variances $\mathbf{z}$ is

$$\mathbf{z} = \log(y_1) - \log(y_2) \sim \mathcal{N}\left(2\log(\sigma_2) - 2\log(\sigma_1), \frac{4}{h}\right). \tag{3.18}$$

Like case 1, the pairwise expectation of $\kappa_h(z)$ can then be used for the expectation of the LVF.

Case 3 and 4: Unequal Correlated Local Sample Variances from Independent Samples.

The last two cases do not have closed form solutions provided, and can not be provided in general due to the expectation and the variance not having closed forms when both the numerator and denominator follow non-central chi-squared distributions. Instead these results should be simulated when needed.

3.2.2 Log Variance Filter Simulation

To understand the utility of the LVF, a Monte Carlo experiment was conducted for the mean shift algorithm. This Monte Carlo experiment was conducted using two separate LCUs, both consisting of 5,000 pixels, with one shared boundary as seen in Figure 3.3. For this experiment three factors were used, the first factor with four levels, and the last two with three. The first factor was the mean of the second LCU μ_2 being set at 0, 1, 5, and 10; the first LCU has a constant location parameter of zero across all factors. The second and third factors are standard deviations of the first and second LCU respectfully with values 1, 5, and 10. Since both LCUs have the same size, only unique combinations of factors are retained.

The treatment of interest is the inclusion or exclusion of the LVF in the mean shift. The local variance in this experiment used a filtered estimated of the mean using a square uniform kernel centered around each pixel with an edge length of 11. Results are compared using the means of the adjusted Rand index (ARI) over all replicates for each distinct

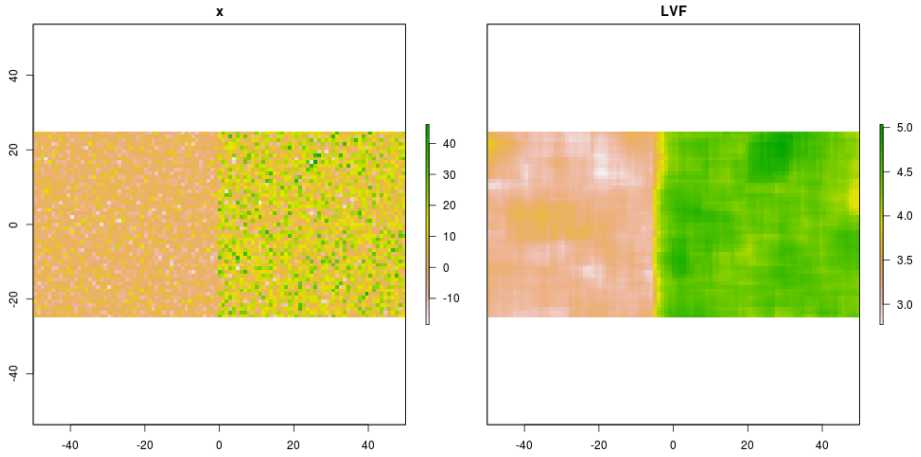

Figure 3.3: The simulated pixels (Left) and local log variance (Right) of two simulated LCUs. The left side of each image was simulated from $\mathcal{N}(0, 5)$, and the right side of each image was simulated from $\mathcal{N}(10, 10)$

combination of levels and treatment, indicated as $\hat{r}_1$ without an LVF component and $\hat{r}_2$ with. Each level is tested for 500 iterations, and the standard deviation for each $\hat{r}_\cdot$ was less than 0.001. Optimal bandwidth with respect to the ARI was determined using 100 iterations over a set of 5 possible bandwidth values for each dimension.

The results of this simulation are straightforward. In the presence of a difference in scale parameters without a substantial difference in location, the means shift with the LVF did much better than the standard mean shift. Likewise, when there was no difference in scale there was no benefit to the LVF. When there was a substantial difference in location, both methods faired equally.

3.3 Sampled Mean Shift

In high resolution images of LCUs, a large amount of the data is largely redundant. Due to this redundancy, a subset of the image may be sufficient to identify the local maxima of the density function of the image; this approach has been seen in Freedman and Kisilev (2009). Although a large amount of this data may be redundant, not all subsets of the pixels within the image will preserve the KDE's features equally well. A particular case

Table 3.2: Results of a Monte Carlo simulation for a variety of location and scale parameters for two simulated LCUs.

μ_1	μ_2	σ_1	σ_2	$\hat{r}_1$	$\hat{r}_2$
0	0	1	1	0.04	0.04
0	0	1	5	0.07	0.88
0	0	1	10	0.06	0.90
0	0	5	5	0.07	0.06
0	0	5	10	0.07	0.83
0	0	10	10	0.06	0.06
0	1	1	1	0.96	0.94
0	1	1	5	0.09	0.88
0	1	1	10	0.06	0.90
0	1	5	5	0.13	0.13
0	1	5	10	0.07	0.77
0	1	10	10	0.06	0.05
0	5	1	1	1.00	1.00
0	5	1	5	0.33	0.80
0	5	1	10	0.36	0.89
0	5	5	5	0.96	0.96
0	5	5	10	0.57	0.50
0	5	10	10	0.44	0.44
0	10	1	1	1.00	1.00
0	10	1	5	0.56	0.61
0	10	1	10	0.30	0.93
0	10	5	5	0.99	0.99
0	10	5	10	0.84	0.86
0	10	10	10	0.90	0.89

would be structures such as roads or paths between fields. These geographic features have little area, but are of high importance

Stratified sampling and interpolation can be used together to classify the entire image at much greater speed with little increase in error, relative to classifying all pixels within an image. Stratified sampling can select representative samples at a constant sampling rate, while ensuring pixels from roads and paths are sampled. Strata in this application are obtained by thresholding based on the LVF filter.

In this dissertation two strata are formed, the first includes units with large LVF values, areas of constant variance and location parameters. The second stratum includes areas with

low LVF values, associated with areas with non-constant variance or location parameters, e.g. roads and areas between fields. For simplicity, in the mean shift algorithm, the sampling rate ,$\pi(s)$, is constant in both strata.

In the high LVF stratum, samples are selected by choosing a pixel randomly within the image, then forming a grid of variable width and height about the image, per Breidt (1995). Variability in width and height is provided to minimize sample inefficiency (with respect to variance) that can occur when period of the sampling coincides with the period of natural phenomena. Because of the irregularity of the small LVF stratum, the local pivotal method was used, see Grafström et al. (2012).

The local pivotal method is a method to create spatially well distributed samples, and is of particular use when the area of interest is not rectangular. The most computationally efficient algorithm to perform the local pivotal method in the literature is LPM2. LPM2, as described in Grafström et al. (2012), has an average computational complexity of $\mathcal{O}(N^2)$.

The order of the computational complexity does not lend itself to high resolution imagery. In this dissertation the average computational complexity of this method was reduced to $\mathcal{O}(N \log(N))$ through a k-d tree.

The current implementation of LPM2 can be found in the *BalancedSampling* R package (see Grafström, 2014). This method doesn't employ k-d trees or other more efficient data structures. Therefore, an update to the algorithm has been provided through the use of k-d trees, and identified as LPM3, providing considerable performance improvement (see Table 3.3). Due to lack of a standard library implementation of k-d trees, a custom implementation with focus on efficient node deletion was created to support LPM3. The software is immediately available under an open source license, see Lisic (2015b).

Table 3.3: A comparison of LPM2 verses LPM3 for a one dimensional data set, timings are in elapsed seconds.

Population Size	Sample Size	LPM2 (s)	LPM3 (s)
100	10	0.001	0.002
1,000	100	0.002	0.002
10,000	1,000	0.163	0.014
100,000	10,000	19.715	0.255
1,000,000	100,000	9,271.350	8.806

3.4 Normal Newton Shift

For Gaussian kernels, there is a simple relationship between Newton's method (NM) and mean shift (MS) iterations; namely in the multivariate case for element k,

$$v_k^{(i+1)} = v_k^{(i)} - \frac{\sum_{j=1}^{N_R}(v_k^{(i)} - x_k)\phi\left(\left(v^{(i)} - x_j\right)^T H^{-2}\left(v^{(i)} - x_j\right)\right)}{\sum_{j=1}^{N_R}\left(1 - \left(h_k^{-1}(v_k^{(i)} - x_k)\right)^2\right)\phi\left(\left(v^{(i)} - x_j\right)^T H^{-2}\left(v^{(i)} - x_j\right)\right)} \tag{3.19}$$

can be rewritten as

$$v_k^{(i+1)} = v_k^{(i)} - \frac{\sum_{j=1}^{N_R}(v_k^{(i)} - x_k)\phi\left(\left(v^{(i)} - x_j\right)^T H^{-2}\left(v^{(i)} - x_j\right)\right)}{\sum_{j=1}^{N_R}\left(1 - \alpha^{(i)}\left(h_k^{-1}(v_k^{(i)} - x_k)\right)^2\right)\phi\left(\left(v^{(i)} - x_j\right)^T H^{-2}\left(v^{(i)} - x_j\right)\right)} \tag{3.20}$$

for $\alpha^{(i)} = 1$. Setting $\alpha^{(i)} = 0$ will produce the mean shift iterator. This relationship is missing in previous comparisons of mean shift and Newton's algorithm in Chiu et al. (2008), Fashing and Tomasi (2005) and Yang et al. (2003a), and will be referred to as NNS (Normal Newton Shift).

In application, the choice of α-sequence $\{\alpha_i\}_i^\infty$ should allow for convergence to the local maxima for all values in the support, while providing for accelerated convergence near the local maxima. Furthermore, the iterator for choice of α-sequence should be bijective over the support, this allows for a "memoryless" mapping from $X \to X$ independent of the

iteration index. For simplicity, only constant valued α-sequences, will be considered in this dissertation. This last requirement significantly simplifies the convergence proof.

A natural course of analysis is to determine the effect of $\left(h_k^{-1}(v_k^{(i)} - x_k)\right)^2$ in direction and step size. This term in the denominator allows for a number of differences between the mean shift algorithm and Newton's method. The first is that the denominator in the mean shift method determines the step size, but not the direction. The step size is a function of the sparsity of close observations; if observations are far from $v^{(i)}$ then the denominator is smaller and the step size is larger. For Newton's method this relationship isn't as clear, since $1 - \left(h_k^{-1}(v_k^{(i)} - x_k)\right)^2$ is smaller than the mean shift denominator, with the amount depending on the scale of x_k and the choice of bandwidth. A second observation is that each step in the Newton's method is not a linear combination of the observations, therefore the method is free to exit the support of the kernel density estimator. The third difference is that the denominator in the mean shift is constant for all elements of the vector being shifted, unlike the Newton's method denominator. This allows for considerably different paths between sequences under these two algorithms. An example of the paths taken for finding local maxima on a subset of Figure 3.11 is provided in Figure 3.4.

The convergence properties for NNS are a bit more difficult to come by then for either NM or MS. The major issues are that NNS is not monotonically increasing with respect to the KDE as in MS; NNS is also not directly related to a Taylor expansion around the local maxima as in NM; finally, it is not a linear operator as required by Newton-Kantorovich generalizations. Instead, a simple and somewhat restrictive proof of convergence of the sequence to a local maxima is provided via the Banach Fixed Point Theorem.

Theorem 3.1 (Normal Newton Shift Sequence Convergence)**.** If $\hat{f}$ is a KDE with a Gaussian kernel over N_R observations with density f:

- With a negative definite Hessian at v_0;
- A is a ball around a unique local maxima v_0 of diameter ϵ such that for any other

local maxima v_1 $\min\{H^{-1}(v_1 - v_0)\} > 2$;

- $\left(\frac{v - x_k}{h_k}\right) < 1$ for all $k \in \{1, \ldots, d\}$ and $v \in A$;
- $||\left(B(x) - B(y)\right)\hat{f}'(v_0)v_0|| \leq ||B(z)\hat{f}'(z)|| = \eta$ for all $x, y, z \in A$;

then for all $\alpha \in [0, 1]$ the NNS sequence converges to the local maxima v_0 for all points in A.

Proof. If $\alpha = 0$ convergence is provided by Chen et al. (2014a). If $\alpha = 1$ convergence is provided by Clarke and Futschik (2007).

For the vector valued function $\hat{f}$ equal to the derivative of the KDE at location v, the Taylor series around the point follows

$$\hat{\hat{f}}(v) = \hat{f}(v_0) + \left(\hat{f}'(v_0)\right)(v - v_0) + \Lambda \tag{3.21}$$

where Λ is a remainder of higher order terms. Due to the bandwidth constraints that H is diagonal the Hessian from Newton's method $\hat{f}'(x)^{-1}\hat{f}(x)$ is diagonal and so is $B(x)\hat{f}(x)$.

$$v - B(v)\hat{f}(v) = v - B(v)\hat{f}'(v_0)(v - v_0) - B(v)\Lambda. \tag{3.22}$$

Each diagonal element of the matrix $B(v)\hat{f}(v_0)$ has the form,

$$\frac{\sum_{j=1}^{N_R}\left(1 - \left(h_k^{-1}(v_{0,k} - x_{j,k})\right)^2\right)\phi\left(\left(v_0 - x_j\right)^T H^{-2}\left(v_0 - x_j\right)\right)}{\sum_{j=1}^{N_R}\left(1 - \alpha\left(h_k^{-1}(v_k - x_{j,k})\right)^2\right)\phi\left(\left(v - x_j\right)^T H^{-2}\left(v - x_j\right)\right)}. \tag{3.23}$$

Therefore, for a sufficiently large bandwidth such that $h_k^{-1}\left(v_k - x_j\right) < 1$ the numerator term is smaller then the denominator for all values of α.

From this

$$||B(x)\hat{f}(x) - B(y)\hat{f}(y)|| \leq \lambda||x - v_0 - y - v_0|| = \lambda||x - y|| \tag{3.24}$$

where $\lambda = \max\left\{\text{diag}\left(1 - B(v)\hat{f}'(v_0)\right) - \eta : v \in A\right\}$, the maximal diagonal element of $1 - B(v)\hat{f}'(v_0)$ plus η.

Therefore, the mapping in the ball about v_0 is a contraction, and by Banach fixed point theorem the fixed point iteration converges to v_0.

□

Performance and accuracy of this method were tested over a Gaussian mixture. The Gaussian mixture is generated from the convex combination of four bivariate normal distributions with means at $(-1, 1)$, $(1, 1)$, $(1, -1)$, and $(-1, -1)$; each distribution has covariance 0, and variance 0.25. MS and NNS are both performed on the KDE from 40 draws from each distribution, Figure 3.5.

The accuracy of the NNS is calculated on a per iteration basis by comparing the final local maxima of MS against each iteration of NNS through mean absolute error (MAE). In this comparison, the bandwidth is set to 0.75 in each direction for both MS and NNS. α for NNS was set to 0.1, 0.3, 0.5, and 0.7.

The results of this test are provided in Figure 3.6. Here it can be seen that small values of α, less than or equal to 0.3, can substantially decrease the number of iterations of the algorithm to arrive at the same MAE. When α was larger than 0.3, a small number of query points diverged, Figure 3.7. This caused the MAE to increase as the number of iterations increase. The sudden drops in MAE seen for all levels of α is due to the non-linear path taken by the steepest ascent algorithm to the local maxima 3.7.

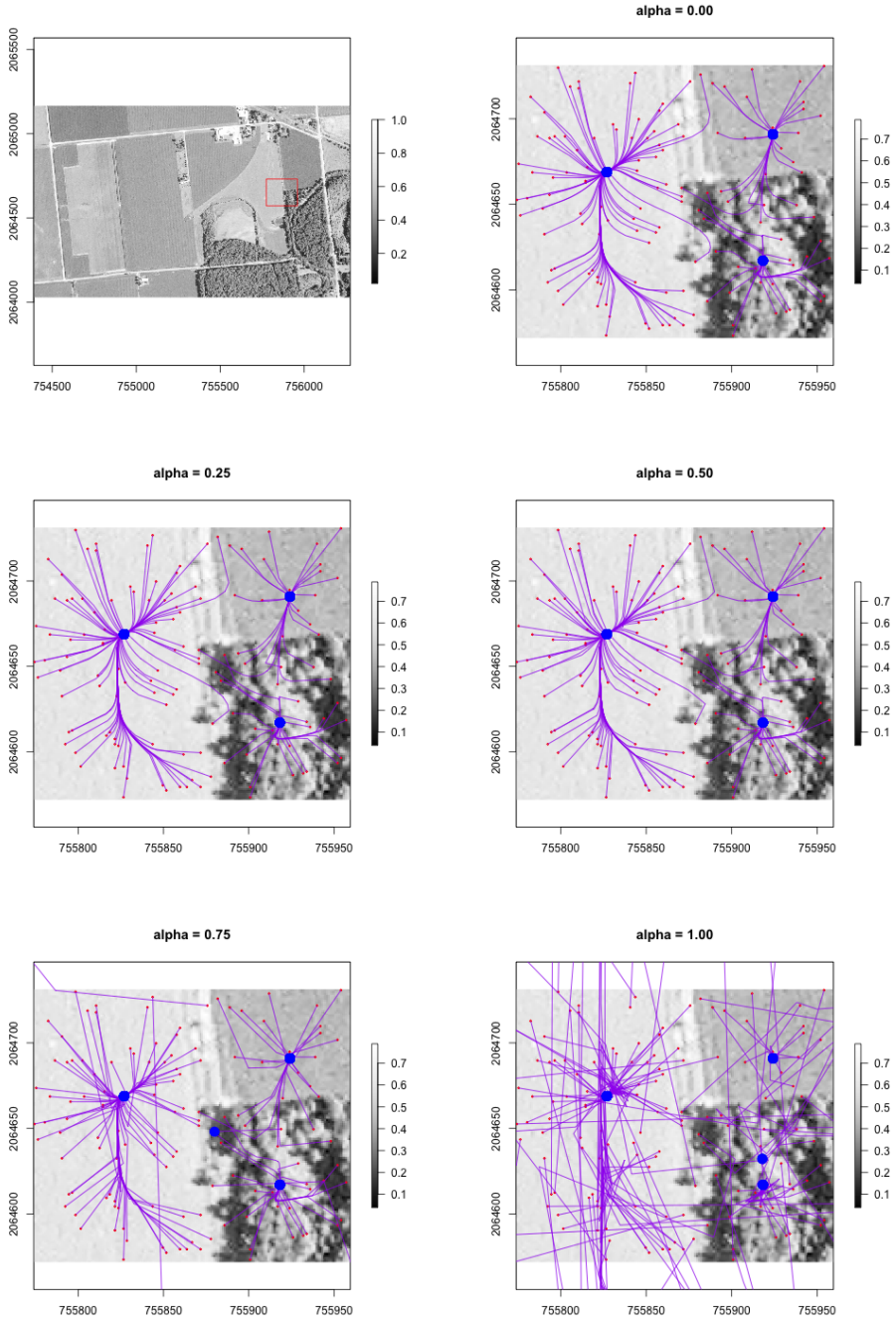

Figure 3.4: Paths of randomly selected pixels within the spatial support for various values of α under Normal Newton Shift (NNS) in a subset of Figure 3.11.

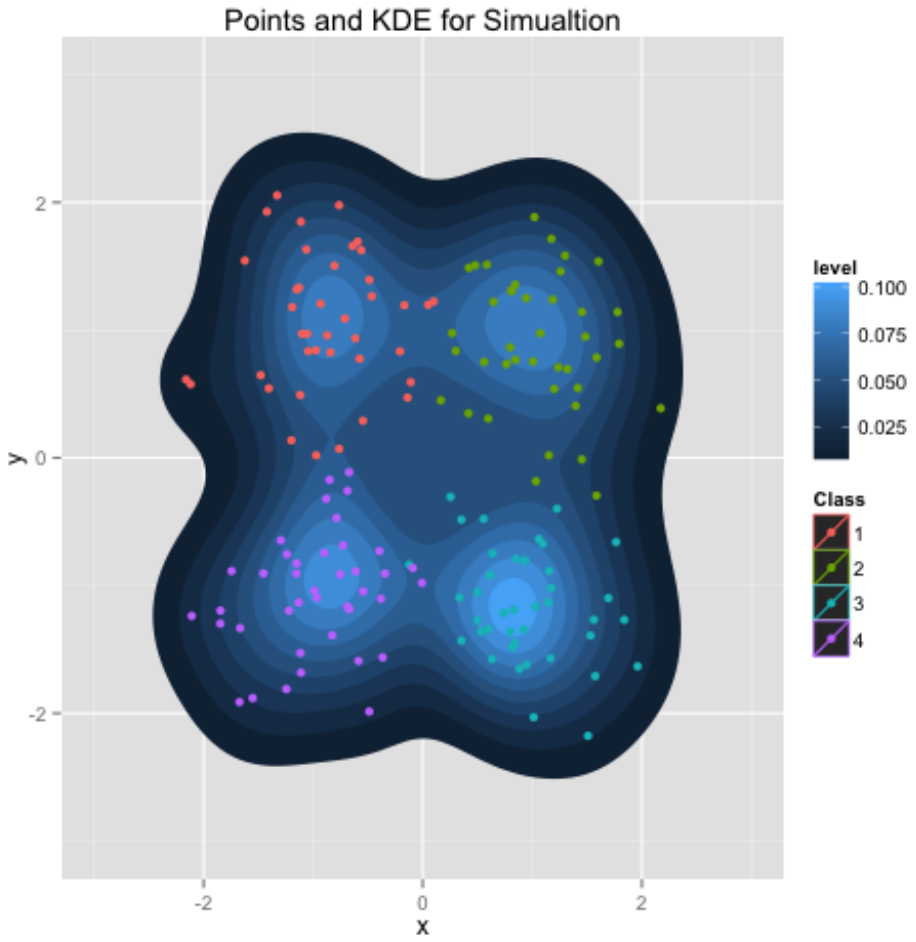

Figure 3.5: Observations from a Gaussian mixture generated from four bivariate normal distributions, and associated kernel density estimate.

3.5 Dual Tree Merge-Path Algorithm

In the dual tree mean shift implementation of Wang et al. (2007), based on the more general Gray and Moore (2000), a reference tree and a query tree are used to perform high-dimension binning for kernel density estimates. An alternative dual-tree approach, presented here, develops a mapping between $\mathbb{R}^d \to \mathbb{R}^d$ by exploiting the fact that mean shift mapping is independent of the sequence index, and the observation that on approach to the local maxima the mean shift sequences follow similar paths. The approach is rather

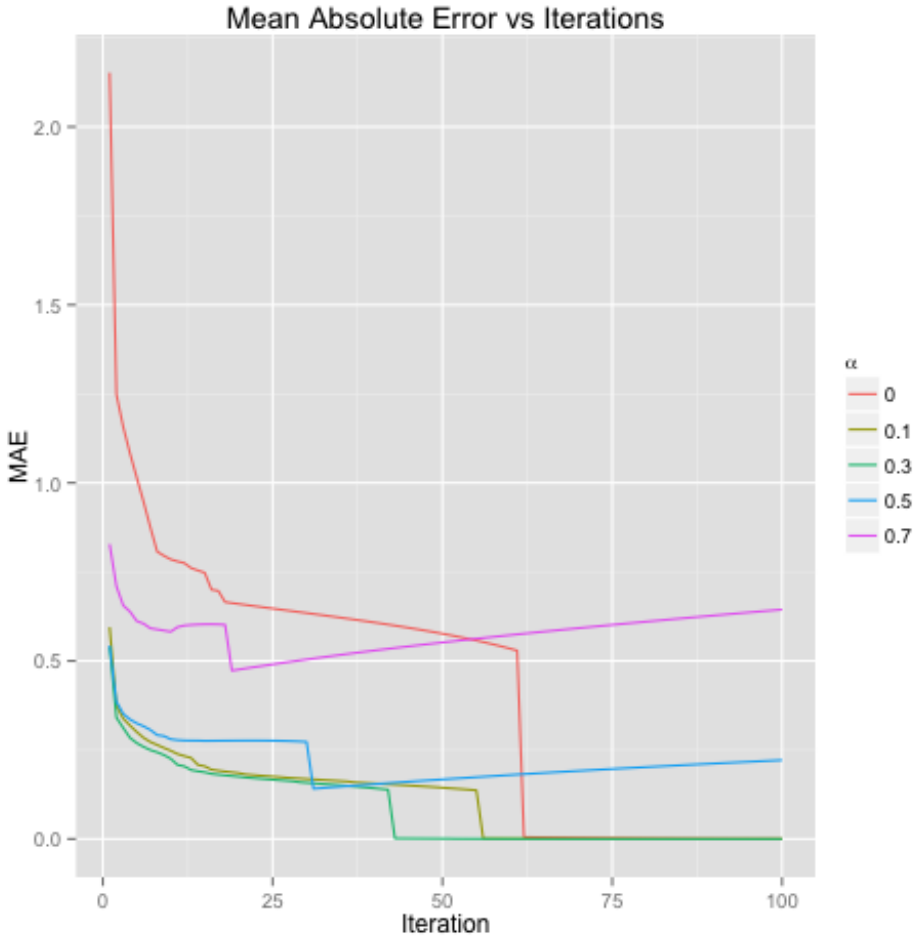

Figure 3.6: Mean Absolute Error as a function of Iterations.

simple, and the algorithm is outlined in Algorithm 2. This approach merges sequences that are sufficiently close to each other, where the distance between sequences at iteration m is defined by

$$d(x_m, y_m) = \min \{d(x_i, y_j) : i \in \{1, \cdots, m\}, j \in \{1, \cdots, m\}\} \tag{3.25}$$

for sequences $x_m = \{x_i\}_{i=1}^m$ and $y_m = \{y_i\}_{i=1}^m$.

In the algorithm below the query and reference points, Q and R are within an array. If sampling is used, the final T_Q tree may be used to classify the non-sampled units through nearest neighbor interpolation.

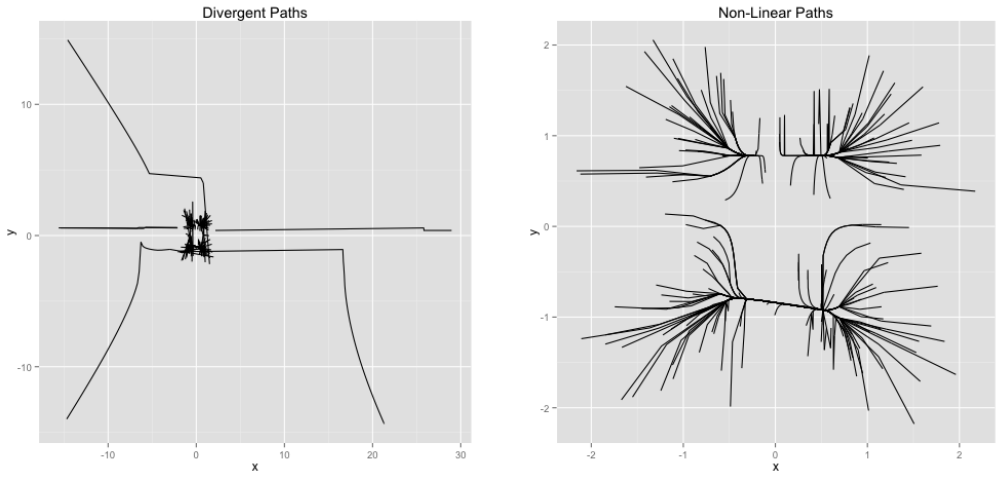

Figure 3.7: Divergent paths for $\alpha = 0.7$ (Left), and convergent nonlinear paths for $\alpha = 0$ (Right).

Algorithm 2. $T_Q \leftarrow \text{tree}(Q)$ {Build k-d tree from Q}

$T_R \leftarrow \text{tree}(R)$ {Build k-d tree from R}

$Q^{(0)} \leftarrow Q$

while $i \leq m$ **do**

$Q_{\text{KNN}} \leftarrow \text{query}(T_R, Q^{(i-1)}, k)$ {Get k-nearest neighbors.}

$Q^{(i)} \leftarrow \hat{f}(Q^{(i-1)}, Q_{\text{KNN}})$ {Perform NNS.}

$Q_{\text{1NN}} \leftarrow \text{query}(T_Q, Q^{(i)}, 1)$ {Get one nearest neighbor.}

while $j \leq \text{length}(Q_{\text{1NN}})$ **do**

$q^* \leftarrow Q_{\text{1NN}}[j]$

if $(||q - q^*|| < \epsilon)$ and $\left(\hat{f}(q) < \hat{f}(q^*)\right)$ **then**

$Q_{\text{MERGE}} \leftarrow Q_{\text{MERGE}} \cup \{q\}$

end if

$j = j + 1$

end while

$Q^{(i)} \leftarrow Q^{(i)}\ Q_{\text{MERGE}}$

$T_Q \leftarrow \text{tree}\left(\cup_{j=1}^{i} Q^{(i)}\right)$

$i = i + 1$

end while

The computational burden of the algorithm for building the k-d tree and searching is of the order

$$\mathcal{O}\left(\sum_{i=1}^{m} M_i log(M_i) + m_i log(M_i) + k m_i log(N_R)\right), \tag{3.26}$$

where $M_i = \sum_{j=0}^{i-1} m_j$, k is the number of neighbors to find, and $m_0 = N_Q$. While a typical single k-d tree would have computational cost of $\mathcal{O}\left(mkN_Q log(N_R)\right)$. However, the computational burden for computing the kernel density estimate in the merge tree algorithm of the order $\mathcal{O}\left(kM_m\right)$ while the single k-d tree has computational burden for kernel computation of order $\mathcal{O}\left(kmN_Q\right)$. Therefore, the advantage of this method over a single k-d tree implementation is a function of the cost of computing the kernel. In application, for k-d tree based implementations the cost of computing the kernel exceeds the search cost Wang et al. (2007). This gives an advantage for the dual-tree merge path algorithm over a single k-d tree.

A short performance test between a naïve implementation, a single k-d tree implementation and the merge path implementations was performed. Testing does not include the dual-tree Wang et al. (2007) due to difficulty in reproducing the method, and lack of availability of the original source code. Each point in the data set is an observation from $\mathcal{N}(0, I_6)$ where I_6 is a six-by-six identity matrix. In this performance test, the bandwidth, number of observations, and number of neighbors for the k-d and merge path algorithm were factors. Comparisons were all done using the same data set for 10 iterations.

The most obvious result in Table 3.4 is the non-ignorable overhead to construct and search k-d trees. This is most apparent when the number of neighbors equal the number of observations. The overhead disappears in the case of 5,000 observations when using 1,250

Table 3.4: Elapsed time in seconds of three implementations of the mean shift algorithm run for 10 iterations on an i7-4790K processor.

Implementation	Bandwidth		
	0.1	1.0	2.0
1,000 Observations, 1,000 neighbors			
Naïve	2.8	2.3	2.4
K-D	6.6	5.4	5.4
K-D Merge	3.7	4.5	1.6
1,000 Observations, 500 neighbors			
K-D	2.7	2.1	2.0
K-D Merge	1.3	2.3	1.7
5,000 Observations, 5,000 neighbors			
Naïve	62.2	49.6	51.6
K-D	184.4	150.7	151.7
K-D Merge	106.5	106.9	45.8
5,000 Observations, 2,500 neighbors			
K-D	73.5	58.3	56.7
K-D Merge	37.2	58.6	38.1
5,000 Observations, 1,250 neighbors			
K-D	33.9	26.9	26.6
K-D Merge	16.5	28.1	26.6

neighbors. In most situations, the merge path implementation out-performs the k-d tree, the exception occurs under cases of sufficiently small neighbors where the advantages of merging are diminished. When compared to the naïve implementation the advantage goes to the merge tree algorithm for a sufficiently small number of neighbors. The advantage also goes to the merge tree when a large number of iterations are required for convergence, Table 3.5, this is a common occurrence in this application to remote sensing.

Determining the appropriate number of neighbors is not obvious. Choosing insufficient neighbors may result in excessive number of local maxima, and too many will not provide any performance benefit. This problem is not directly addressed in this research, instead target run times are chosen for the mean shift operation. Target run times are specified in seconds per square mile, and the number of neighbors are chosen *a priori* to fit this target. The *a priori* target is determined by running the mean shift implementation over a set of test images. Literature on neighbor selection is limited, but relative error for mean shift

Table 3.5: Elapsed time in seconds of three implementations of the mean shift algorithm run for 20 iterations with bandwidth set at 3 for each dimension on two Xeon 5335 processors.

Implementation	Cores			
	1	2	4	8
2,000 Observations, 2,000 neighbors				
LPCM - ms	16.57			
meanShiftR K-D	27.73	13.83	6.93	3.69
meanShiftR K-D Merge	4.31	2.36	1.36	0.90
2,000 Observations, 1,000 neighbors				
meanShiftR K-D	10.02	5.08	2.56	1.34
meanShiftR K-D Merge	4.98	2.63	1.73	1.09
2,000 Observations, 500 neighbors				
meanShiftR K-D	4.66	2.44	1.27	0.68
meanShiftR K-D Merge	2.93	1.77	1.23	0.93

approximations for a number of images can be found in Wang et al. (2007).

3.6 Mean Shift R Package

An R package, *meanShiftR*, has been made available using the Algorithm 2, and a traditional k-d tree algorithm. This software is available through online sources, Lisic (2015c). The only other currently supported mean shift algorithm in R, per CRAN (Central R Archive Network), is the *ms* function in the *LPCM* (local principal curve methods) package (see University and Einbeck, 2011). The primary advantages of the meanShiftR package over the LPCM package are: the meanShiftR MS implementation is written entirely in C/C++ with Open MP support; the meanShiftR package supports a variety of approximate nearest neighbor implementations through the FLANN C++ library (see Muja and Lowe, 2009). A short performance comparison between the methods is provided below using the same simulation methods in Table 3.4.

There are some odd results in Table 3.5, where the merge tree run time with full neighbors is less than some subset of neighbors. It is unclear why this occurs, and requires further investigation into the FLANN library. Otherwise there is a fairly notable speed up relative

to the LPCM package's mean shift implementation.

It should be noted, at least one other R package does implement the mean shift algorithm, *r-opencv* (see Zhang, 2014). *r-opencv* is a wrapper for the OpenCV (Open Computer Vision) library, and through this library a binning implementation of the mean shift algorithm is provided. Unfortunately, binning is unsuitable for high dimensional data, making this package unsuitable for this application (see Bradski and Kaehler, 2008). *r-opencv* is also not currently supported, and not available on CRAN.

3.7 Image Segmentation of Remotely Sensed Images

The application of the mean shift segmentation algorithm follows a set of three steps, initial image processing, estimating filtered estimates, and mean shift classification. Initial image processing includes downloading the images, projecting multiple years of images to the same projection, and grayscale conversion. Estimating filtered images includes smoothing the image, calculating the local variance, and stratified sampling using the LVF. Mean shift classification includes applying the mean shift algorithm to the NAIP imagery, and post processing by merging segments below a threshold and applying a modal filtered estimate to smooth edges.

To make use of cheap general purpose computing, the mean shift program is run over regions of land geographically bounded by known roads. These geographically bounded regions are called *parts*, and are described on page 78. The parts are loaded onto a PostGIS database server. PostGIS is a GIS adapted version of the popular open source database server PostgreSQL. Computing nodes, other computers, each independently run an R program performing the steps in 3.8, with the exception of classification. Classification, is performed on a single computing node. Most image processing steps are cached to avoid re-computation. Final results are stored in the PostGIS database.

Total run time excluding prediction is approximately six hours using three compute nodes with Intel i7 quad core processors and 32GB of RAM each. The database and file server ran off a single computer using an Intel i5 quad core processor, also with 32GB of

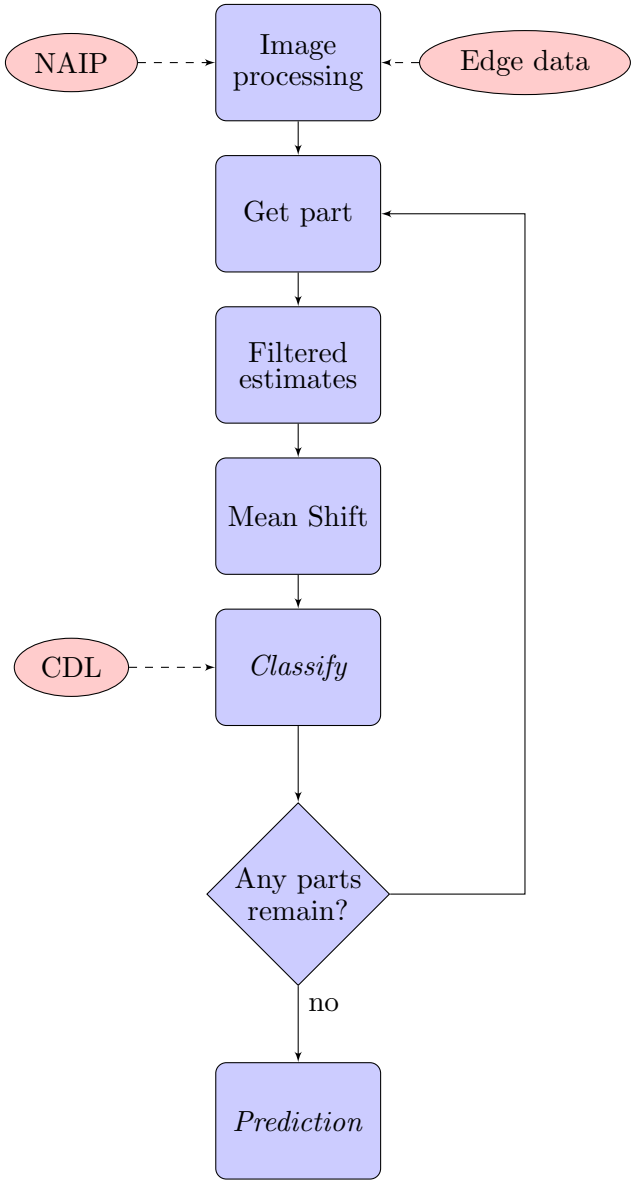

Figure 3.8: A flowchart of the parcel level agricultural prediction application.

RAM.

The accuracy of the initial mean segmentation results relative to a set of 200 hand-drawn LCUs had an adjusted Rand index (ARI) of 0.7. A high ARI value indicates a lack of under or over segmentation, but does not penalize jagged or thin protruding boundaries such as part of a road. The results of the initial LCUs are certainly not perfect. The edges, in

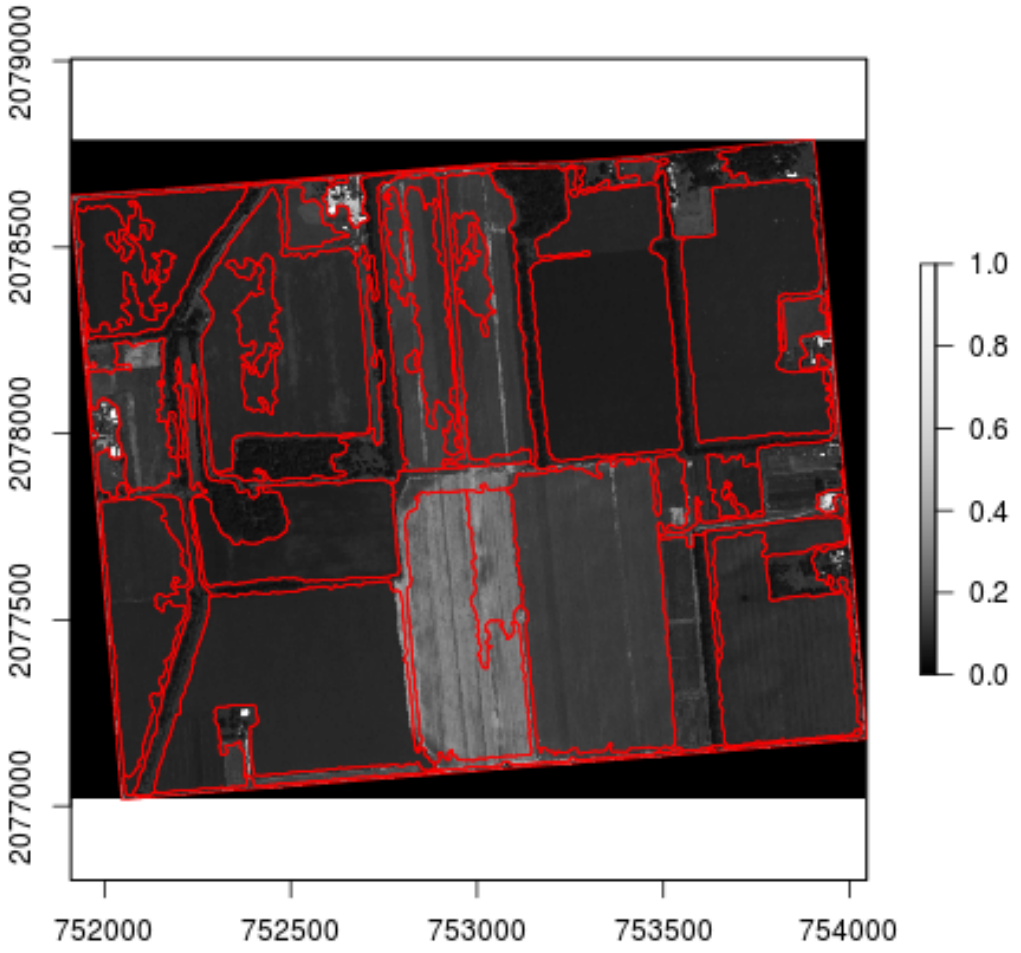

Figure 3.9: Mean shift segmented boundaries over NAIP imagery.

particular edges near trees or other tall structures can be erratic when shadows are present, Figure 3.9.

3.7.1 NAIP Imagery

NAIP pixels are observed at, and indexed by n discrete locations in the continental United States. The surface of the continental United States, formed by the union of NAIP pixels

will be identified simply as R. Each NAIP pixel is projected under an Alber's equal area (AEA) projection, with the same datum for the entire United States. The AEA projection is sufficient for local modeling performed here. Due to the high resolution of NAIP imagery relative to the size of and shape of the LCU used for agriculture, there will be no distinction between an LCU formed from NAIP pixels and the true area covered by LCU.

Each NAIP pixel can be identified through a Cartesian coordinate system starting at the North Western-most part of the United States, with the first coordinate indicating meters to the East from this point, and the second coordinate indicating the number of pixels South from this points. Each observed NAIP pixel is identified as $x(t, s)$, where $s = (s_1, s_2)$ is the vector of length two containing the first and second coordinates in the first and second element positions respectively, and t is a year in the temporal window identified by the set $\{1, \cdots, T\}$. Each location s is indexed in a set D through row major form.

Each NAIP pixel has an image intensity value (amount of light reflected) this value is positive and real. This intensity measure is obtained through a grayscale conversion, of the original RGB (Red, Green, and Blue) channels in a NAIP image. Grayscale conversion is performed through a linear luminance conversion found in Anderson et al. (1996). The grayscale conversion produced near identical segmentation results compared to both the utilization of all RGB channels, or by reprojecting the RGB space into the largest principal component. The NAIP pixels given LCU membership are assumed to be approximately normally distributed with mean $\mu(\xi)$ and variance $\Sigma(\xi)$ where ξ is the index of a LCU. To assist in clustering, each pixel is initially smoothed using a square uniform kernel with bandwidth (edge length) h=11.

Each NAIP pixel is a member of a unique LCU estimate. The mapping between classified NAIP pixels and their associated LCU estimates will be through a binary valued matrix G, where each row is a spatial index for the NAIP pixels, and each column is an LCU estimate index. Since only LCUs of sufficient size for commercial land use will be considered, LCU estimates only greater than eight 56 m^2 square meter CDL pixels(approximately 4 acres or 12,544 square meters) are considered. An overlay of NAIP imagery with CDL pixels can

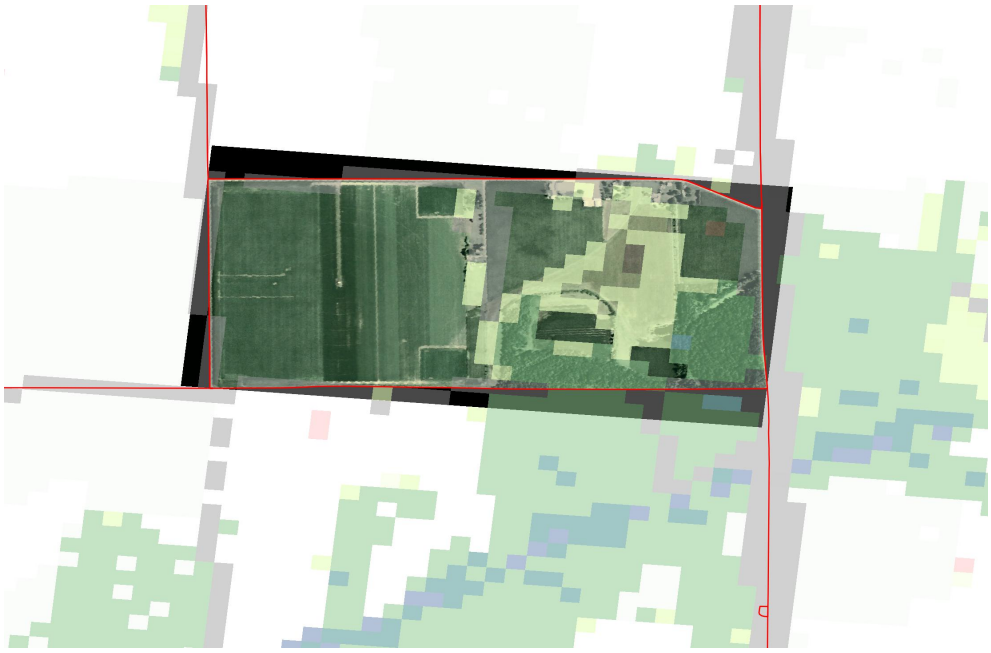

Figure 3.10: A Translucent CDL layer overlaid over a NAIP Imagery layer; The red edges indicate boundaries for classification formed via 2010 Tiger Edge Shape files from the U.S. Census Bureau.

be seen in Figure 3.10.

3.7.2 U.S. Census Bureau GIS Data and Problem Reduction

To reduce the memory requirements of segmenting large areas of land, permanent boundaries from the U.S. Census Bureau's edge data will be used to divide the problem into smaller areas of land. U.S. Census Bureau's Edge Data is a collection of polygon boundaries of geographic features (e.g. lakes), linear features such as roads and hydrography (e.g. rivers and streams) (freely available through the US Census FTP site). The edges in this data are used to partition the continental United States into areas of land similar in size to U.S. Census blocks, approximately one square mile for agriculturally dense areas. Any error within the U.S. Census Bureau's GIS data will be considered insignificant and ignored. These smaller areas will be called *parts*, and are simply connected areas of land that are bounded by roads or permanent hydrographic features (Figure 3.12). For the purposes of

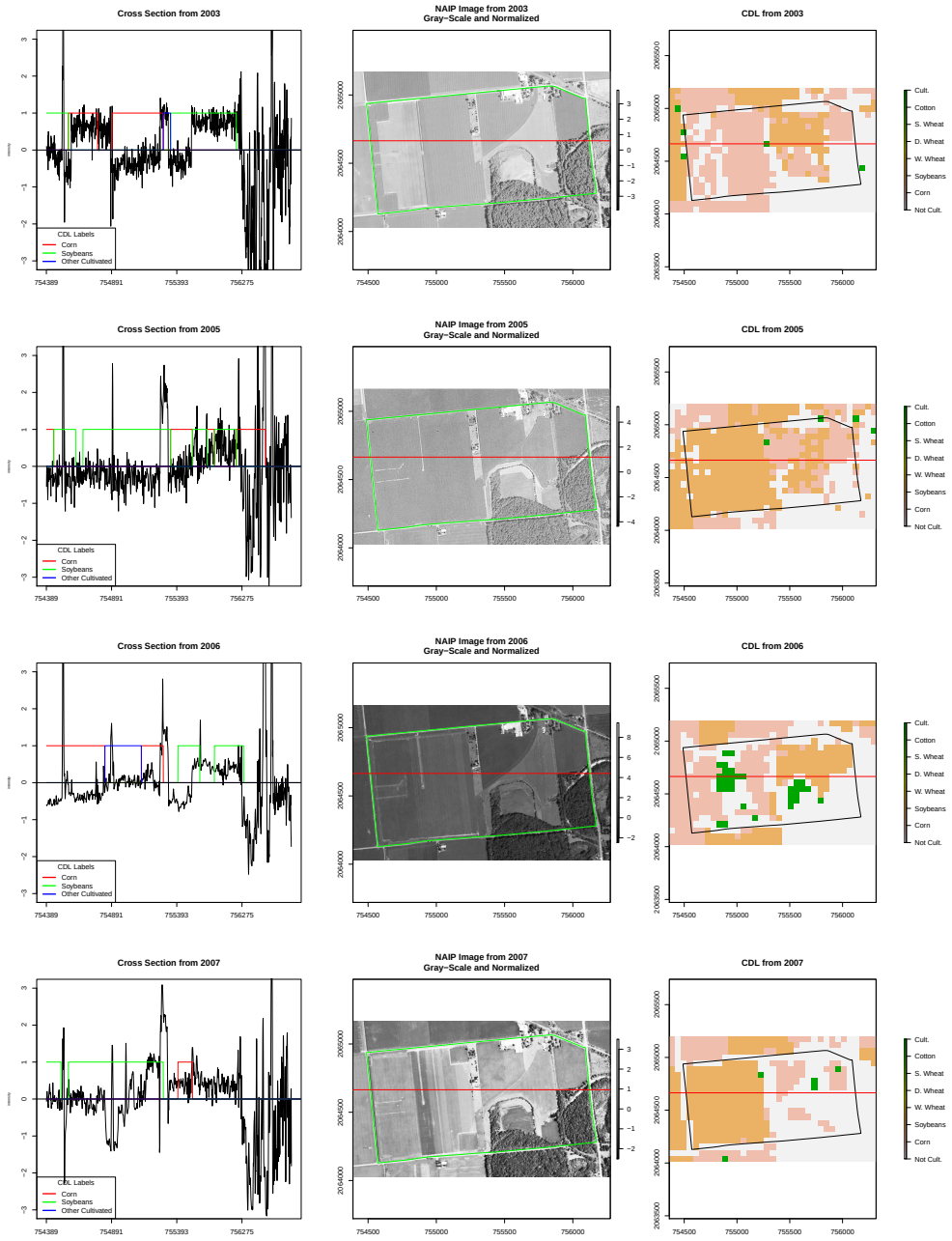

Figure 3.11: Cross section and digital image of a bounded region (Green for NAIP, Black for CDL) in Indiana, over several years. The red line indicates the location of the cross section.

segmentation, parts are identified as $R(\mathfrak{s})$ where $\mathfrak{s} \in \mathfrak{D}$ the index of all parts.

The partitioning of R is largely due to computational constraints. As an example,

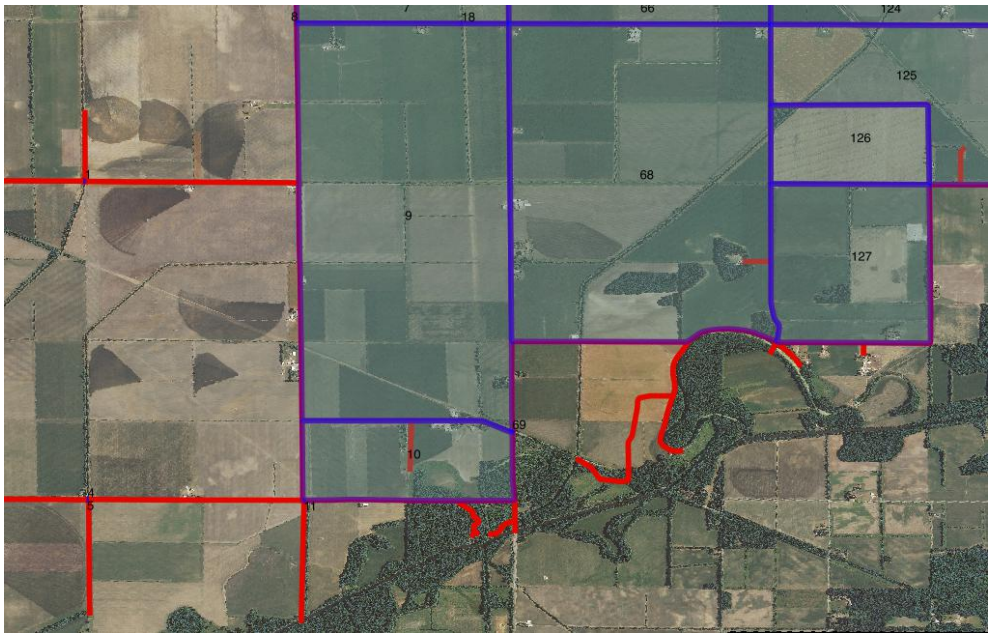

Figure 3.12: An example of U.S. Census Tiger GIS data used to estimate LCU boundaries plotted on top of NAIP imagery; LCU boundary estimates are in blue, unused edges are in red.

the Indiana county used to evaluate the approaches in this dissertation, La Porte County, Indiana, has 13 CDL images and six NAIP images. A CDL layer of $30m^2$ CDL raster image has a size of 1296×2051 pixels, as a compressed geotiff (Geographic Tagged Image Format) this is 2.7MB in size. In R due to integers being only 32-bit (regardless of the architecture), this raster image would use 10MB of memory. The equivalent NAIP image would take 900 times the amount of memory (9GB) regardless of the system being 32-bit or 64-bit since the NAIP images are stored in double precision format. To predict land cover in La Porte County, Indiana, for 2011, using all available prior NAIP imagery (6 years worth), would require 108GB of memory. The compressed NAIP geotiff image of the entire state of Indiana for a single year is approximately 500GB, making implementation difficult on inexpensive hardware.

3.7.3 Filtered Estimates

The filtered estimates include a local mean and variance calculation, and the LVF. The local mean and variance are calculated over square neighborhoods of 11 by 11 pixels. The square shape of the neighborhoods was due to computational simplicity, and the size was determined by visually assessing how smooth the image was, and the preservation of features such as boundaries.

Both the mean shift and LVF were written in C with an R interface, and make efficient use of multiple cores. Stratification was performed using LPM3 and the method described in Breidt (1995). The sampling rate was kept constant at 1/400. Both sampling methods were written in C with an R interface.

3.7.4 Mean Shift

One large advantage mean shift has over k-means, a popular classification method, is that it does not require the number of classes to be known *a priori*. A consequence of this is if the United States is broken into a set of parts of sufficient size there exists a bandwidth parameter of h that should work reasonably well for all parts. This can be seen by considering the rule-of-thumb bandwidth for the derivative of a kernel density estimate for six years of NAIP data, in this case is proportional to $n^{-1/18}$ for the Gaussian kernel. Therefore, large changes in n only slightly affect the optimal bandwidth.

In this implementation of the mean shift algorithm, the properties of the bandwidth along the boundary between LCUs is of the utmost importance. Since each year, is scaled the bandwidth selection problem is reduced to a global bandwidth selection problem where global implies over all NAIP images within a county. In this situation two bandwidth values will be selected, h_μ and h_σ, that refer to the NAIP intensity and NAIP log variance respectfully. The natural way to approach this problem is through cross validation (CV), this requires the identification from secondary sources or construction of LCUs and an objective function. To implement cross validation 200 hand-drawn LCUs were used, with the adjusted Rand index used as an objective function. The final bandwidth selected by

cross validation was h_μ and $h_\sigma = 1$.

The mean shift algorithm was run using the cross validated bandwidth on a set of sampled data. Each sampled point was an element of a 12 dimensional space including six years of smoothed NAIP data and local variances. Each pixel in the original image was provided a classification through nearest neighbor interpolation with the sampled response. The distances for the nearest neighbor classification were calculated using Euclidean distance in the 12 dimensional space used for shifting. The spatial support was not used for the mean shift or the interpolation due to the difficulty of finding a suitable bandwidth and the quality of the output.

A modal filtered estimate was applied to the final classification. The modal filtered estimate is used over an 11 by 11 square neighborhood, and replaces the center pixel with the most popular class in the 121 pixel neighborhood. The effect of this filtered estimate is the removal of single pixel edges, and smoothing of corners. The application of this filtered estimate substantially reduces issues that can occur when the pixels are turned into polygons. This filtered estimate provides an alternative to erosion and dilation functions found in computational morphology, and used in Yan and Roy (2014).

Chapter 4: Classification

Classification and post-segmentation adjustments are applied through three steps. These three steps are detection and re-segmentation of under segmented LCUs, identification of non-agricultural LCUs, and merging over-segmented LCUs. The application of all three classification steps has been evaluated using the Adjusted Rand Index (ARI) and a set of geospatial measures of error. USDA's Cropland Data Layer was used in the development of all three steps, but the methods employed are sufficiently general to work with other thematic crop maps.

4.1 CDL and LCU Terminology

The number of CDL pixels in R (3.7.1) is n^* and each Cartesian pair is identified with a vector s^*. Similar to NAIP pixels s^* is indexed by a set D^*. The CDL pixel for a given year t is identified by $x^*(t, s^*)$. CDL pixels are observed and categorical, represented by an integer in the set $\{1, \cdots, C\}$ of all possible CDL classifications (up to 256). For prediction, this set are reduced.

Depending on the modeling need, a CDL pixel may also be represented by a binary valued vector with length equal to the total number of CDL classes. To distinguish between the integer and binary vector values, the binary vector values will use the notation $\underline{x}^*(t, s^*)$. The binary vector value is zero for all values except the element equal to the integer value of the class.

CDL pixels are available at 30m^2 and $56m^2$, depending on year and state. For simplicity, all CDL pixels are resampled to 56m^2 this ensures that the index s^* is constant over all years studied. This resampling is done via nearest neighbor interpolation using Euclidean

distance applied to pixel centroids. Ties are handled by randomly selecting a tied nearest neighbor.

4.2 Segmentation of ALCU Estimates

Detection of under segmentation allows for the inclusion of missing boundaries by allowing a subset of the segments to be re-segmented. The advantage of re-segmentation is that the segmentation process is applied exclusively to a single segment, allowing for easier identification of local maxima relative to the initial segmentation. The disadvantage of re-segmentation is that re-segmentation may admit erroneous edges and has non-zero computational cost. The cost of re-segmentation is application specific, and the level of the tests should be set based on both desired quality of the output and the computational burden.

Spatial dispersion of error is an important characteristic of an LCU estimate, since multiple homogeneous clusters of pixels within an LCU may imply under segmentation while well distributed errors imply noisy data (Figure 4.1). In this section a test for under segmentation, or homogeneity is presented.

The Black-Black join count statistic provides a means to detect two types of dispersion, clustering and autocorrelation. The former type of dispersion, clustering, is of interest, while autocorrelation is not of interest. Kulldorff et al. (2003) calls this first type of dispersion first order clustering, and autocorrelation as second order clustering. In spatial data analysis the Black-Black join count test is traditionally used as a test of autocorrelation, where the assumption of second order stationarity is used to avoid the clustering alternative. Second order stationarity implies that over a spatial domain the mean and variance are constant, in this application the spatial domain is the interior of an estimated LCU. In the classification section of this dissertation only the first order clustering is of interest.

The null hypothesis for the Black-Black join count test for binary data is that all the pixels belong the same class, and discrepancies from this class should have iid error. In this problem the classification error can be from multiple classes. A conservative approach to

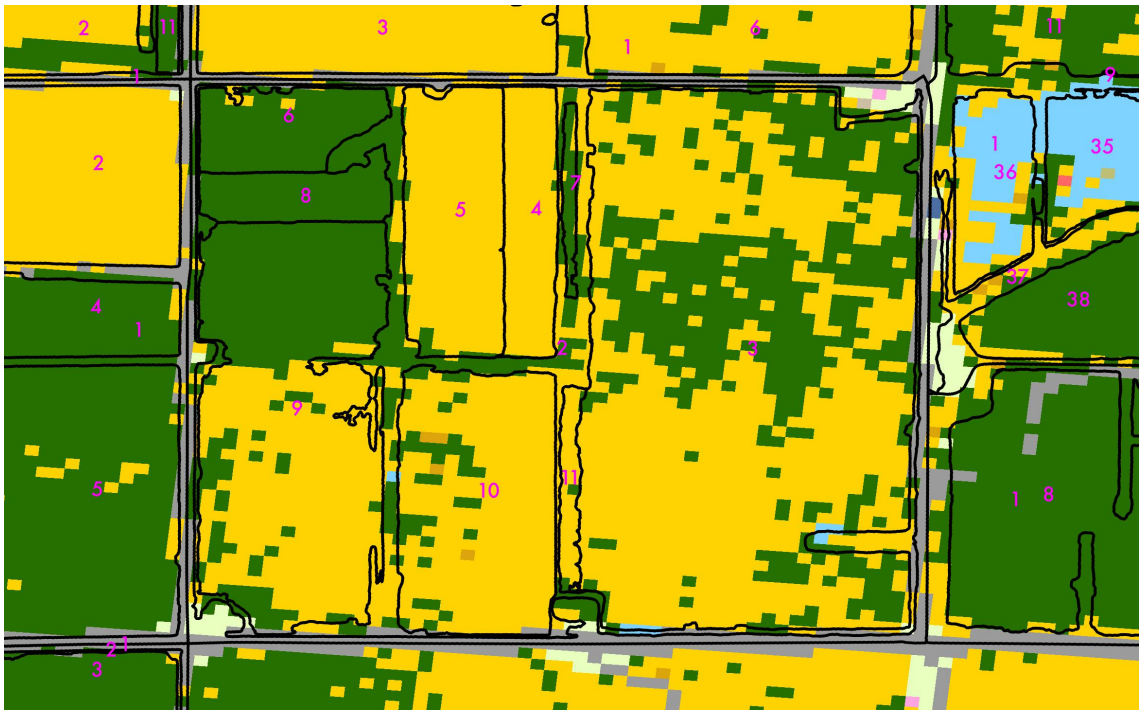

Figure 4.1: LCU estimate three and nine have roughly the same proportion of soybean classified pixels (green), however LCU estimate three's soybean pixels are likely classification errors due to the spread of the pixels.

this problem is to assume that the most popular class within a set of observations is the true class, and test the distributions of the discrepancies from this class. This approach is taken in this dissertation.

The tests for homogeneity considered are applied to single years independently, and the classes considered are crop types. The choice of majority class based on crop type is reasonable given both the low error rate for CDL classification for major crops, and the acreage of major crops relative to minor crops. The presence of under segmentation within an LCU for a single year implies that the LCU is under segmented.

Rejection of the null hypothesis for any year will imply the presence of under-segmentation of an LCU. Multiple comparison are based on the Šidák correction (see Šidák, 1967). This adjustment for the level of a test under multiple comparisons is $\alpha = 1 - (1 - \alpha_0)$, where α is the overall level of the test and α_0 is the level of each test. As an example, if the level of

the test is 0.05 for each year, then the level of the test for a four year sequence is close to 0.19, and for 10 years the level of the test is 0.40 (assuming independence of errors between years). Care should be taken in choosing the number of years to test, since the level of the test for all years is a function of the level of the test for each year. As the number of years increase the power of the test decreases. Therefore, the number of years used should be limited to the shortest sequence of years, such that there is a high probability of observing a change in sequence. This probability is application specific and crop specific, and is outside the scope of this research. For this particular application, only five years of CDL data are used, 2007-2011. Years before 2007 were excluded due to the due to the relatively poor quality of CDL classification before 2007.

Under conditions of high spatial correlation, an alternative to using the Black-Black join count test is to ignore the spatial distribution of the pixels and just compare the local error rate verse the global error rate for the majority crop. The rational behind this approach, is that if under segmentation occurs, then the error rate should be higher than the global error rate for the majority crop due to the presence of multiple homogeneous regions within an LCU. A global error rate is available for the CDL through an estimate of the producer's accuracy statistic, and similar quality measures should be available for any set of pre-classified pixels. *Producer's accuracy* is the probability that a classified pixel is in class c given the true class is c. In the CDL the producer's accuracy is estimated using the misclassification rate from the decision tree used to classify pixels using ground truth from FSA CLUs, see Boryan et al. (2011). Since this estimated producer's accuracy statistic is calculated over a large number of LCU's the precision is quite high relative to a local estimate of error rate from within an LCU. A binomial test with the null hypothesis that the local error rate is less than or equal to the global error rate for the majority crop can therefore be a useful test for the presence of clusters.

Both the Black-Black join count test and the binomial test will only be considered under a one-tail test. In the case of the Black-Black join count test this excludes alternatives with checkerboard like patterns in favor of clustering. The binomial test is uniformly most

powerful (UMP) under the one-tail test with binary distributed iid error assumptions. For major crops, where the global error rate is quite good, and the assumption that the error rate is constant over all LCUs, the binomial test is appropriate. For cases where the error rate is not reasonably constant, or the global error rate is poor, such as in minor crops, the Black-Black join count test statistic is reasonable. A useful test would be conditioned on the majority crop estimate, where a majority of pixels in a major crop would use the binomial test, otherwise the Black-Black join count test. The p-value for this test would be,

$$\text{p-value} = \begin{cases} 1 - F_0(\hat{Y}^*(t,\xi)) & c \in C_0 \\ 1 - F_1(J_{BB}(t,\xi)) & c \in C_1 \end{cases} \tag{4.1}$$

where

c	=	the majority crop for year t;
$x^*(t,\xi)$	=	the set of interior CDL pixels for LCU estimate indexed by ξ;
$\hat{Y}^*(t,\xi)$	=	the sum of majority crop pixels from LCU ξ;
$J_{BB}(t,\xi)$	=	Black-Black join count statistic (2.26) from LCU ξ;
F_0	=	the CDF for a binomial distribution with parameter $p_c(t)$ from the global error rate $1 - p_c(t)$;
F_1	=	the CDF of the BB join count test statistic, C_0 is the subset of C for major crops, and C_1 for minor crops.

A Monte Carlo test is used to compare the power of the binomial and Black-Black test statistic under the alternative hypothesis cases of the presence of interior LCUs and autocorrelation. This test is performed for a crop sequence of length one, and the level of the test is set at $\alpha = 0.05$. In the simulations with interior LCUs, square LCUs are generated at three levels of LCU sizes, three sizes of interior LCUs, and three levels of classification error, $q = 1 - p$. For the simulations under autocorrelation, the same levels of LCU sizes and classification errors are used, with the notable exception that three levels of autocorrelation are provided instead of sizes of interior LCUs. The autocorrelation is

generated under a SAR model with autocorrelation parameter equal to ρ, mean zero, and variance equal to one. Each deviate from the SAR model is broken into binary response, according to the normal quantile for p, values greater than this quantile are set to zero, and values less than or equal to this quantile are set to one. Each combination of levels was run 5,000 times for each experiment.

The results are straightforward, the binomial test is much more powerful at detecting the presence of an LCU except in the largest values of p, where both are equally powerful. The binomial test is also more robust against spatial autocorrelation, except in the cases of extreme autocorrelation where the autocorrelation is sufficiently high to affect the observed number of positive outcomes.

To test observed data, a set of LCUs were compared against Farm Service Agency (FSA) parcel like units known as common land units (CLUs). Although CLUs are not complete for a given region, they do provide a means of identifying under segmentation. To perform this test two groups are formed:

- The first group represented under segmented LCUs. This group was formed by identifying LCU estimates that overlap multiple FSA CLUs. Overlap is defined as sharing at least nine pixels with an LCU.

- The second group was formed by LCU estimates that had a high degree of mutual overlap with a single FSA CLUs. Where a high degree of mutual overlap is defined as greater than 70% and less than eight overlapping pixels from other CLUs.

Only CLUs from 2011 were available, and CDL data from 2007-2011 was used for test statistics.

The results of this analysis were interesting. In the density plot of the p-values from the Black-Black join count statistic, Figure 4.2, it is clear that there is little sensitivity to the choice of level of the test. This insensitivity is due to two reasons. The first reason is that there were a large number of LCUs that clearly included multiple crops, or had excessive misclassification. The second reason reason is due to the large number of LCUs, primarily

Table 4.1: Statistical power under the alternative of the presence of an LCU of size m with constant error rate of p for a square segment of size n.

p	n	m	J_{BB}	Binomial
0.60	25	4	0.07	0.47
0.70	25	4	0.15	0.97
0.80	25	4	0.37	1.00
0.90	25	4	0.77	1.00
0.60	49	4	0.07	0.83
0.70	49	4	0.14	1.00
0.80	49	4	0.35	1.00
0.90	49	4	0.79	1.00
0.60	81	4	0.06	0.95
0.70	81	4	0.11	1.00
0.80	81	4	0.28	1.00
0.90	81	4	0.77	1.00
0.60	25	9	0.09	0.29
0.70	25	9	0.26	0.85
0.80	25	9	0.57	1.00
0.90	25	9	0.91	1.00
0.60	49	9	0.10	0.73
0.70	49	9	0.28	1.00
0.80	49	9	0.67	1.00
0.90	49	9	0.97	1.00
0.60	81	9	0.09	0.93
0.70	81	9	0.25	1.00
0.80	81	9	0.64	1.00
0.90	81	9	0.97	1.00
0.60	25	16	0.11	0.13
0.70	25	16	0.24	0.43
0.80	25	16	0.50	0.82
0.90	25	16	0.84	1.00
0.60	49	16	0.12	0.57
0.70	49	16	0.38	0.99
0.80	49	16	0.80	1.00
0.90	49	16	0.99	1.00
0.60	81	16	0.12	0.86
0.70	81	16	0.39	1.00
0.80	81	16	0.86	1.00
0.90	81	16	0.99	1.00

small LCUs, with no error, causing zero valued p-values.

Table 4.2: Statistical power under the alternative of the presence of three levels of autocorrelation, for a square segment of size n.

p	n	ρ	J_{BB}	Binomial
0.60	25	0.00	0.07	0.03
0.70	25	0.00	0.06	0.04
0.80	25	0.00	0.06	0.05
0.90	25	0.00	0.07	0.03
0.60	49	0.00	0.06	0.04
0.70	49	0.00	0.06	0.04
0.80	49	0.00	0.06	0.02
0.90	49	0.00	0.07	0.02
0.60	81	0.00	0.05	0.03
0.70	81	0.00	0.05	0.05
0.80	81	0.00	0.06	0.04
0.90	81	0.00	0.07	0.03
0.60	25	0.50	0.30	0.14
0.70	25	0.50	0.29	0.17
0.80	25	0.50	0.27	0.18
0.90	25	0.50	0.25	0.19
0.60	49	0.50	0.46	0.16
0.70	49	0.50	0.44	0.17
0.80	49	0.50	0.39	0.15
0.90	49	0.50	0.33	0.15
0.60	81	0.50	0.61	0.15
0.70	81	0.50	0.57	0.19
0.80	81	0.50	0.52	0.21
0.90	81	0.50	0.41	0.18
0.60	25	0.90	0.62	0.48
0.70	25	0.90	0.63	0.57
0.80	25	0.90	0.65	0.63
0.90	25	0.90	0.68	0.75
0.60	49	0.90	0.86	0.46
0.70	49	0.90	0.88	0.52
0.80	49	0.90	0.89	0.58
0.90	49	0.90	0.91	0.65
0.60	81	0.90	0.95	0.46
0.70	81	0.90	0.96	0.52
0.80	81	0.90	0.95	0.57
0.90	81	0.90	0.95	0.66

It should also be noted that there is little difference between the binomial and Black-Black join count tests in this application. According to Figure 4.2, a test with level, α,

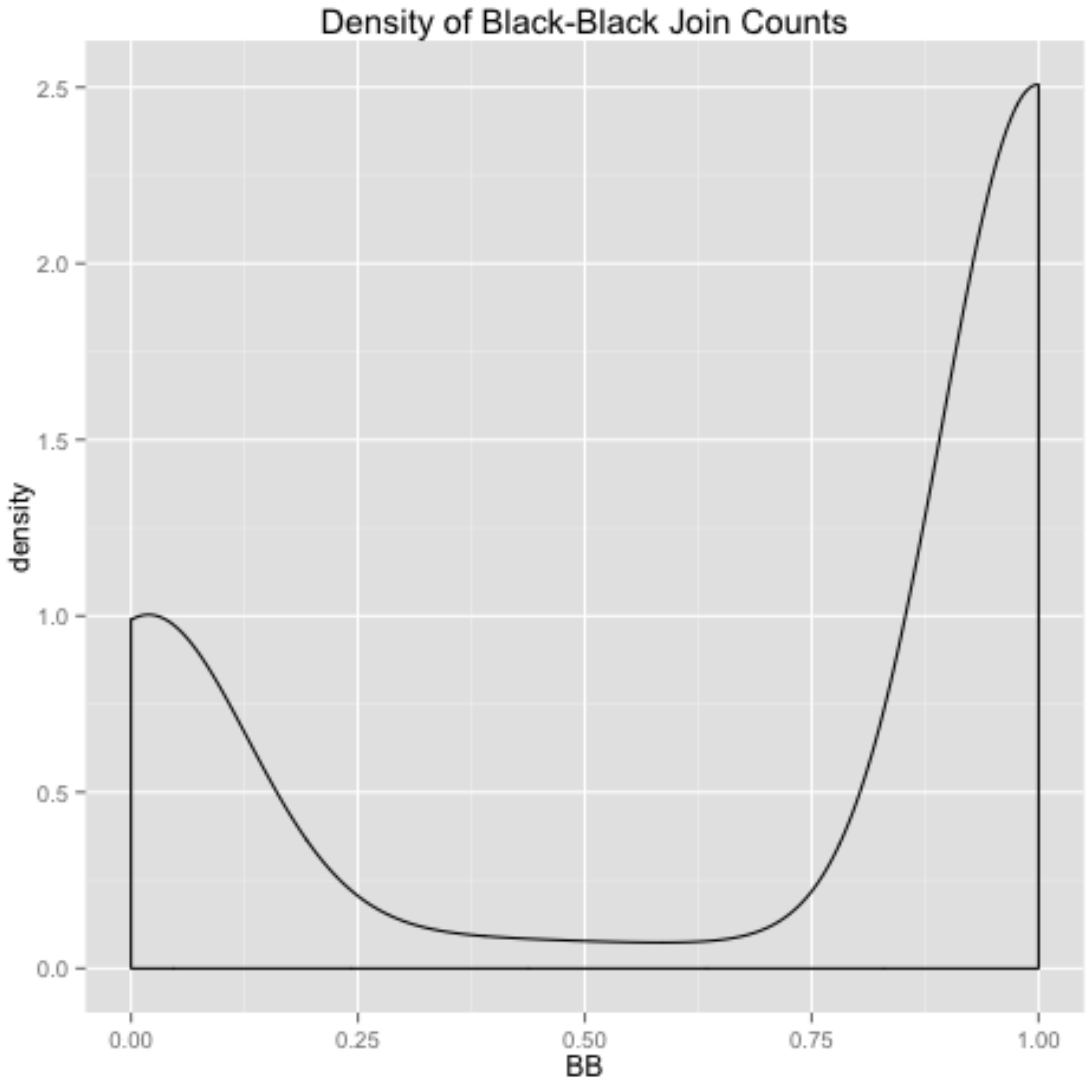

Figure 4.2: Density plot of the Black-Black statistic over a set of segmented LCUs in La Porte County, Indiana

slightly less than one would yield good results. Figure 4.3 Shows the result of splitting applied to LCU estimates for $\alpha = 0.40$ using CDL data from 2007-2011.

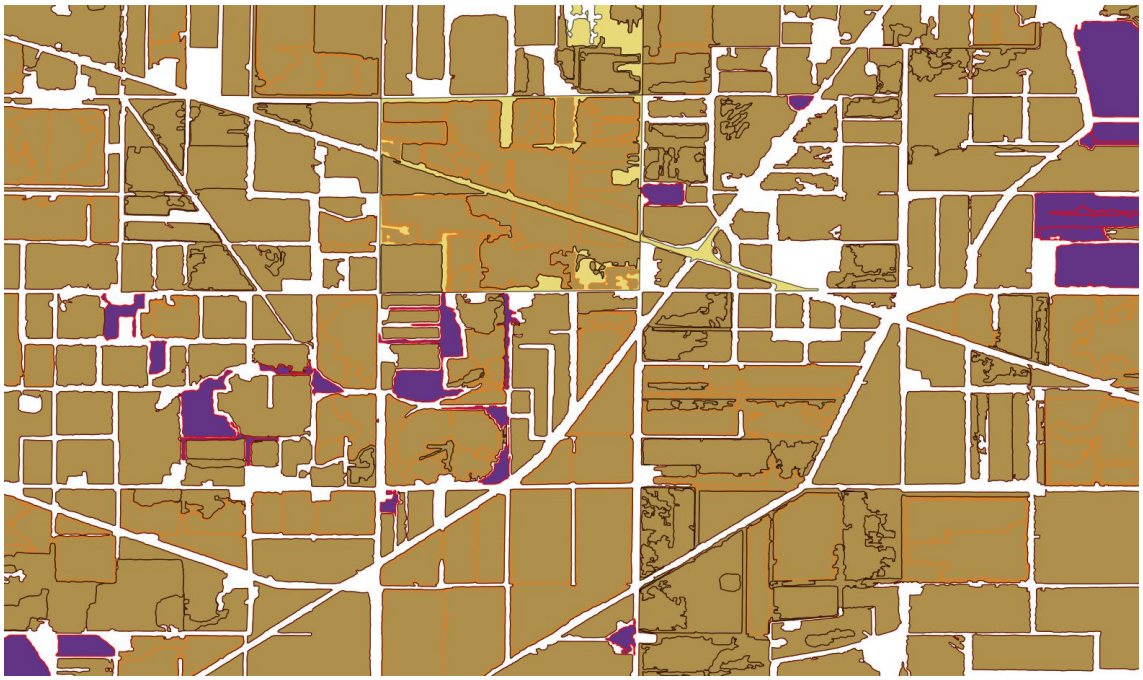

Figure 4.3: Two sets of LCU estimates are included in this image, the first set (yellow) simply had classified non-agricultural LCU estimates removed, the second set (purple) was re-partitioned based on the Black-Black join count statistic at level 0.40, and had non-agricultural LCU estimates removed. Brown pixels indicate overlap between the LCU estimates, and orange lines indicate new LCU estimates formed by re-partitioning.

4.3 Identification of Non-Agricultural LCU Estimates

Non-agricultural LCUs tend to follow roads and boundaries between fields. These LCUs are quite narrow, and often contain no interior CDL pixels or only CDL pixels associated with neighboring LCUs. Under these circumstances classification via CDL pixels is problematic at best, since few of these LCUs have interior pixels.

Instead, LCU boundary properties are useful in identifying if an LCU is likely to be used for agriculture. This can be seen through a logistic regression with L_1 regularization (see Friedman et al., 2001) using 200 hand-drawn LCUs supplemented by 50 LCUs manually identified as roads, houses, forests, and other non-agricultural land covers. In this model selection procedure only the intercept and coefficients for the ratio of LCU boundary estimate length to LCU area and the presence of an intersection with permanent GIS features,

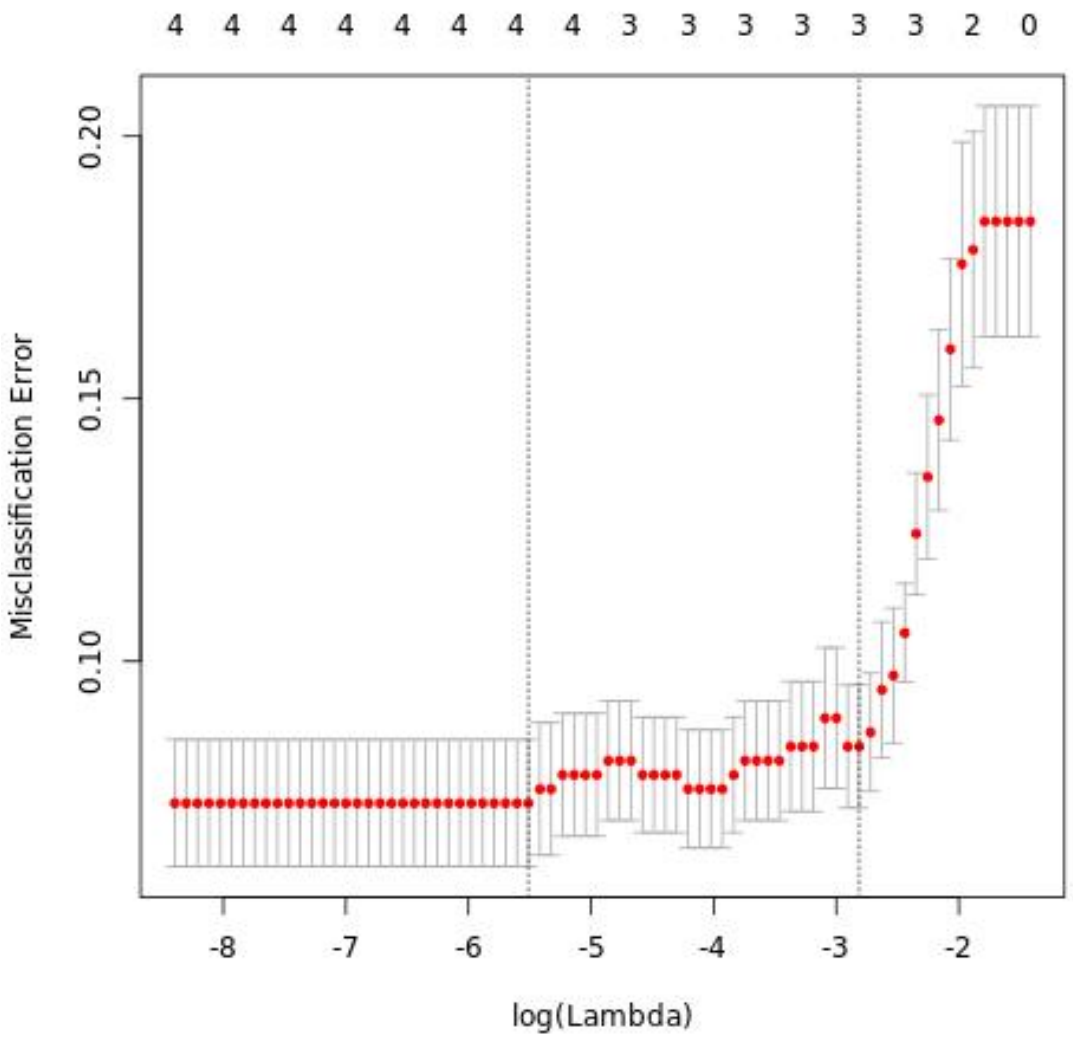

Figure 4.4: Misclassification error for various values of the regularization parameter λ.

were not shrunk to zero (Figure 4.4). The covariates that had their coefficients shrunk to zero included CDL pixel counts by land cover class, LCU mean intensity, LCU intensity variance, LCU area, and LCU edge length. The misclassification rate for this procedure was performed using cross validation in the R package *glmnet* (see Friedman et al., 2010), with a value of 5.0%.

4.4 Merging ALCUs Estimates

To prevent over-segmentation, adjacent ALCU (agricultural LCU) estimates not transected by a known geographic feature with identical land cover sequences and intensity characteristics (mean and variance) should be merged. This operation is problematic since these parameters of the ALCU estimates are not known, and both the CDL and the NAIP data contain measurement error. Due to this uncertainty, merges should be based on the likelihood of adjacent ALCUs sharing the same parameters.

Under a likelihood framework, the problem reduces to identifying neighboring ALCUs, and determining whether merging them together increases the likelihood relative to an alternative set of merges. Neighbors in this section follow the "rook" based definition of neighbors, where neighbors share a non-degenerate edge (length greater than zero). Since this problem is a combinatorial optimization problem with respect to the likelihood a, simulated annealing based approach is used.

To perform this simulated annealing approach a likelihood needs to be specified. The likelihood that an ALCU estimate's true land cover sequence, intensity, and mean; under the assumption of known CDL miss-classification error is provided in (4.2). This likelihood was formed based on the assumption that conditioned on the ALCU boundaries the classification and NAIP pixel intensity are independent. This is a reasonable assumption, since the NAIP intensity is independent of crop type, see Figure 4.5. The lack of relationship between the NAIP intensity and crop type is due to different planting times of individual crops within LCUs, and is further compounded by varying image acquisition periods of NAIP imagery.

$$\begin{aligned} \ell\left(c, \mu(\xi), \Sigma(\xi) | x^*, x, \hat{G}, p_c\right) &\propto \left(\frac{n_\xi^*!}{\prod_{c \in \{1,\cdots,C\}^T} x_c^*(\xi)}\right) \\ &\cdot \prod_{c' \in \{1,\cdots,C\}^T} p_{c|c'}^{x_{c'}^*(\xi)} \prod_{s \in D_\xi} \phi\left(x(s), \mu(\xi), \Sigma(\xi)\right) \end{aligned} \tag{4.2}$$

where

c = a sequence of crops in $\{1, \cdots, C\}^T$ the product space of C crop classes over T years;

D_ξ = the set of all interior NAIP pixels within ALCU $y(\xi)$;

$x_c^*(\xi)_c$ = the count of all interior CDL pixels of ALCU ξ that have land cover sequence c;

n_ξ^* = the number of all interior CDL pixels for ALCU ξ;

ϕ = the normal density;

$\mu(\xi)$ = the mean of the ALCU with index ξ;

$\Sigma(\xi)$ = the variance of the ALCU with index ξ;

G = is the NAIP pixel LCU assignment, estimated by $\hat{G}$, page 76;

$p_{c|c'}$ = is the probability that a CDL pixel belonging to class c given that its observed value was c'.

In this research it is assumed that there exists a true, fixed land cover sequence for each each ALCU, $y(\xi)$. Since the $y(\xi)$ are not observed, interior CDL pixels are treated as independent measurements of the land cover sequence of the estimated ALCU. Therefore, it is possible to classify an estimated ALCU by selecting the most likely crop sequence given the observed CDL pixel sequences. One problem with this approach is that conditioned on G, the probability that a CDL pixel has land cover sequence c given that we have observed sequence c', $p_{c|c'}$, is unknown.

For major crops we do have a good estimate for $p_{c|c}$ $t \in \{1, \cdots, t\}$, the yearly probability that a classified CLU pixel belongs to the correct class, $\Pr(\mathbf{x}^*(t, s^*) = c|\hat{\mathbf{x}}^*(t, s^*) = c)$. In GIS parlance this probability is called *user's accuracy*. What we do not have a good estimator for is $p_c c|c'$ where $c \neq c'$. Under the assumption that the misclassification rate is invariant to the class being misclassified, then the values of c' not equal to c can be considered a single class. Using the producer's accuracy, $\Pr(\hat{\mathbf{x}}^*(t, s^*) = c|\mathbf{x}^*(t, s^*) = c)$, the probability that a CDL pixel is classified into class c, and the probability that a pixel belongs

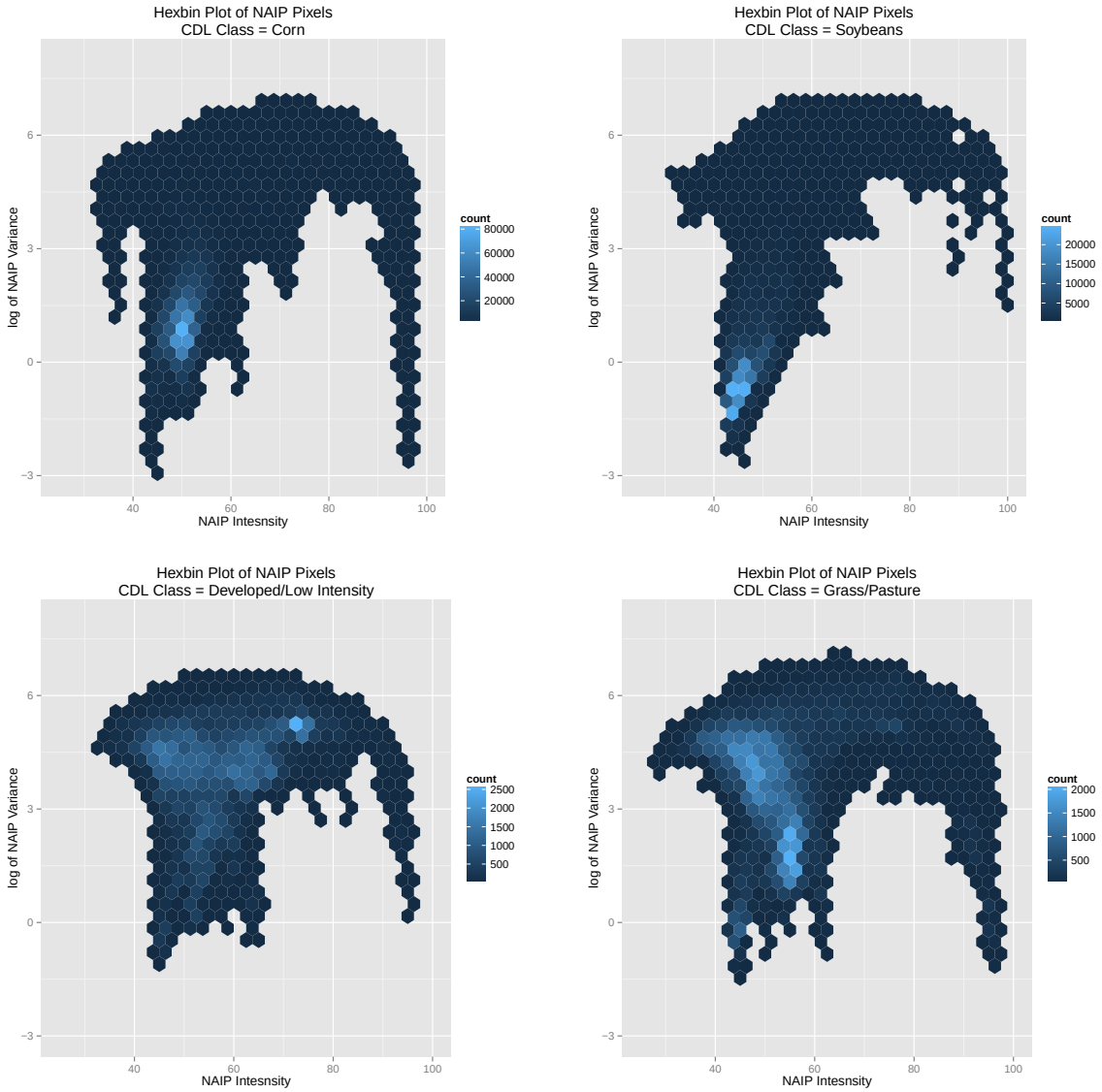

Figure 4.5: Hexbin plots of the NAIP pixels intensity for the bounded region of Indiana in Figure 3.11, conditioned on the CDL classification of the pixels; the log of the variance is calculated over a 15-by-15 pixel neighborhood centered around the target pixel.

to class c, it is possible to calculate

$$p_{c|c'} = \frac{\Pr\left(\hat{\mathbf{x}}^*(t,s^*) = c | \mathbf{x}^*(t,s^*) = c\right) \Pr\left(\hat{\mathbf{x}}^*(t,s^*) = c\right)}{\Pr\left(\mathbf{x}^*(t,s^*) = c'\right)} \tag{4.3}$$

where $c \neq c'$. $\Pr\left(\hat{\mathbf{x}}^*(t, s^*) = c\right)$ is calculated from the CDL response, and $\Pr\left(\mathbf{x}^*(t, s^*) = c\right)$ is calculated using FSA CLU data, "Ground Truth."

Assuming that the misclassification rate, one minus user's accuracy is temporarily independent, then it is possible to determine merges based on the likelihood in (4.2),

$$\begin{aligned} \ell\left(c, \mu(\xi), \Sigma(\xi) | x^*, x, \hat{G}, p_c\right) \quad &\propto \quad \left(\frac{n_\xi^*!}{(n_\xi^* - x^*(\xi)_c) x^*(\xi)_c}\right) p_{c|c}^{x_c^*(\xi)} p_{c|c'}^{n_\xi^* - x_c^*(\xi)} \\ &\cdot \quad \prod_{s \in D_\xi} \phi\left(x(s), \mu(\xi), \Sigma(\xi)\right) \end{aligned} \tag{4.4}$$

where $c \neq c'$.

Unlike CDL pixels, NAIP pixels are observed through remote sensing, not an estimator of land cover. A NAIP pixel's observed intensity is also likely to change over the growing season, even within a day due to weather and production practices. This change in intensity is captured by the intensity mean and variance associated with each ALCU. Because measurement error from the NAIP pixels is low relative to the variation of intensity within an ALCU, the measurement error is ignored and each NAIP pixel within an estimated ALCU are treated as an independent realization of the ALCU estimate's intensity. Since NAIP pixels aren't necessarily collected every year, and only a small number of yearly observations are available for each NAIP pixel, the NAIP pixels are assumed to be temporally independent. This assumption greatly simplifies the covariance structure of the NAIP pixels (4.2) and (4.4),

$$\begin{aligned} \ell\left(c, \mu(\xi), \Sigma(\xi) | x^*, x, \hat{G}, p_c\right) \quad &\propto \quad \left(\frac{n_\xi^*!}{(n_\xi^* - x^*(\xi)_c) x^*(\xi)_c}\right) p_{c|c}^{x_c^*(\xi)} p_{c|c'}^{n_\xi^* - x_c^*(\xi)} \\ &\cdot \quad \prod_{s \in D_\xi} \prod_{t \in \{1, \cdots, T\}} \phi\left(x(s), \mu(t, \xi), \Sigma(t, \xi)\right). \end{aligned} \tag{4.5}$$

where $c \neq c'$.

The final class assigned to each ALCU is the maximum likelihood estimator under (4.4), and evaluated for each year. In practice, there is little difference between applying

Table 4.3: USDA, National Agricultural Statistics Service, 2007 Indiana cropland data layer for major crops.

Cover Type	Correct Pixels	Producer's Accuracy	Omission Error	User's Kappa	User's Accuracy	Kappa
Corn	1333209	96.17%	3.83%	0.9248	95.63%	0.9147
Sorghum	502	81.63%	18.37%	0.8161	23.44%	0.2342
Soybeans	892534	94.90%	5.10%	0.9235	94.28%	0.9144
Tobacco	0	n/a	n/a	n/a	0.00%	0
Barley	0	n/a	n/a	n/a	0.00%	0
Winter Wheat	28044	86.63%	13.37%	0.8647	82.73%	0.8253

Table 4.4: Contingency table of correct and incorrect merges of the merging algorithm

	Needed Merging	Did Not Need Merging	Total
Merged	26	14	40
Not Merged	6	156	160
Total	32	168	200

the likelihood estimator under (4.4) and letting the ALCU assignment be equal to the most popular class. In application to La Porte County, Indiana, there was no difference in assignment between using (4.4) and assigning the most popular class from the interior CDL pixels.

To test the merging method against observed data, the set of hand-drawn ALCUs from the segmentation portion of this dissertation were used. The FSA CLUs were not used due to difficulty in identifying FSA CLUs that meet the ALCU definition. Of the 200 hand-drawn ALCUs, 32 had intersecting estimated ALCUs from the mean shift segmentation that required merging. After performing the likelihood based method 40 of the hand-drawn ALCUs had intersecting estimated ALCUs that were merged, including 14 that should not have been merged.

Many of the incorrect mergers occurred between ALCUs with identical land use sequences without prominent boundaries between them. An example of this can be seen in Figure 4.6 where an LCU was correctly split (red), and then incorrectly merged back together (blue). Almost all of the estimated ALCUs needing merges, that were not merged,

Figure 4.6: Incorrectly merged boundaries (blue), from split boundaries (red).

were due to poor boundaries. The ALCUs with poor boundaries tended to overlap multiple fields, but were not split apart in the re-segmentation step.

Chapter 5: Prediction

In this chapter, a spatial-temporal autoregressive probit model is introduced for modeling crop rotations. This model differs from prior spatial-temporal autoregressive probits in the literature both on application and model specification. This model specification follows a traditional Bayesian hierarchical approach with a SAR prior placed on the coefficient parameter in the latent model. The spatially autoregressive prior allows for crop rotations to be spatially correlated. An application of this method to a subset of La Porte County, Indiana, is presented. In this application, predictions are made for 2010 to 2013, using data from 2001 to one year before the predicted year. Computational and theoretical issues are discussed.

5.1 Model Specification

The MNP and SAR MNP models described in Section 2.3 provide a way to link categorical response to a linear model with multivariate normal error structure. These models still need to be linked with the crop rotation phenomena used for prediction. To do this, it is assumed that the categorical response is temporally stationary and conditionally independent given a prior state, sequence of prior crops, on the same ALCU. Therefore, by specifying a set of prior states (rotations) $\{1, \cdots, P\}$, it is then possible to create a design matrix to include the prior state information for each ALCU.

In the proposed crop rotation model, it is also assumed that the crop rotations are spatially correlated, e.g. corn-to-soybean rotations are close to corn-to-soybean rotations. This model differs from the SAR MNP models of LeSage and Pace (2009) and Wang et al. (2012), where the response is considered auto-correlated, e.g. corn grows next to corn. The autocorrelated crop rotation model approach has the form

$$\mathbf{Z}^{**} = U^{**}\boldsymbol{\beta}^{**} + X\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{5.1}$$

$$\boldsymbol{\beta}^{**} = \mathbf{B}\boldsymbol{\beta}^{**} + (I - \mathbf{B})\beta_0 + \boldsymbol{\upsilon}, \tag{5.2}$$

where

$\mathbf{Z}^{**}$	=	$(\mathbf{z}_{1,1,1}, \cdots, \mathbf{z}_{J,1,1}, \cdots, \mathbf{z}_{J,n,1}, \cdots, \mathbf{z}_{J,n,T-l})$, a latent response vector of length $nJ(T-l)$,
U^{**}	=	$nJ(T-l) \times nPJ$ design matrix,
T	=	the number of observed years,
l	=	the lag, number of prior years,
$\boldsymbol{\beta}^{**}$	=	nPJ vector of covariates,
X	=	other covariates,
$\mathbf{B}$	=	spatial correlation matrix $I - \boldsymbol{\rho}W$,
$\boldsymbol{\epsilon}$	=	$\boldsymbol{\epsilon}\|\Omega \sim \mathcal{N}(0, \Omega)$,
$\boldsymbol{\Omega}$	=	$1_{T-l} \otimes I_n \otimes \boldsymbol{\Sigma}$,
$\boldsymbol{\Sigma}$	=	the covariance of $\epsilon_{\cdot,i,t}$,
$\boldsymbol{\gamma}$	=	other covariates coefficients,
β_0	=	hyper parameter for β_0, and
$\boldsymbol{\upsilon}$	=	random vector of length nJ with distribution $\mathcal{N}(0, \Sigma_0)$.

A problem with this model is the large number of parameters introduced; however, for larger values of ρ, the number of effective parameters is actually much lower than npJ, avoiding excessive over-fitting.

Similar to Wang et al. (2012) prior state information is provided through a covariate matrix U^{**}, unlike Wang et al. (2012) conditional temporal independence is assumed for each observation given a set of prior states. Prior states are a particular crop rotation pattern, e.g. an observed corn-to-soybeans rotation. To avoid the need for generalized inverses to calculate the inverse of $(U^{**})^T(U^{**})$, it is assumed that all prior states have been observed at least once. U^{**} is an $nJ(T-l) \times nPJ$ design matrix. The design matrix

U^{**} has the form

$$U^{**}\begin{bmatrix} P_{1,1} & 0 & \cdots & 0 \\ 0 & P_{2,1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_{n,1} \\ P_{1,2} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_{n,T-l-1} \\ P_{1,T-l} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_{n,T-l} \end{bmatrix} \tag{5.3}$$

where $P_{i,t}$ is a $J \times PJ$ matrix equal to $\underline{P}_{\xi,t} \otimes I_J$. $\underline{P}_{\xi,t}$ is a row vector of length P with zeroes everywhere except, at the index of the observed prior state for observation ξ for year t. For the purposes of this dissertation, no additional covariate information beyond U^{**} is used, therefore, the properties of $\boldsymbol{\gamma}$ will not be explored.

Parameter estimation as in other SAR MNP approaches is through gibbs sampling, however unlike LeSage and Pace (2009) and Wang et al. (2012) that use conditional data augmentation, the more computationally efficient data augmentation method of Imai and van Dyk (2005) is used. Due to the SAR component of the model only interacting with the rest of the model through β^{**}, the implementation is fairly straightforward.

The likelihood of the posterior distribution without identified parameters, (β and Σ), is

$$\begin{aligned} P(\beta, \Sigma, \rho, \lambda, Z^{**}|Y^{**}) \quad \propto \quad & \exp\Big(\left(Z^{**} - U^{**}\beta^{**}\right)^T \Omega \left(Z^{**} - U^{**}\beta^{**}\right) \\ + \quad & \left(\beta - \beta_0\right)^T \left((I-B)^{-1}\Sigma_0(I-B^T)^{-1}\right)^{-1} \left(\beta - \beta_0\right) \\ + \quad & \log|\Sigma|(\nu + J + 1) + \operatorname{tr}\left(S\Sigma^{-1}\right)\Big) \\ \cdot \quad & \mathbb{I}_{\{\rho \in (a_1, b_1)\}} \end{aligned} \tag{5.4}$$

The hyper parameters for $\boldsymbol{\beta}^{**}$ are β_0 and Σ_0. β_0 is a vector of length nJ and Σ_0 is a covariance matrix of dimension $nJ \times nJ$. The hyper parameters for $\boldsymbol{\Sigma}$ are the degrees of freedom parameter $\nu > J$, and S, the expectation of $\boldsymbol{\Sigma}$.

5.1.1 Properties of $\mathbf{Z}^{**}$

The posterior distribution of $\mathbf{Z}^{**}$ is identical to that found in Imai and van Dyk (2005) and McCulloch and Rossi (1994). This is the distribution provided in (2.36) with the exception that deviates are generated for each of the $T - l$ years of data in the model.

5.1.2 Properties of $\boldsymbol{\beta}^{**}$

In this model there is a unique β^{**} for each ALCU, prior state, and crop type; $\boldsymbol{\beta}^{**}$ is arranged such that

$$\boldsymbol{\beta}^{**} = \left(\boldsymbol{\beta}^{**}_{1,1}, \ldots, \boldsymbol{\beta}^{**}_{J,1,1}, \ldots, \boldsymbol{\beta}^{**}_{J,P,1}, \ldots, \boldsymbol{\beta}^{**}_{J,P,n}\right). \tag{5.5}$$

This particular configuration of β^{**} is a consequence of the spatial autocorrelation imposed on β^{**}. The spatial correlation of β^{**} allows for spatially close ALCUs to have similar regression coefficients for each prior crop sequence.

Using marginal data augmentation, the identified parameter $\tilde{\beta}^{**}$ under a Bayesian approach has a conditional distribution for gibbs sampling similar in form to Algorithm 2 in Imai and van Dyk (2005) when $\beta_0 \neq 0$ and Algorithm 1 when $\beta_0 = 0$. For Algorithm 2 in Imai and van Dyk (2005) the posterior distribution of the identified parameter is,

$$\tilde{\boldsymbol{\beta}}^{**} | \Theta_{-\beta^{**}} \sim \mathcal{N}\left(\hat{\beta}, (\alpha^2)^* \left(A + \sum_{t \in \{l, \cdots, T-l\}} (U^{**})^T \Sigma^{-1} U^{**}\right)^{-1}\right) \tag{5.6}$$

where,

$$\hat{\beta} = \left(A + \sum_{t \in \{l, \cdots, T-l\}} (U^{**})^T \Sigma^{-1} U^{**} \right)^{-1} \left(A\beta_0 + \sum_{t \in \{1, \cdots, T-l\}} (U^{**})^T \Sigma^{-1} Z^{**} \right), \tag{5.7}$$

A is $\left((I-B)^{-1}\Sigma_0(I-B^T)^{-1}\right)^{-1}$ and, Θ_{-a} is the set of parameters minus parameter a. The temporal domain is specified over $l+1$ to T, where l is the lag, length of the crop sequence used for this model. $\mathbf{\Sigma}$, Z, and $(\alpha^2)^*$ all follow their respective conditional distributions in Imai and van Dyk (2005).

5.1.3 Properties of $\mathbf{\Sigma}$

Under marginal data augmentation $\alpha^2\Sigma = \tilde{\Sigma}$. In the model (5.1) the conditional distribution for the identified parameter $\tilde{\mathbf{\Sigma}}$ is identical to the specification given in Imai and van Dyk (2005). This equivalence is due to the spatial correlation being provided through $\boldsymbol{\beta}$ instead of $\mathbf{Z}^{**}$ as in LeSage and Pace (2009).

5.1.4 Properties of $\boldsymbol{\rho}$

The parameter $\boldsymbol{\rho}$ from $\mathbf{B} = \boldsymbol{\rho} W \otimes I_{np}$ follows the same distribution as in LeSage and Pace (2009), and implemented using the sparse techniques found in Bivand (2015). The conditional distribution of $\boldsymbol{\rho}$ given the rest of the parameters in the posterior distribution follows,

$$\begin{aligned} \boldsymbol{\rho}|\Theta_{-\rho} &\sim f(\boldsymbol{\rho}|\Theta_{-\rho}) \\ &\propto \exp\left(\tfrac{-1}{2} (\beta - \beta_0)^T \left((I - \mathbf{B})^{-1} \Sigma_0 (I - \mathbf{B}^T)^{-1} \right)^{-1} (\beta - \beta_0) \right) \mathbb{I}_{\{\rho \in (a_1, b_1)\}}. \end{aligned} \tag{5.8}$$

In this dissertation $\boldsymbol{\rho}$ is sampled using a Metropolis-Hastings based approach. This approach generates deviates based on a stationary Markov chain.

5.2 Computational Burden

Most of the parameters in (5.4) are quite sparse and special handling must be used for even a small number of observations, prior states, and crops. The greatest computational burden comes from two sources, the first source is the computational cost to invert the covariance of Z^{**}. The second source is from the log determinant used in the Metropolis-Hastings sampling of $\boldsymbol{\rho}$.

Sparse matrix handling, with optimized Cholesky decompositions and inversions were used through the *Matrix* package in R (see Bates and Maechler, 2015). Further reduction in computational time was achieved through two implementation decisions. The first decision, was to assume independence of the prior states in the covariance matrix. This makes the covariance matrix separable over the prior state, and each prior state can then be handled separately for calculating inverses.

The second decision was to break the spatial support into a set of overlapping regions. Estimation and prediction is done independently in each region. Since multiple estimates are produced for ALCUs where regions overlap, the ALCU estimate at the region with the closes centroid is used.

5.3 Model Diagnostics

To test the convergence of the gibbs sampling, the process as described in Imai and Van Dyk (2005) was followed. In this process, three Monte Carlo Markov (MCMC) chains from the gibbs sampler were generated with different initial parameters, and compared against each other through the Gelman-Rubin convergence diagnostic statistic (Gelman and Rubin, 1992). The Gelman-Rubin statistic is simply an analysis of variance (ANOVA), with the null Hypothesis that all the chains for the parameter of interest have the same location parameter.

5.4 Prediction of Land Cover

The posterior predictive distribution is estimated through gibbs sampling to predict future land use. The posterior predictive distribution is

$$\Pr\left(\mathbf{Y}_{T+1} = c|\mathbf{Y}_T^*\ast, \cdots, \mathbf{Y}_{T-l+1}\right) = \int \Pr\left(\mathbf{Y}_{T+1}^*\ast = c|\mathbf{Y}_T, \cdots, \mathbf{Y}_{T-l+1}, \Theta\right) f(\Theta|Y) d(\Theta) \tag{5.9}$$

where Y_t is the set of ALCU land uses for year t. For computational simplicity and presentation, the maximum a posteriori probability (MAP) estimates of Y_{T+1} are used. The MAP of $Y(\xi)_{T+1}$ is simply the class with with the highest probability at location ξ given the prior state of that ALCU, and is generated through gibbs sampling,

5.5 Application

The subset of La Porte County that was initially segmented and explored through spatial analysis was used to provide some initial results. This county subset consisted of 1486 ALCUs. The set of 1486 ALCUs were split into six overlapping regions. Each overlapping region had approximately 500 ALCUs. After prediction, the results were merged together, only retaining a single ALCU estimate for each ALCU. The ALCU estimate retained was the one from the region with the closest centroid, this was done to minimize edge effects.

The non-spatial MNP model used a small number of two year rotations based on the most popular rotations, Table 5.1. The same rotations were used for the spatial model. In this application the number of categories from the CDL were reduced to the three most popular crops, non-agriculture, and other agriculture. The categories considered for this model include corn, soybeans, winter wheat, non-agriculture and other agriculture.

Run time for parameter estimation was six hours per chain, with a total of three chains. Only the first chain was used for predicting the subsequent year via MAP. The prediction step used 2000 iterations, with a 200 iteration burn-in, and performed gibbs sampling using a Metropolis Hastings routine for the conditional distribution of ρ. Because of excessive

Table 5.1: Popular rotations of major commercial crops for a subset of La Porte County, Indiana, from 2001 to 2011.

Rotations	Percent of Rotations
Corn → Corn	14.9%
Corn → Soybeans	23.8%
Soybeans → Corn	24.2%
Non-Agriculture → Non-Agriculture	12.5%
Other	23.6%

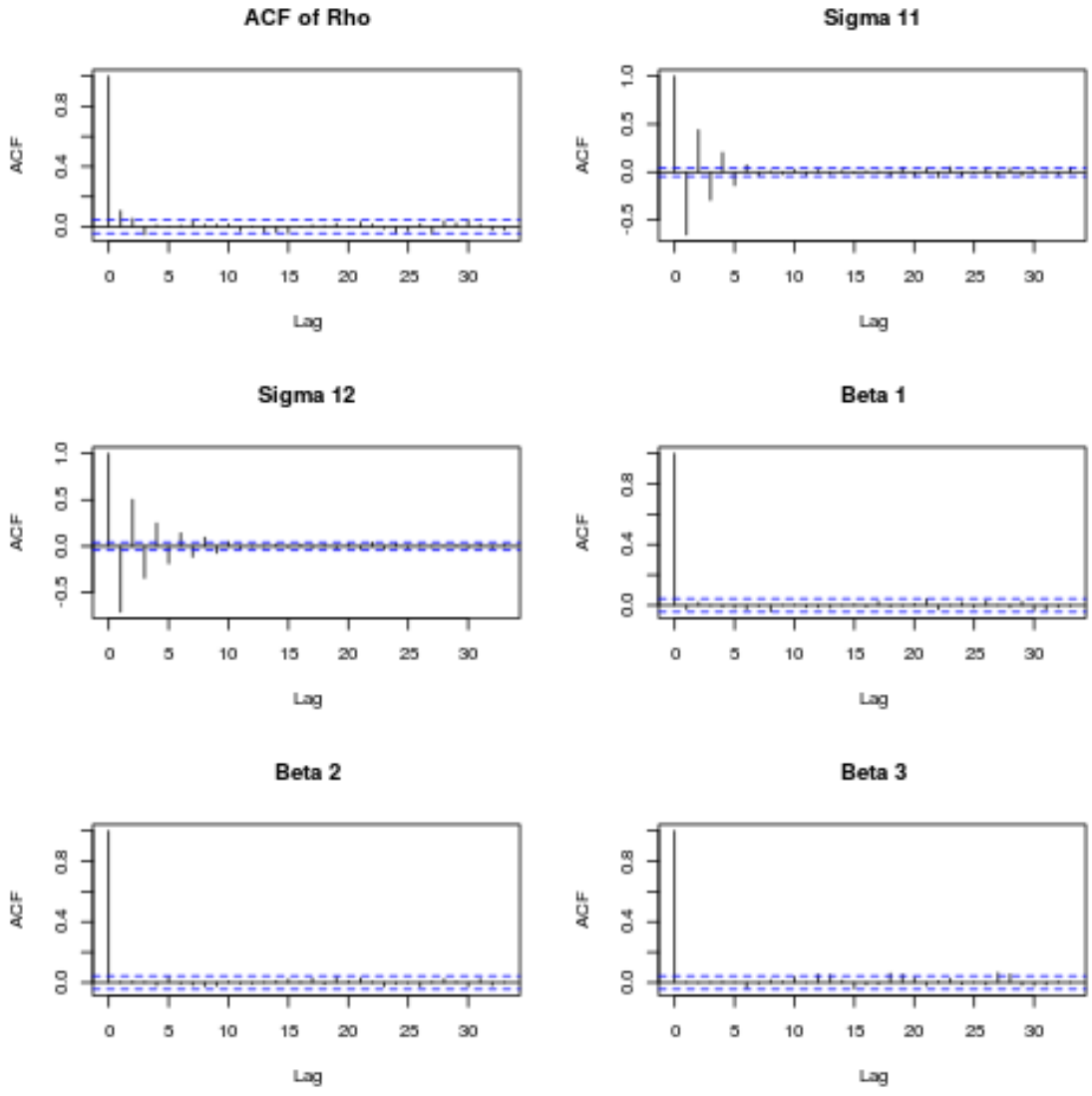

Figure 5.1: Trace plots for six parameters, ρ, Σ_{12}, $\Sigma_{1,2}$, and the first three β coefficients.

Table 5.2: Integrated misclassification rate for ALCU in La Porte County, Indiana, for five classes (corn, soybeans, winter wheat, non-agriculture and other agriculture) through CDL.

	IMR		
Year	Naïve Bayes*	Pooled MAP	Spatial MAP
2010	0.62	0.25	0.29
2011	0.64	0.59	0.54
2012	0.62	0.68	0.63
2013	0.60	0.47	0.44

correlation between draws from the Metropolis Hastings routine, only one out of every 30 draws was retained, no thinning was needed for other parameters.

Unlike in Wang et al. (2012) the parameters were extremely well behaved, with little autocorrelation between consecutive draws, Figure 5.1. There was also no sign of divergence in the β^{**} estimates, Figure 5.3. The parameter ρ did indicate a degree of positive spatial correlation, but it was quite low, Figure 5.2. Using a separate ρ for each parameter may have improved results.

The Gelman-Rubin convergence diagnostic statistic was near one for all parameters using three independent chains. This indicates lack of divergence of parameters, and an overall stable convergence of the distribution of the MCMC to the stationary distribution of this model.

Predicted land use was compared against CDL response for 2010 through 2013. Predictions were compared, conditional on the ALCU land-use sequences, using an integrated misclassification rate (IMR),

$$IMR = 1 - \frac{\sum_{\xi \in D} \mathbb{I}_{\hat{y}_{\xi,j}=y_{\xi,j}} a_\xi}{\sum_{\xi \in D} a_\xi} \tag{5.10}$$

where a_ξ is the area of ALCU ξ.

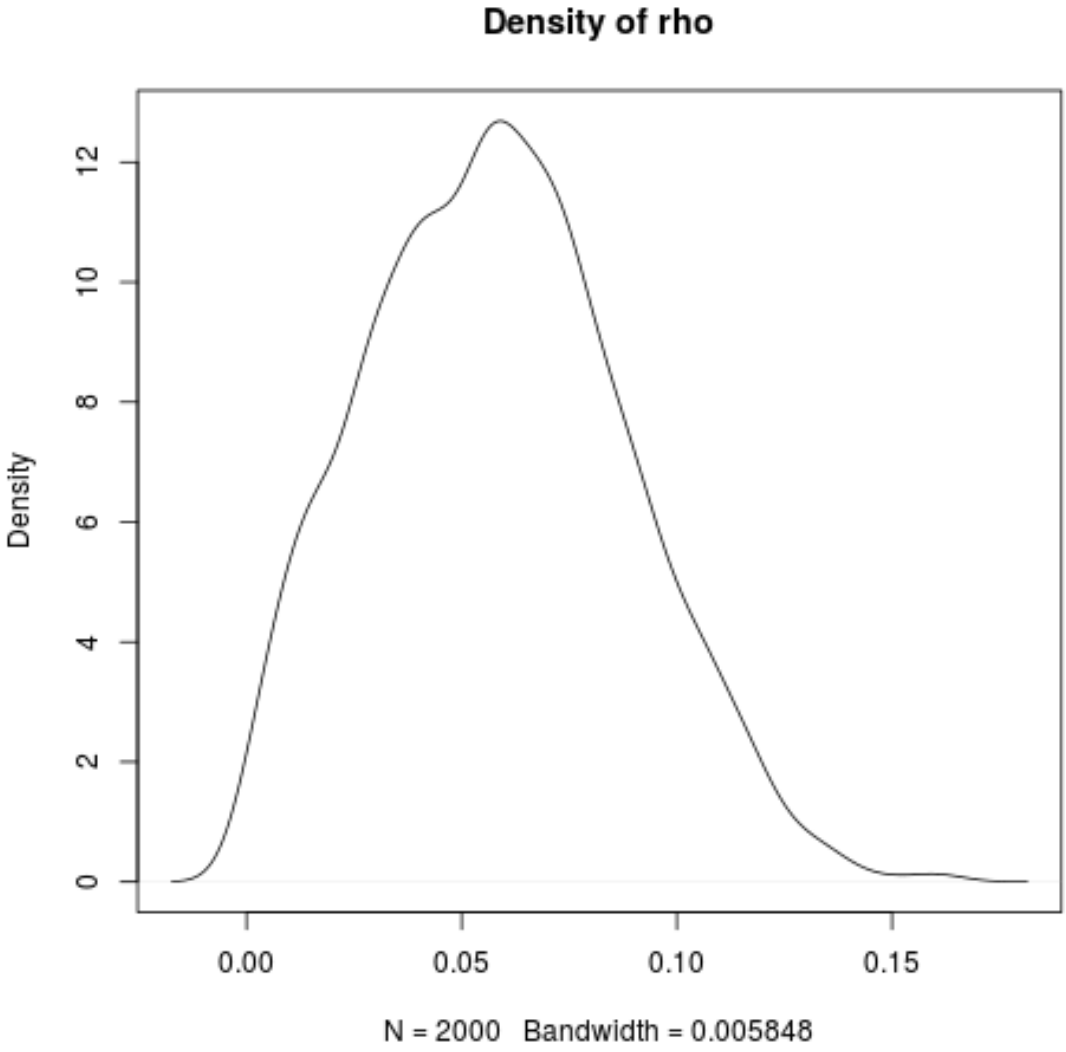

Figure 5.2: A density plot of the marginal posterior distribution of ρ, generated under gibbs sampling (using a Gaussian Kernel).

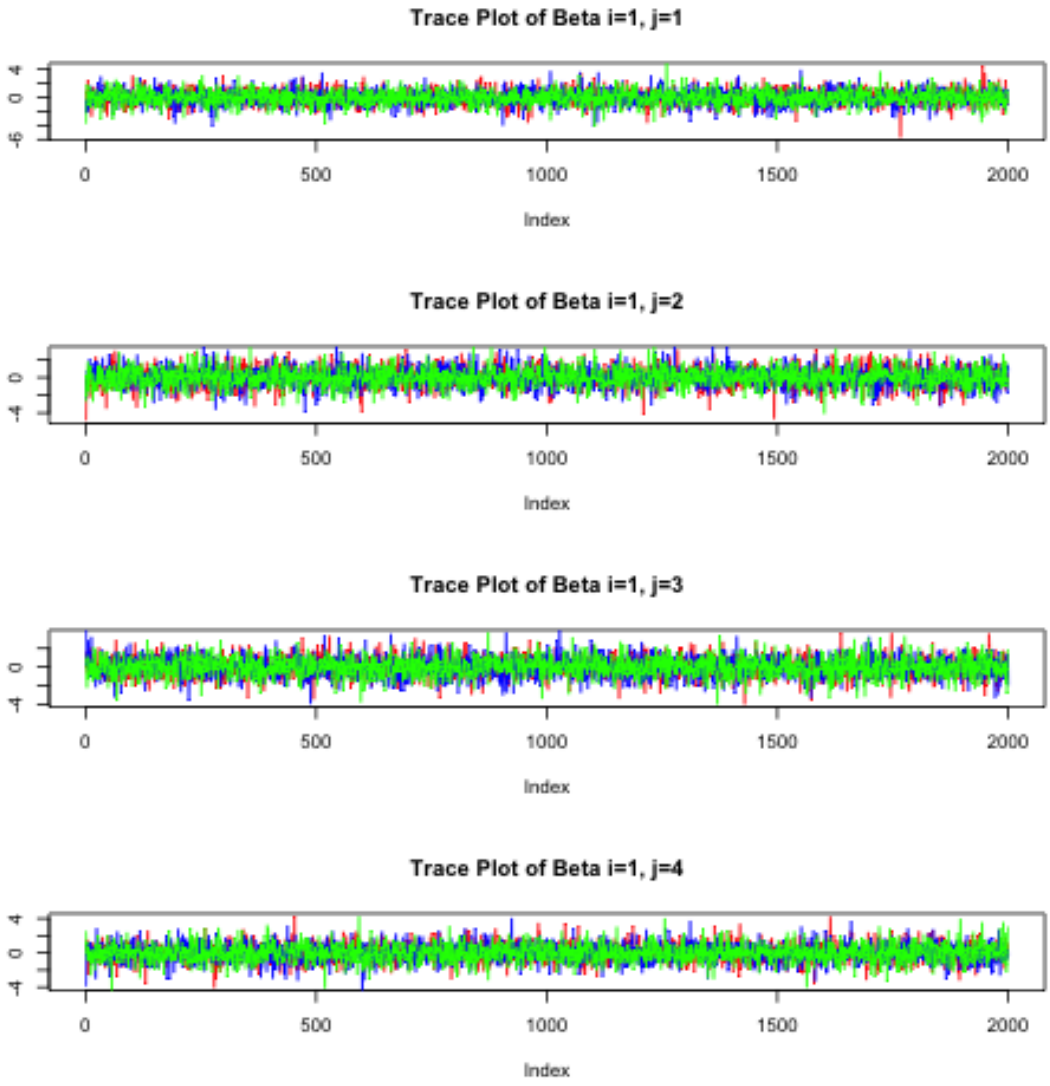

Figure 5.3: Trace plots of the first four regression coefficients β^{**}.

Chapter 6: Discussion and Future Work

In this dissertation prediction of land cover was performed using three steps:

1. Image Segmentation - to estimate LCU boundaries from high resolution imagery through the mean shift algorithm.

2. Classification - to classify LCU estimates based on coarse pre-classified pixels and other exogenous data sources.

3. Prediction - to predict LCU estimates' land cover contents using crop rotations through a spatially auto-regressive process.

Each of these steps were approached as three separate problems, each requiring a set of novel methods and applications to complete the overall task of agricultural prediction. The application of these tasks should aid in future research in agricultural production, natural resource management, survey development, and other uses of land cover and land use at a local scale. Many of the contributions have application outside the scope of land cover prediction, therefore the contributions and future work will be identified within each of the three problems. An overview of the contributions has been provided below:

- **Segmentation:**

    1. A novel well defined spatial-temporal land cover unit;
    2. A novel approach to separating edge detection from high variance structures through local variances;
    3. A novel combination of the mean shift fixed point iterator and Newton's method fixed point iterator under a Gaussian kernel;

4. A novel implementation of the mean shift algorithm using the observed property that the mean shift sequences tend to merge into a few distinct paths approaching a stationary point;
5. A novel image stratification and sampling method for mean shift segmentation of spatial images;
6. An improvement to the existing local pivotal method used for spatially balanced sampling (Grafström et al., 2012), reducing the computational complexity from $\mathcal{O}\left(n^2\right)$ to $\mathcal{O}\left(\log(n)n\right)$.

- **Classification:**

1. A test using the Black-Black join count statistic and the binomial distribution;
2. An analysis of useful properties for identifying non-agricultural land cover through penalized logistic regression;
3. A method to detect and merge of LCU estimates through maximization of the joint likelihood of adjacent LCU classifications.

- **Prediction:**

1. A novel spatial-temporal model for crop rotation is presented;
2. Computational methods to handle the sparse structure of the proposed model are presented;
3. An evaluation of the proposed model against a real data set is provided.

6.1 Segmentation

A key issue with prior GIS work is the lack of well defined spatial-temporal units. In this dissertation an explicitly well defined spatial-temporal unit was provided based on land cover sequences. This unit, called an LCU, was defined. The difficulty of creating well defined temporal spatial units was discussed. This well defined unit is of great use in land

cover and land use research, where there is a lack of papers addressing well defined spatial units, let alone spatial-temporal land cover units.

Future work on creating well defined spatial-temporal units is certainly needed, as the choice of units is application specific. One potential avenue of research is the development of dynamic units, in particular the description of spatial-temporal units through stochastic functions such as random sets, see Cressie (1993).

In this dissertation, a number of novel improvements are provided for the mean shift classification method. These improvements include generalized and application specific improvements.

In this dissertation, the log variance filter (LVF) was introduced, and a Monte Carlo study was performed. The Monte Carlo study showed that the LVF can considerably improve the performance of mean shift based classification when local variance describes particular features such as trees and buildings. This filtered estimate also provides edge detection.

A novel combination of the mean shift fixed point iterator and Newton's method for fixed point iterator under a Gaussian kernel was introduced. The combined method known as Normal Newton Shift (NNS) provided under simulation close to 33% reduction in the number of iterations required for convergence of the algorithm with little loss in quality.

A novel algorithm for merging mean shift sequences, the "Dual Tree Merge-Path Algorithm", based on the path of ascent to the local maxima was presented. This algorithm was shown to improve performance considerably when more than 10 iterations are required for convergence of the algorithm. An implementation of this algorithm was provided through an R package, and supports multiple CPU's for improved performance.

A method to perform stratified sampling on an image was presented. This stratified sampling was performed to decrease the computational burden of the mean shift algorithm, and to ensure important features such as roads are included in the sampled image. These strata were created using a novel technique employing the LVF and thresholding.

A balanced sampling method was used for stratified sampling within the low LVF stratum. The current "fast" implementation of this method has an average computational complexity of order $\mathcal{O}\left(n^2\right)$ (Grafström et al., 2012). A k-d tree based implementation was introduced in this dissertation, providing average computational complexity of order $\mathcal{O}\left(\log(n)n\right)$. The new implementation has been made available as an R package.

The implementation of a method to use U.S. Census Bureau edge data to reduce the image segmentation problem was presented. This method allows for computationally efficient image segmentation over large areas of land. An evaluation of the entire procedure was performed using hand segmented data, providing an adjusted Rand index of 0.7.

Future work in this area involves the immediate application to more diverse areas of agriculture, and evaluation of other filtered estimates, such as wavelets. A method to smooth out the edges in field boundaries should be considered, and a method to deal with shadows should be identified.

The theoretical background of NNS still remains fairly weak, and future work should be performed to relax the conditions of convergence. The "Dual Tree Merge-Path Algorithm" should be evaluated on other images, and for other kernels.

The local pivotal method sampling algorithm improved in this dissertation can be further sped up through the use of approximate nearest neighbors. Research on the computational gains verse the quality of the approximation should be performed, where quality is defined by the B statistic (2.10).

6.2 Classification

An evaluation of a test to identify under segmentation was performed on simulated and real data. This test incorporates an adjustment for handling multiple years of data, and is conditioned on both the majority crop type and the presence of precise classification rate. This test was shown to be fairly powerful under the presence of an interior LCU in simulation. Applications to real data identified limitations of the test, namely the test is limited by the presence of non-identical crop sequences of adjacent fields.

Logistic regression with an $L1$ penalty function was explored and tested as a means to identify non-agricultural LCUs from agricultural LCUs (ALCUs). The $L1$ penalty provides model selection, and helped identify a small number of useful parameters. These parameters were not associated with either the high resolution imagery used, nor the classified pixels. Instead the parameters identified included LCU properties, namely the ratio of LCU edge length to area and the presence of a GIS feature intersection.

A method to merge ALCUs was explored. This method used the likelihood that the NAIP imagery and CDL pixels describe both a single crop type and the same LCU. This merging used simulated annealing to maximize the likelihood that two neighboring ALCUs has the same parameters If the likelihood of the merged parcels exceed the less restrictive joint likelihood of the separate parcels, then the merge was accepted. This likelihood was applied using NAIP intensity properties and CDL pixel classifications. This method was evaluated against 200 hand-drawn LCUs.

The tests for under and over segmentation provided here are still quite primitive, and further work should be performed to improve these tests or find alternatives. More object based identification methodology (see Yan and Roy, 2014) may aid in ALCU identification from non-agricultural LCUs.

6.3 Prediction

The temporal spatial multinomial probit model is extremely well behaved and provides a slight increase in classification rates for the application to La Porte County, Indiana, relative to non-spatial methods. This model is novel in form and describes the phenomena well, and should provide an excellent model for use in agricultural economics related to agricultural land cover and land use change. The model also provides a working SAR MNP approach that has been difficult to accomplish in prior papers (Wang et al., 2012).

The multiple support method used to retain computational feasibility is currently ad-hoc, and future research should be performed to identify more optimal ways to share the support under multiple models. The model itself was only tested on a small number of

crops and prior states, future work should be done to both try this method on more diverse crops and expand the number of prior states. Computational feasibility may be improved through GPU computing to accelerate the calculation of the inverse of the covariance of $\mathbf{Z}^{**}$. Additional covariates should be identified and applied to improve this model, such as rainfall or soil data. The addition of trends, and the ability to add constraints on maximum total acres for each crop may help improve the model.

Appendix A: Appendix - Methodology Review

A.1 Kernel Density Estimation

The mean shift algorithm is based on the derivative of a kernel density estimate, therefore some knowledge about kernel density estimation is required to understand the algorithm. A minimal overview of the methodology is provided here, classical texts Silverman (1986) and Wand and Jones (1994) provide excellent and more in depth discussions of the subject. Kernel density estimation is a non-parametric statistical method to estimate the density f of a random variable $\mathbf{x}$. The kernel density estimator takes the form,

$$\hat{f}_{n,H}(v) = \sum_{j=1}^{N_R} |H|^{-1} n^{-1} \kappa \left(H^{-1} \left(v - x_j \right) \right) \tag{A.1}$$

where $\{x_i\}_{i=1}^{n}$ is a sequence of iid random variables and the function κ is known as a kernel. This can be simplified to

$$\hat{f}_{n,H}(v) = \sum_{j=1}^{N_R} n^{-1} \kappa_H \left((v - x_j) \right) \tag{A.2}$$

where

$$\kappa_H \left(u \right) = |H|^{-1} \kappa \left(H^{-1} u \right). \tag{A.3}$$

In this research kernels will be restricted to second order kernels, where the kernel order corresponds to the first non-zero moment of the kernel function. The p^{th} moment of the kernel is

$$\int u^p \kappa(u) du \tag{A.4}$$

with $u = H^{-1}(v - x_j)$, where H is a bandwidth parameter. Furthermore, to ensure that the asymptotic properties of the kernels hold four conditions are imposed:

Table A.1: Common kernels used for KDE estimators.

Kernel	Equation	Derivative	Support
Uniform	$\frac{1}{2}$	0	$u \in [-1, 1]$
Epanechnikov	$\frac{3}{4}\left(1-u^2\right)$	$-\frac{3}{2}u$	$u \in [-1, 1]$
Biweight	$\frac{15}{16}\left(1-u^2\right)^2$	$-\frac{15u}{8}\left(1-u^2\right)$	$u \in [-1, 1]$
Gaussian	$\rho(u)$	$-\frac{u}{2}\rho(u)$	$u \in \mathbb{R}$

K1 $\kappa(u) > u$ for $u \in \mathbb{R}^d$;

K2 $\int \kappa(u)du = 1$;

K3 $\int u\kappa(u)du = 0$;

K4 $\int u^2\kappa(u)du = c < \infty$.

Common kernels include Uniform, Epanechnikov, biweight and Gaussian kernels, (see Table A.1). In this research u will be vector valued with length d, under these conditions the bandwidth is a matrix. To ensure that the $\hat{f}_{n,H}$ exists and to simplify results, the bandwidth matrix H will be diagonal with diagonal elements h_k for $k \in \{1, \cdots, d\}$, and

K5 the bandwidth matrix H is symmetric and positive definite; for a scalar $a > 0$ and diagonal matrix $A : aA = H, |A| = 1$, as $a \to 0$, and the sample size n tends to ∞ and $an \to \infty$.

In this dissertation only kernels of the form $|H|^{-1}g(||H^{-1}(v - x_j)||_2^2)$ will be considered, and for simplicity $||x - v_j||_H^2 = ||H^{-1}(v - x_j)||_2^2$.

Since f is a function, properties of estimators are given at single points in the domain of f or over the entire domain through integration. The pointwise properties include the bias, variance and MSE, to provide simple workable forms of these properties the function f is approximated through a Taylor series and hence the properties are prefixed "asymptotic." Because these are asymptotic results the utility of these approximations are a function of the local linearity about v, and in the global sense for the properties integrated over the

domain. The asymptotic pointwise properties of $\hat{f}_H$ under K1–K5 are

$$\text{bias}\left(\hat{f}_{n,H}(v)\right) = \frac{c}{2}\sum_{k=1}^{d} h_k^2 \frac{\partial^2 f(v)}{\partial v_k} + \mathcal{O}\left(\text{tr}\left(H^2\right)\right), \tag{A.5}$$

$$\text{var}\left(\hat{f}_{n,H}(v)\right) = n^{-1}|H|^{-1}R(\kappa)^d f(v) + o\left(n^{-1}|H|\right) \tag{A.6}$$

with $R(f) = \int f(x)^2 dx$, and

$$\text{MSE}\left(\hat{f}_{n,H}(v)\right) = \left(\frac{c}{2}\sum_{k=1}^{d} h_k^2 \frac{\partial^2 f(v)}{\partial v_k}\right)^2 + n^{-1}|H|^{-1}R(K)^d f(v). \tag{A.7}$$

Of the global characteristics the one of greatest interest is the AMISE, asymptotic integrated squared error, this has the closed form of

$$\text{AMISE}\left(\hat{f}_{n,H}\right) = \left(\frac{c}{2}\right)^2 R(f'')h^4 + \frac{R(\kappa)}{nh^d} \tag{A.8}$$

with the restriction $h = h_k$. The derivation of these quantities can be found in Silverman (1986) and Wand and Jones (1994).

The value of h that minimizes (A.8) is $h \propto n^{-2/(d+2)}$ and is termed the AMISE optimal bandwidth. Under these conditions $\hat{f}_{n,H}(v)$ is a consistent estimator of $f(v)$ at point v with order $\mathcal{O}_p\left(n^{-4/(d+4)}\right)$.

It should be noted that as d increases the slower the bias approaches 0, this is one of several issues with kernel density estimates in high dimensions. Another issue is the computation of high dimensional kernel density estimators. Binning is one simple way to increase the computational speed of kernel density estimation, where the bin sizes are set to guarantee some level of numerical precision. However, as the number of dimensions increase the number of bins required to retain that level of precision increases exponentially.

Table A.2: Common kernels and profiles

Name	Notation	Profile	Use
Kernel	γ	g	Mean Shift Iterator
Shadow Kernel	κ	k	KDE used for clustering

A.2 Mean Shift

Mean shift is a mode searching algorithm similar to the typical Newton's Method for finding roots of a smooth function but applied to a kernel density estimator. A short overview of the algorithm and its properties used in this dissertation are provided in this section. For further study Chen et al. (2014a) provides a detailed overview of the algorithm with an extensive set of applications.

As a clustering method a set of *query* points Q are classified by steepest ascent to a mode from the kernel density estimator of another set of *reference* points R where each point v in R or Q is assumed to be an iid observation of the random variable $\mathbf{x}$ with pdf f with support in $\mathbb{R}^d$. Q and R are assumed to both have the same support, and frequently $Q = R$. Because this is a root searching method, there is no dependency on a fixed number of clusters, instead the choice of both kernel and bandwidth parameters are determine the number of modes in the kernel density estimate (KDE). For convenience, the cardinality of the sets Q and R will be denoted by N_Q and N_R respectfully, and the KDE will be rewritten as

$$\hat{f}_{N_R,H}(v) = \sum_{j=1}^{N_R} |H|^{-1} n^{-1} \kappa\left(H^{-1}\left(v - x_j\right)\right) = \sum_{j=1}^{N_R} c_{j,H} k\left(||v - x_j||_H^2\right) \tag{A.9}$$

where $c_{j,H}$ is a constant ensuring that $k\left(||v^i - x_j||_H^2\right) c_{j,H} = \kappa\left(||v^i - x_j||_H^2\right)$. In the context of mean shift, the kernel κ is known as the *shadow kernel*; k is known as the *shadow profile*; and $g = -k'$, is the profile of γ, where γ is the kernel used for iterating in (A.10). Table A.2 provides a brief description of the kernels, profiles and uses.

The mean shift algorithm creates a set of candidate stationary points via fixed point

iterations of

$$v^{i+1} = \frac{\sum_{j=1}^{N_R} x_j c_{j,H} g\left(||v^i - x_j||_H^2\right)}{\sum_{j=1}^{N_R} c_{j,H} g\left(||v^i - x_j||_H^2\right)}. \tag{A.10}$$

This iteration continues until either a fixed number of iterations occur or the difference between two iterations are below a threshold.

The algorithm was first proposed by Fukunaga and Hostetler (1975) but did not see wide spread adoption until the publication of Cheng (1995). Cheng used the term "Natural Clustering" to describe the procedure, as the data points themselves self-organize around the mode of the density. One advantage of this method over others such as K-means is that the clustering only relies on the bandwidth parameter which depending on the problem may be easier to specify. In the case of K-means where choice of initial centroids may have considerable influence on the clustering, mean shift has the further benefit of being entirely deterministic.

A.2.1 Algorithm

The mean shift algorithm is an iterative gradient ascent algorithm (see Figure A.2.1). In the mean shift algorithm a kernel density estimate of a random variable $\mathbf{x}$ is searched for a stationary point, ideally a mode.

The locations of the local maxima in the density of $\mathbf{x}$ can be estimated by the set of v such that $\nabla \hat{f}_v(v) = 0$, where the gradient of the kernel density estimator at location v has the form

$$\nabla \hat{f}_{N_R,H}(v) = \sum_{j=1}^{N_R} -2c_{j,H} H^{-1} (v - x_j)\, k\left(||v - x_j||_H^2\right). \tag{A.11}$$

.

In the case of a normal kernel $\kappa(u) = \phi(u)$, $\frac{\partial k(u)}{\partial u} \propto uk(u)$ the roots occur where

$$v = \frac{\sum_{j=1}^{N_R} xg\left(||v - x_j||_H^2\right)}{\sum_{j=1}^{N_R} g\left(||v - x_j||_H^2\right)} = \frac{\sum_{j=1}^{N_R} x\phi\left(||v - x_j||_H^2\right)}{\sum_{j=1}^{N_R} \phi\left(||v - x_j||_H^2\right)}. \tag{A.12}$$

This is form is similar to the Nadaraya-Watson estimator from kernel regression, but differs in that the $\mathbf{x}$ is being "regressed" on neighboring values of $\mathbf{x}$. It should be noted that the Gaussian kernel is the only kernel where the kernel and the shadow kernel are equivalent (see Cheng, 1995). Given an initial location v^0, a local mode can be found by fixed point iteration,

$$v^{(i)} = \frac{\sum_{j=1}^{N_R} xg\left(||v^{(i-1)} - x_j||_H^2\right)}{\sum_{j=1}^{N_R} g\left(||v^{(i-1)} - x_j||_H^2\right)} \tag{A.13}$$

until $||v^{(i)} - v^{(i-1)}|| < \delta$ where δ is an acceptable tolerance. Each iteration of (A.13) is a convex combination of the points in the domain of g, therefore all movements of the algorithm are necessarily bounded for finite samples from $\mathbf{x}$ and occur within the convex hull of the samples. The closer $v^{(i)}$ is to the local maxima for a sufficiently smooth shadow kernel, the smaller the steps are; this avoids the complication of line searching in the similar Newton Method. Although the mean shift is considerably more stable than Newton's method, it does so with a considerable computational complexity ,$\mathcal{O}\left(n^3\right)$.

A.2.2 Convergence of the Mean Shift Algorithm

Convergence of the Mean Shift algorithm to a stationary point in the KDE over R was first presented in the literature by Cheng (1995). This was done with fairly large restrictions on the kernel choice and underlying distribution f that the KDE was estimating. Further work by Comaniciu and Meer (2002) generalized the proof to all convex kernels with strictly decreasing profiles, corrections to this proof were later provided by Li et al. (2007) introducing the requirement of finite and separate stationary points. Comaniciu and Meer (2002)

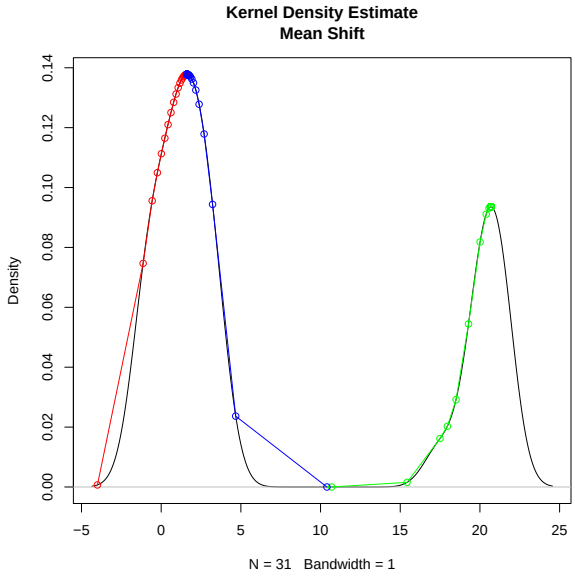

Figure A.1: Simple mean shift example on a bivariate normal density sample.

also showed that the estimate of the mode from the mean shift is an M-estimator. Clarke and Futschik (2007) for the more general case of M-estimators and later by Chen et al. (2014b) for the mean shift algorithm provided the necessary conditions for a.s. convergence of the mean shift sequence to the mode of f. Both of these results come from Giné and Guillou (2002), which provides conditions for strong uniform consistency for multivariate kernel density estimators.

A set of regularity conditions used in Li et al. (2007), Cheng (1995), and Ghassabeh et al. (2012), for example, provide sufficient conditions for convergence of the mean shift sequence $\left\{v^{(i)}\right\}_{i=0}^{\infty}$ to a stationary point within the kernel density estimate on R with kernel κ.

M1. κ meets the regularity conditions for kernels K1–K5.

M2. κ has a convex, differentiable and strictly decreasing profile k.

M3. The stationary points must be finite and each separated by a non-degenerate ball in $\mathbb{R}^d$, $0 < d < \infty$.

Under these regularity conditions it can be shown that the mean shift sequence converges to a stationary point of the KDE on R. Ghassabeh et al. (2012) provides a lemma that ensures regularity condition M3 can be met for Gaussian kernels by checking that Hessian matrix of the KDE at each stationary point is of full rank. Proofs for can be found in found in Li et al. (2007), and Wang et al. (2007) follow.

Theorem A.1 (Mean Shift sequence is a Monotone Increasing Sequence)**.** Under the prior regularity conditions, the sequence of KDEs $\left\{\hat{f}(v_{i+1}) - \hat{f}(v_i)\right\}_{i=0}^{\infty}$ is a monotone increasing sequence for any initial point v^0 in the support of $\mathbf{x}$, R, where $\{v_i\}_{i=0}^{\infty}$ follow from the mean shift algorithm.

Since the sequence is bounded and monotonically increasing it therefore trivially converges to either a point in the support or the boundary of the support $\mathbf{x}$.

Corollary 1. Under regularity conditions M1–M3, the sequence of KDEs

$$\lim_{i\to\infty} \left\{\hat{f}(v_{i+1}) - \hat{f}(v_i)\right\}_{i=0}^{\infty} = 0 \tag{A.14}$$

Theorem A.2 (Convergence of Adjacent Mean Shift Terms)**.** Under the regularity conditions M1–M3, the sequence $||v^{(i+1)} - v^{(i)}||_H^2$ converges to 0 for any initial point v^0 in the support of $\mathbf{x}$, as $i \to \infty$.

Theorem A.3 (Convergence of the Mean Shift Sequence to a Critical Point)**.** Under the regularity conditions M1–M3, the sequence $\left\{v^{(i+1)} - v^{(i)}\right\}_{i=0}^{\infty}$ converges to a critical point for any initial point v^0 in the support of $\mathbf{x}$, as $i \to \infty$.

Chen et al. (2014a) provides a set of necessary conditions and associated proof to show that the estimate of the mode from the mean shift algorithm is a consistent estimator for the mode in the density of $\mathbf{x}$. These conditions C1–C2 and the associated theorem are provided below, this result can be considered an extension of Giné and Guillou (2002).

C1. The kernel density functions are bounded and continuously differentiable up-to the third order and

$$\int x^2 \kappa^{\alpha}(x) dx < \infty \tag{A.15}$$

and

$$\int \left(\kappa^{\alpha}(x)\right)^2 dx < \infty \tag{A.16}$$

for $\alpha \in \{1, 2, 3\}$.

C2. Let P be be a probability measure on the measurable space $(S, \mathcal{S})$ where $\mathcal{S}$ is the minimal sigma algebra of S, $\mathcal{F}$ is a class of uniformly bounded functions including κ, then for $\mathrm{N}(T, D, \pi)$, the τ-covering number of the metric space (T, d) there exists some $\epsilon > 0$ and scalar $A > 0$ such that $\mathrm{N}\left(\mathcal{F}, L_2(P), \tau||F||_{L_2(P)}\right) \leq \left(\frac{A}{\tau}\right)^{\epsilon}$ for every P on $(S, \mathcal{S})$.

C1 simply allows for the Taylor series to be used to obtain a rate of convergence, while C2 admittedly is a bit terse, it just ensures that the kernel function is sufficiently smooth. Per Chen et al. (2014b) the Guassian kernel and other smooth kernels with compact support satisfy these conditions.

Theorem A.4 (Consistency of Estimating Local Modes)**.** Under the regularity conditions M1–M3, and C1–C2 the sequence $\left\{v^{(i+1)} - v^{(i)}\right\}_{i=0}^{\infty}$ converges to a critical point for any initial point v^0 in the support of $\mathbf{x}$, as $i \to \infty$.

A.2.3 Bandwidth Selection

Silverman Rule-of-Thumb is a method to choose the optimal bandwidth of a kernel density estimator based on the assumption that the "roughness" $R(k)$ can be estimated through a normal density, the net result is the choice of $h = \hat{\sigma} n^{-1/5}$ for second order kernels. The r^{th} derivative on the other hand has Rule-of-thumb bandwidth of $h \propto n^{-1/(2r+5)}$ for second

order kernels when using the derivative of the KDE as an estimator. For multivariate kernels, Chacón et al. (2011) showed that under C1 and $f \in BC^3$ that the MISE optimal bandwidth to estimate the density by the gradient of a kernel density estimator is $h = Cn^{-1/(d+6)}$ for some constant C.

Shifting from AMISE conditions to an alternative norm as seen in Arias-Castro et al. (2013), and Chen et al. (2014a),

$$||\nabla \hat{f}_n - \nabla||_{\max,\infty} = \sup_x ||\nabla \hat{f}_n(x) - \nabla f||_\infty \tag{A.17}$$

The optimal bandwidth is $h \propto \left(\frac{\log(n)}{n}\right)^{1/(d+6)}$.

The issue with bandwidth selection, from KDE based methods is that the objectives differ. In mean shift for classification the goal is to minimize the miss-classification rate, while the KDE based methods attempt to minimize the AMISE. The miss-classification may admit more bias, as the true location of the mode and density values at any given point are less important than the path of ascent. In fact, for the purposes of computational efficiency steeper paths are preferred.

Other empirical works on bandwidth selection for mean shift include University and Einbeck (2011). University and Einbeck (2011) proposed a method of determining the bandwidth parameter by first fitting curves to the data, then determining the proportion of total data points within a fixed distance from the curve.

A.2.4 Kernel Choice

Choice of a shadow kernel under the mean shift is restricted by the conditions K1–K5. In Clarke and Futschik (2007) the Gaussian kernel was identified as meeting the conditions for convergence of the Newton's Method sequence to the mode of the pdf of $\mathbf{x}$, while it is mentioned that the biweight kernel in simulation does present good results it does not meet the criteria for convergence. Chen et al. (2014a) states that the conditions for convergence of the mean shift method to both a mode of the KDE or a mode of the density of the

sample should holds for all continuous KDE's with a compact support. The use of the biweight kernel is of particular interest since is optimal in MSE for a fixed bandwidth for estimating the derivative of the pdf of $\mathbf{x}$ (see Wand and Jones, 1994 and Silverman, 1986). Empirical results under an Epanechnikov kernel were obtained in Wang et al. (2007) showing considerable performance increase over the Gaussian case.

Bibliography

S. Aksoy, I. Z. Yalniz, and K. Tasdemir. Automatic detection and segmentation of orchards using very high resolution imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(8):3117–3131, 2012.

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

M. Anderson, R. Motta, S. Chandrasekar, and M. Stokes. Proposal for a standard default color space for the internetsrgb. In *Color and imaging conference*, pages 238–245. Society for Imaging Science and Technology, 1996.

E. Arias-Castro, D. Mason, and B. Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Unpublished manuscript*, 2013.

D. Bates and M. Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2015. URL `http://CRAN.R-project.org/package=Matrix`. R package version 1.2-0.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

S. Beucher and C. Lantuéjoul. Use of watersheds in contour detection. In *International Workshop on Image Processing, Real-Time Edge and Motion Detection*, 1979.

R. Bivand. Spatial weights objects as sparse matricies and graphs, 2015. URL `https://cran.r-project.org/web/packages/spdep/vignettes/nb_igraph.html`.

C. Boryan, Z. Yang, R. Mueller, and M. Craig. Monitoring us agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, 26(5):341–358, 2011.

G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, Inc., 2008.

F. J. Breidt. Markov chain designs for one-per-stratum spatial sampling. In *Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC*, pages 356–361, 1995.

L. F. Burgette and E. V. Nordheim. The trace restriction: An alternative identification strategy for the bayesian multinomial probit model. *Journal of Business & Economic Statistics*, 30(3):404–410, 2012.

T. Büschenfeld and J. Ostermann. Edge preserving land cover classification refinement using mean shift segmentation. *Proceedings Of The 4th GEOBIA*, 2012.

R. Cai, J. D. Mullen, M. E. Wetzstein, and J. C. Bergstrom. The impacts of crop yield and price volatility on producers cropping patterns: A dynamic optimal crop rotation model. *Agricultural Systems*, 116:52–59, 3 2013. ISSN 0308521X. doi: 10.1016/j.agsy.2012.11.001.

J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679–698, 11 1986. ISSN 0162-8828. doi: 10.1109/TPAMI.1986.4767851.

V. Caselles, F. Catté, T. Coll, and F. Dibos. A geometric model for active contours in image processing. *Numerische Mathematik*, 66(1):1–31, 1993.

M. S. Castellazzi, J. Matthews, G. A. Wood, P. J. Burgess, K. F. Conrad, and J. N. Perry. Landsfacts: Software for spatio-temporal allocation of crops to fields. In *Proceedings of 5th Annual Conference of the European Federation of IT in Agriculture, Glasgow, UK*, 2007.

M. S. Castellazzi, G. A. Wood, P. J. Burgess, J. Morris, K. F. Conrad, and J. N. Perry. A systematic representation of crop rotations. *Agricultural Systems*, 97(1-2):26–33, 4 2008. ISSN 0308521X. doi: 10.1016/j.agsy.2007.10.006.

J. E. Chacón, T. Duong, and M. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 21:000–000, 2011.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Enhanced mode clustering. *arXiv preprint arXiv:1406.1780*, 2014a.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Generalized mode and ridge estimation. *arXiv preprint arXiv:1406.1803*, 2014b.

Y. Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790–799, 1995.

S.-Y. Chiu, J.-R. Zhang, and L.-S. Lan. A dual-mode mean-shift algorithm. In *Circuits and Systems, 2008. MWSCAS 2008. 51st Midwest Symposium on*, pages 334–337. IEEE, 2008.

C. K. Chu and J. S. Marron. Choosing a kernel regression estimator. *Statistical Science*, pages 404–419, 1991.

C. K. Chu, I. K. Glad, F. Godtliebsen, and J. S. Marron. Edge-preserving smoothers for image processing. *Journal of the American Statistical Association*, 93(442):526, 6 1998. ISSN 01621459. doi: 10.2307/2670100.

B. R. Clarke and A. Futschik. On the convergence of newton's method when estimating higher dimensional parameters. *Journal of Multivariate Analysis*, 98(5):916–931, 5 2007. ISSN 0047259X. doi: 10.1016/j.jmva.2006.12.002.

A. D. Cliff and J. K. Ord. *Spatial Processes: Models & Applications*, volume 44. Pion London, 1981.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002.

N. A. C. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993. ISBN 0471002550.

N. A. C. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, N.J., 2011. ISBN 9780471692744.

A. M. Dean and G. M. Smith. An evaluation of per-parcel land cover mapping using maximum likelihood class probabilities. *International Journal of Remote Sensing*, 24 (14):2905–2920, 2003.

N. Detlefsen. Crop rotation modelling. In *Proceedings of the EWDA-04 European workshop for decision problems in agriculture and natural resources. Silsoe Research Institute, England*, pages 5–14, 2004.

J.-C. C. Deville and Y. Tille. Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85(1):89–101, 1998.

V. Dey, Y. Zhang, and M. Zhong. *A Review on Image Segmentation Techniques with Remote Sensing Perspective*. na, 2010.

M. J. Falkowski and J. A. Manning. Parcel-based classification of agricultural crops via multitemporal landsat imagery for monitoring habitat availability of western burrowing owls in the imperial valley agro-ecosystem. *Canadian Journal of Remote Sensing*, 36(6): 750–762, 2010.

M. Fashing and C. Tomasi. Mean shift is a bound optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):471–474, 2005.

D. Freedman and P. Kisilev. Fast mean shift by compact density representation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1818–1825. IEEE, 2009.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*. Springer series in statistics Springer, Berlin, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

L. Friedman, N. S. Netanyahu, and M. Shoshany. Mean shift-based clustering of remotely sensed data with agricultural and land-cover applications. *International Journal of Remote Sensing*, 34(17):6037–6053, 9 2013. ISSN 0143-1161. doi: 10.1080/01431161.2013.793866.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1): 32–40, 1975.

R. M. Fuller, G. B. Groom, and A. R. Jones. Land cover map of great britain. an automated classification of landsat thematic mapper data. *Photogrammetric Engineering and Remote Sensing*, 60, 1994.

Y. Gao, N. Kerle, J. Mas, A. Navarrete, and I. Niemeyer. Optimized image segmentation and its effect on classification accuracy. *Spatial Data Quality*, 2007.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

B. Georgescu, I. Shimshoni, and P. Meer. Mean shift based clustering in high dimensions: A texture classification example. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 456–463. IEEE, 2003.

Y. A. Ghassabeh, T. Linder, and G. Takahara. On the convergence and applications of mean shift type algorithms. In *Electrical & Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference on*, pages 1–5. IEEE, 2012.

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

F. Godtliebsen and E. Spjsotvoll. Comparison of statistical methods in mr imaging. *International Journal of Imaging Systems and Technology*, 3(1):33–39, 1991.

C. A. Gotway and L. J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.

A. Grafström. *BalancedSampling: Balanced and spatially balanced sampling*, 2014. URL `http://CRAN.R-project.org/package=BalancedSampling`. R package version 1.4.

A. Grafström, N. L. P. Lundström, and L. Schelin. Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–20, 6 2012. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2011.01699.x.

A. Grafström, S. Saarela, and L. T. Ene. Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. *Canadian Journal of Forest Research*, 44(10):1156–1164, 2014.

A. G. Gray and A. W. Moore. N-body'problems in statistical learning. In *NIPS*, volume 4, pages 521–527. Citeseer, 2000.

D. A. Hennessy. On monoculture and the structure of crop rotations. *American Journal of Agricultural Economics*, 88(4):900–914, 2006.

X. Huang and L. Zhang. An adaptive mean-shift analysis approach for object extraction and classification from urban hyperspectral imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(12):4173–4185, 2008.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

K. Imai and D. A. van Dyk. A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, 124(2):311–334, 2005.

K. Imai and D. A. Van Dyk. Mnp: R package for fitting the multinomial probit model. *Journal of Statistical Software*, 14(3):1–32, 2005.

Y. Ji and S. Rabotyagov. Estimating adoption of cover crops using preferences revealed by a dynamic crop choice model. In *AAEA annual meeting, San Francisco, California*, 2015.

B. Johnson and Z. Xie. Unsupervised image segmentation evaluation and refinement using a multi-scale approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(4): 473–483, 2011.

G. Kauermann. Edge preserving smoothing by local mixture modelling. In *Sonderforschungsbereich 386*, 2001.

H. Kipka, T. R. Green, O. David, L. A. Garcia, J. C. Ascough, and M. Arabi. Development of the land-use and agricultural management practice web-service (lamps) for generating crop rotations in space and time. *Soil and Tillage Research*, 155:233–249, 2016.

M. Kulldorff, T. Tango, and P. J. Park. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, 42(4):665–684, 2003.

J.-S. S. Lee. Digital image enhancement and noise filtering by use of local statistics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pages 165–168, 1980.

J.-S. S. Lee. Digital image smoothing and the sigma filter. *Computer Vision, Graphics, and Image Processing*, 24(2):255–269, 1983.

J. P. LeSage and R. K. Pace. *Introduction to Spatial Econometrics*. CRC Press, Boca Raton, 2009. ISBN 142006424x.

X. Li, Z. Hu, and F. Wu. A note on the convergence of the mean shift. *Pattern Recognition*, 40(6):1756–1762, 6 2007. ISSN 00313203. doi: 10.1016/j.patcog.2006.10.016.

J. Lisic. Github repository. https://github.com/jlisic/MNP, 2015a.

J. Lisic. Github repository. https://github.com/jlisic/lpm3, 2015b.

J. Lisic. Github repository. https://github.com/jlisic/meanShiftR, 2015c.

M. Livingston, M. J. Roberts, and J. Rust. Optimal corn and soybean rotations. In *AAEA annual meeting, Orlando, Florida*, pages 27–29, 2008.

M. Livingston, M. J. Roberts, and Y. Zhang. Optimal sequential plantings of corn and soybeans under price uncertainty. *North Carolina State University, Working paper*, 2012.

J. A. Long, R. L. Lawrence, P. R. Miller, and L. A. Marshall. Changes in field-level cropping sequences: Indicators of shifting agricultural practices. *Agriculture, Ecosystems & Environment*, 189:11–20, 5 2014. ISSN 0167-8809. doi: 10.1016/j.agee.2014.03.015.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*, volume 37. CRC press, 1989.

R. McCulloch and P. E. Rossi. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1):207–240, 1994.

R. E. McCulloch, N. G. Polson, and P. E. Rossi. A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99(1):173–193, 2000.

X.-L. Meng and D. A. Van Dyk. Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86(2):301–320, 1999.

M. Muja and D. G. Lowe. Flann, fast library for approximate nearest neighbors. In *International Conference on Computer Vision Theory and Applications (VISAPP'09)*, 2009.

A. Nobile. A hybrid markov chain for the bayesian analysis of the multinomial probit model. *Statistics and Computing*, 8(3):229–242, 1998.

A. Nobile. Comment: Bayesian multinomial probit models with a normalization constraint. *Journal of Econometrics*, 99(2):335–345, 2000.

M. Oesterle and R. Wildmann. Land parcel identification as a part of the integrated administration and control system (iacs). *Wide Angle*, 153:6–120, 2004.

J. Osman, J. Inglada, and J.-F. Dejoux. Assessment of a markov logic model of crop rotations for early crop mapping. *Computers and Electronics in Agriculture*, 113:234–243, 4 2015. ISSN 0168-1699. doi: 10.1016/j.compag.2015.02.015.

J. Panaretos, S. Psarakis, E. Xekalaki, and D. Karlis. The correlated gamma-ratio distribution in model evaluation and selection. Technical report, University Library of Munich, Germany, 2005.

M. H. Quenouille. Problems in plane sampling. *The Annals of Mathematical Statistics*, pages 355–375, 1949.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

M. Richardson and P. Domingos. Markov logic networks. *Mach Learn*, 62(1-2):107–136, 1 2006. ISSN 0885-6125. doi: 10.1007/s10994-006-5833-1.

P. A. Rogerson. The detection of clusters using a spatial version of the chi-square goodness-of-fit statistic. *Geographical Analysis*, 31(2):130–147, 1999.

R. Sahajpal, X. Zhang, R. C. Izaurralde, I. Gelfand, and G. C. Hurtt. Identifying representative crop rotation patterns and grassland loss in the us western corn belt. *Computers and Electronics in Agriculture*, 108:173–182, 2014.

O. Schabenberger and C. A. Gotway. *Statistical Methods for Spatial Data Analysis*. CRC Press, 2004.

M. Schönhart, E. Schmid, and U. A. Schneider. Croprota a crop rotation model to support integrated land use assessments. *European Journal of Agronomy*, 34(4):263–277, 5 2011. ISSN 11610301. doi: 10.1016/j.eja.2011.02.004.

N. Sharma, R. Bedi, and A. kumar Dogra. An approach to extract road from colour image using vectorization. *International Journal*, 3(11), 2013.

L. H. Shoemaker. Fixing the f test for equal variances. *The American Statistician*, 57(2): 105–114, 5 2003. ISSN 0003-1305. doi: 10.1198/0003130031441.

Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, Boca Raton, 1986. ISBN 0412246201.

G. M. Smith and R. M. Fuller. An integrated approach to land cover classification: an example in the island of jersey. *International Journal of Remote Sensing*, 22(16):3123–3142, 2001.

M. Song and D. Civco. Road extraction using svm and image segmentation. *Photogrammetric Engineering & Remote Sensing*, 70(12):1365–1371, 2004.

S. E. Spielman, D. Folch, and N. Nagle. Patterns and causes of uncertainty in the american community survey. *Applied Geography*, 46:147–157, 2014.

D. L. Stevens and A. R. Olsen. Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465):262–278, 2004.

K. Taşdemir and C. Wirnhardt. Neural network-based clustering for agriculture management. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–13, 2012.

K. Taşdemir, P. Milenov, and B. Tapsall. A hybrid method combining som-based clustering and object-based analysis for identifying land in good agricultural condition. *Computers and Electronics in Agriculture*, 83:92–101, 2012.

M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528, 6 1987. ISSN 0162-1459. doi: 10.2307/2289457.

C. Thenail, A. Joannon, M. Capitaine, V. Souchère, C. Mignolet, N. Schermann, F. Di Pietro, Y. Pons, C. Gaucherel, V. Viaud, and J. Baudry. The contribution of crop-rotation organization in farms to crop-mosaic patterning at local landscape scales. *Agriculture, Ecosystems & Environment*, 131(3-4):207–219, 6 2009. ISSN 01678809. doi: 10.1016/j.agee.2009.01.015.

C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision*, pages 839–846. IEEE, 1998.

R. Trias-Sanz, G. Stamon, and J. Louchet. Using colour, texture, and hierarchial segmentation for high-resolution remote sensing. *ISPRS Journal of Photogrammetry and remote sensing*, 63(2):156–168, 2008.

M. Turker and E. H. Kok. Field-based sub-boundary extraction from remote sensing imagery using perceptual grouping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 79: 106–121, 5 2013. ISSN 09242716. doi: 10.1016/j.isprsjprs.2013.02.009.

D. University and J. Einbeck. Bandwidth selection for mean-shift based unsupervised learning techniques: a unified approach via self-coverage. *Journal of Pattern Recognition Research*, 6(2):175–192, 6 2011. ISSN 1558884X. doi: 10.13176/11.288.

D. A. Van Dyk and X.-L. Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1), 2001.

M. M. Wall. A close look at the spatial structure implied by the car and sar models. *Journal of Statistical Planning and Inference*, 121(2):311–324, 4 2004. ISSN 0378-3758. doi: 10.1016/s0378-3758(03)00111-3.

L. A. Waller and C. A. Gotway. *Applied Spatial Statistics for Public Health Data*, volume 368. John Wiley & Sons, 2004.

M. P. Wand and M. C. Jones. *Kernel Smoothing*, volume 60. Crc Press, 1994.

P. Wang, D. Lee, A. G. Gray, and J. M. Rehg. Fast mean shift with accurate and stable convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 604–611, 2007.

X. Wang and K. M. Kockelman. Application of the dynamic spatial ordered probit model: Patterns of land development change in austin, texas. *Papers in Regional Science*, 88(2): 345–365, 2009.

X. C. Wang, K. M. Kockelman, and J. D. Lemp. The dynamic spatial multinomial probit model: analysis of land use change using parcel-level data. *Journal of Transport Geography*, 24:77–88, 2012.

P. Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.

G. Wilkinson. Results and implications of a study of fifteen years of satellite image classification experiments. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3): 433–440, 1999. ISSN 0196-2892. doi: 10.1109/TGRS.2004.837325.

C. Xiao and M. Liu. Efficient mean-shift clustering using gaussian kd-tree. In *Computer Graphics Forum*, volume 29, pages 2065–2073. Wiley Online Library, 2010.

L. Yan and D. P. Roy. Automated crop field extraction from multi-temporal web enabled landsat data. *Remote Sensing of Environment*, 144:42–64, 3 2014. ISSN 0034-4257. doi: 10.1016/j.rse.2014.01.006.

C. Yang, R. Duraiswami, D. DeMenthon, and L. Davis. Mean-shift analysis using quasinewton methods. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II–447. IEEE, 2003a.

C. Yang, R. Duraiswami, N. Gumerov, and L. Davis. Improved fast gauss transform and efficient kernel density estimation. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 664–671. IEEE, 2003b.

H. Zhang, J. E. Fritts, and S. A. Goldman. Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2):260–280, 2008.

J. D. Zhang. *Ropencv: Wrapper of OpenCV in R*, 2014. URL `https://www.openhub.net/p/r-opencv`. R package version 0.1.1.

J.-S. Zhang. Farmland parcel extraction based on high resolution remote sensing image. *Spectroscopy and Spectral Analysis*, 29(10):2703–2707, 2009.

Q. Zhang and I. Couloigner. Automated road network extraction from high resolution multi-spectral imagery. In *ASPRS 2006 Annual Conference, Reno, Nevada*, pages 1–10, 2006.

S. Zimmer, J.-K. Kim, and S. Nusser. A hierarchical clustering algorithm for multivariate stratification in stratified sampling. *proceedings of the Survey Research Methods Section, American Statistical Association*, pages 4790–4800, 2012.

Biography

Jonathan J. Lisic received his Bachelor of Arts in Mathematics from Western Washington University in 2003. He went on to receive his Masters of Science in Statistics at the University of Akron in 2005. He has worked for the Bureau of Labor Statistics from 2005-2009, and the United States Department of Agriculture's National Agricultural Statistical Service from 2009 to present (NASS). After finishing his Doctor of Philosophy in Computational Science and Informatics in 2015, he will continue employment at NASS.