Modeling road network time series using a hidden Markov random field

Ricardo Batista and Hao Sun
Dr. Zhengyuan Zhu

Dept. of Statistics
Iowa State University

November 11, 2021

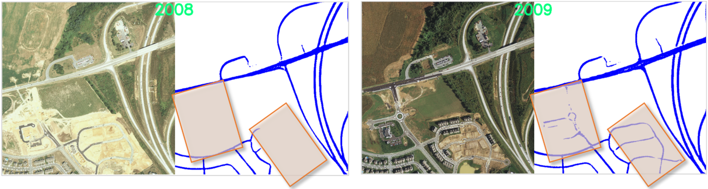
Agenda

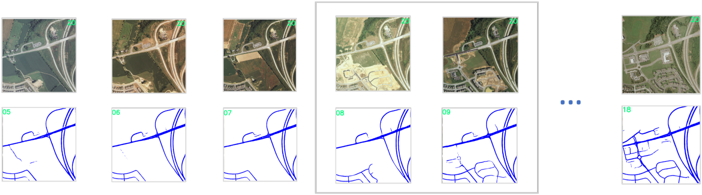
1. Motivation
2. Model formulation
3. Undirected Graph
4. Random variable definition
5. Parameter estimation

Motivation

Goal

Automated road extraction and change detection.

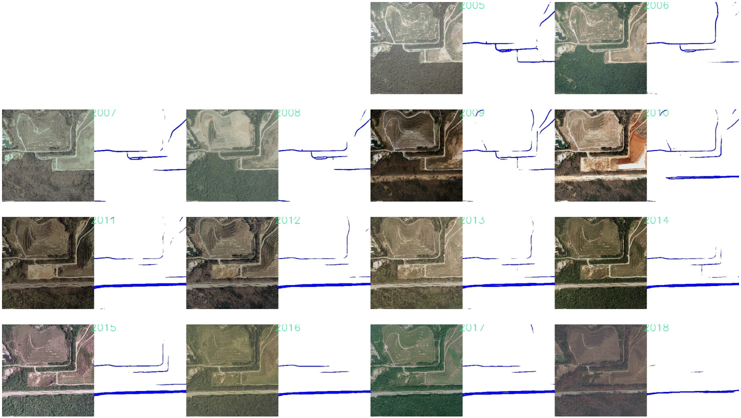




2005 2006 2007 2008 2009 ... 2018

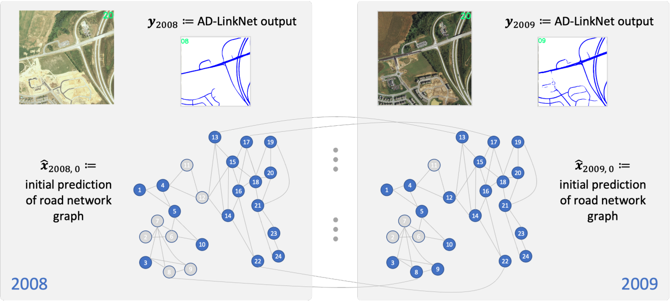
Approach combining available best-in-class methods

Current best possible approach for road extraction, i.e., **CNN segmentation** → **time-specific post-processing** → **pixelwise temporal smoothing**, leaves room for improvement.



Proposed approach

Leverage *nature of road network*: it is **graph-like** (i.e., road segments are nodes and junctions are edges) and relatively **invariant** over time.



$\hat{x}$:= most likely state (i.e., most likely evolution of road network over time) obtained from fitting a **HMRF**

Better predictions

Model formulation

Hidden Markov Random Field

Suppose we model the road network time series at location k with a

Hidden Markov Random Field (HMRF)

A bivariate probability distribution $P_{\boldsymbol{X},\boldsymbol{Y}}$ defined on an random graph (here understood as an undirected graph generated by a random process) with node set S defined by

- $P_{\boldsymbol{X}}$ is a MRF
- the conditional independence property $f(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i \in S} f(y_i|x_i)$

where $\boldsymbol{X}$'s behavior is not directly observable ("hidden").

Hidden Markov Random Field

That is, for node $i \in S = \{1, \dots N\}$,

- $P_{\boldsymbol{X}}$ is defined by $P(x_i \mid \boldsymbol{x}_{S-\{i\}}) = P(x_i \mid x_{N_i})$ where neighborhoods are defined by the edges of the underlying undirected graph, and
- $P(y_i \mid \boldsymbol{y}_{S-\{i\}}, \boldsymbol{x}) = P(y_i \mid x_i)$.

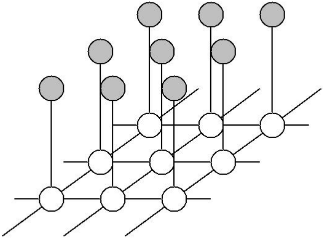


Figure 1: Abstracted HMRF where $\boldsymbol{X}$ is defined on the graph nodes (white) and $\boldsymbol{Y}$ (grey nodes) depends solely on the values of $\boldsymbol{X}$.

Computationally convenient Gibbs formulation

The **Hammersley-Clifford Theorem** states that the joint probability distribution of a Markov field $P_{\boldsymbol{X}}$ is a *Gibbs distribution* (for which we use the notation P_G) given by density

$$P_G(\boldsymbol{x}) = \frac{1}{W(\beta)} \exp\{-\beta H(\boldsymbol{x})\} = \frac{1}{W(\beta)} \exp\left\{-\beta \sum_c V_c(\boldsymbol{x}_c)\right\}$$

where

- the sum is over all cliques c in the graph
- $\boldsymbol{x}_c = \{x_i \in \boldsymbol{x} : i \in c\}$
- each V_c is a *clique potential* (i.e., any positive function that depends only on the nodes in clique c)
- $W(\beta) := \sum_{\boldsymbol{x}} \exp\{-\beta H(\boldsymbol{x})\}$ is just the normalizing constant

Implication of Hammersley-Clifford

The Gibbs formulation of $P_{\boldsymbol{X}}$ allows us to easily compute the probability of X_i given its neighbors

$$P_G(x_i|\boldsymbol{x}_{N(i)}) = \frac{\exp\left\{-\beta \sum_{c \ni i} V_c(\boldsymbol{x}_c)\right\}}{\sum_{x_i} \exp\left\{-\beta \sum_{c \ni i} V_c(\boldsymbol{x}_c)\right\}}$$

where $N(i) \subset S$ are the neighbors of i. We will leverage this fact repeatedly.

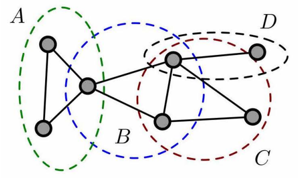
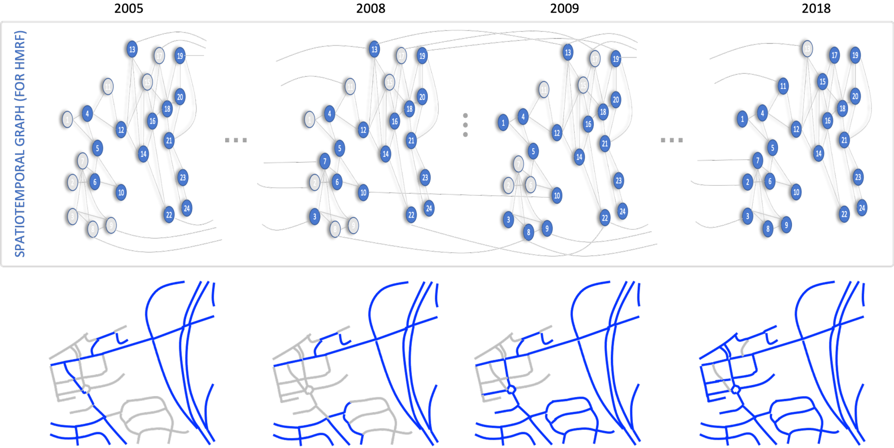


Figure 2: Recall that a **clique** is a vertex set such that every pair of vertices is adjacent (so A, B, C, and D are all cliques).

Undirected Graph

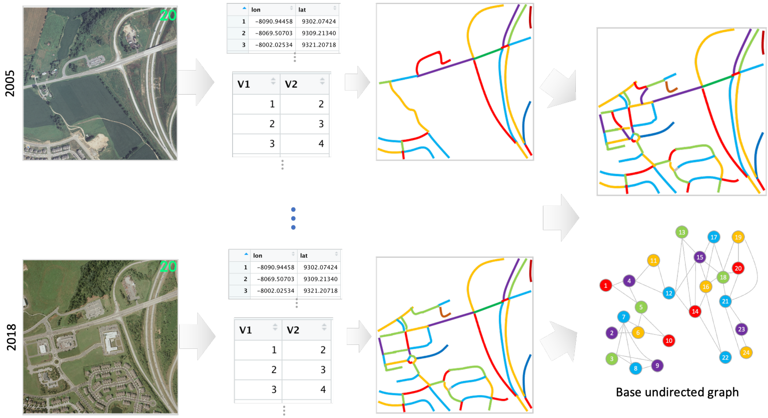
Spatiotemporal graph

Each node $s_{(t,i)}$ is the **stretch of land** at position i at time t. Note $s_{(t,i)}$ *may not be covered by road* at time t, i.e. $X(s_{(t,i)}) := X_{(t,i)}$ might equal 0 (shown in grey below).



Base undirected graph

The base undirected graph is the largest partition created from the algebra of the union of the $t = 2005, \ldots, 2018$ road fragments (but is built sequentially in practice). We can produce each G_t either based on the output of **RoadTracer** or **post-processing**.



Base undirected graph

RoadTracer

RoadTracer [Bastani et al., 2018] uses an iterative search process guided by a CNN-based decision function to derive the road network graph directly from the image and a (good) starting point for image exploration.

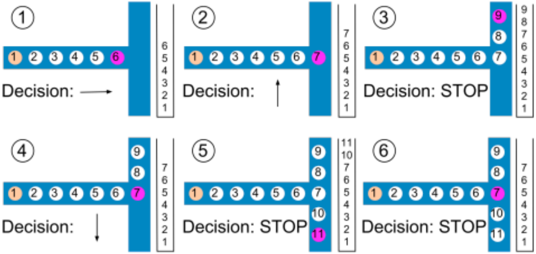


Figure 4. Exploring a T intersection in the search process. The blue path represents the position of the road in the satellite imagery. Circles are vertices in G, with S_{top} in purple and v_0 in orange. Here, the decision function makes correct decisions on each step.

Random variable definition

Unobserved r.v.

- Unobserved variable $x_{t,i}$: a binary variable, i.e. $x_{t,i} \in \{0, 1\}$, representing the stretch of land s_i is covered by road or not in year t
- Neighbor set of $x_{t,i}$: define $N^s_{ti} = \{x_{t,j} : (x_{t,i}, x_{t,j}) \in E\}$ as its spatial neighbour set and $N^t_{ti} = \{x_{t-1,i}, x_{t+1,i}\}$ as its temporal neighbour set, so that $P(x_{t,i} \mid \boldsymbol{X}/\{x_{t,i}\}) = P(x_{t,i} \mid N^s_{ti}, N^t_{ti})$.
- Conditional distribution: $x_{t,i} \mid N^s_{ti}, N^t_{ti} \sim \text{Ber}(p_{ti})$ where
$$p_{t,i} = \frac{\exp[\beta_1 \sum_{x_j \in N^s_{ti}} \delta(x_{t,i} - x_j) + \beta_2 \sum_{x_j \in N^t_{ti}} \delta(x_{t,i} - x_j)]}{\sum_{x=0}^{1} \exp[\beta_1 \sum_{x_j \in N^s_{ti}} \delta(x - x_j) + \beta_2 \sum_{x_j \in N^t_{ti}} \delta(x - x_j)]}$$

Observed r.v.

Choices for observed random variable (vector) $y_{t,i}$:

- $y_1 \in \{0,1\}$: AD-LinkNet/RoadTracer prediction
- $y_2 \in [0,1]$: Averaged conformal probability
- $y_3 \in [0,1]^{d_i}$: Raw RGB vector from satellite image, or pixel-wise conformal probability output, pixel-wise sigmoid output

Based on the type of $y_{t,i}$, We can assume the following conditional model of $f(y_{t,i} \mid x_{t,i})$

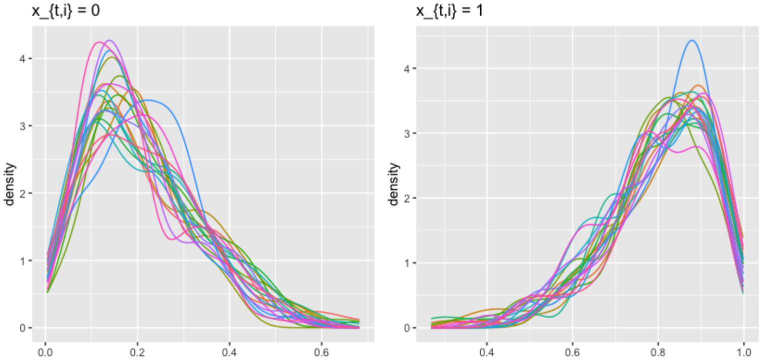
- $y_{t,i} \mid x_{t,i} \sim \text{Bern}(p_{t,i})$, if $y_{t,i} \in \{0,1\}$
- $y_{t,i} \mid x_{t,i} \sim \text{Beta}(\alpha_{t,i}, \beta_{t,i})$, if $y_{t,i} \in [0,1]$
- $y_{t,i} \mid x_{t,i} \overset{i.i.d}{\sim} f_{x_{t,i}}(r, \theta_{t,i})$ if $y_{t,i} \in [0,1]^{d_i}$

Functional Data Example

In the third case, we assume that the density function $f_{x_{t,i}}(r, \theta_{t,i})$ has form like

$$f_{x_{t,i}}(r, \theta_{t,i}) = \theta'_{t,i}\beta_{t,i}(r) + \epsilon(r), \quad r \in [0, 1]$$

and the d_i dimension random vector $y_{t,i}$ is assumed as a sample draw from $f_{x_{t,i}}(r, \theta_{t,i})$.



Parameter estimation

Context

We make the following standard assumptions about the HMRF:

- $P_{\boldsymbol{X},\boldsymbol{Y}}$ is a parametric family with parameter $\Psi = (\theta, \beta)$ and density

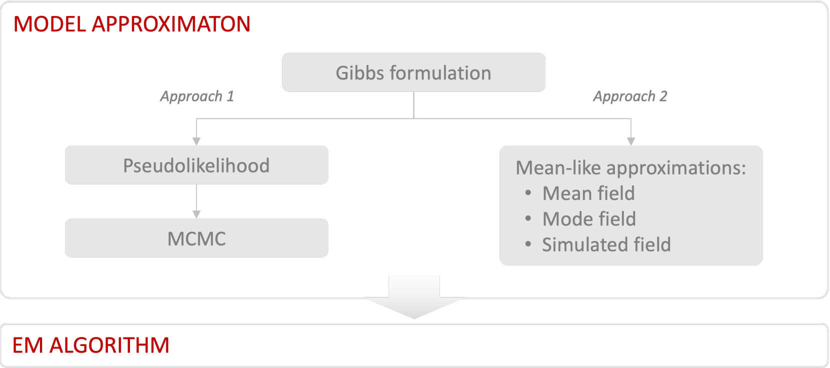
$$p(\boldsymbol{x}, \boldsymbol{y} | \Psi) = f(\boldsymbol{y} | \boldsymbol{x}, \theta) p(\boldsymbol{x} | \beta)$$

  where $f_i(y_i | x_i, \theta)$ have known forms.

- each $x_i \in L = \{1, ..., l\}$ so $\boldsymbol{x} \in L^{|S|}$ (in our case, $L = \{0, 1\}$)
- each y_i can be multivariate
- θ is also of general form, typically $\theta = (\theta_1, ..., \theta_l)$ (e.g., $\theta = (\theta_0, \theta_1)$)
- $\beta \in \mathbb{R}$ (scalar β for simplicity; in our case $\beta = (\beta_1, \beta_2)$) at least

Overview of estimation methods

Suppose we observe $\boldsymbol{Y} = \boldsymbol{y}$ and seek to perform inference on Ψ and $\boldsymbol{x}$. This is an overview of the common estimation approaches (Celeux et al. [2003]):



Section organization

- We will *walk through* the EM algorithm starting with the exact distributions and, when the procedure becomes intractable, we resort to the relevant model approximations.

- Focus is on "**approach 2**" since the MCMC simulations of "approach 1" require a large amount of computation.

General EM algorithm

Suppose we have random variables $\boldsymbol{X}$ (unobserved) and $\boldsymbol{Y}$ (observed) with (complete data) likelihood

$$L(\theta, \beta) = p(\boldsymbol{y}, \boldsymbol{x} | \theta, \beta) = f(\boldsymbol{y} | \boldsymbol{x}, \theta) p(\boldsymbol{x} | \beta)$$

Let $\Psi = (\theta, \beta)$. The **EM algorithm** seeks the MLE of the marginal likelihood $p(\boldsymbol{y}|\Psi)$ by iteratively applying these two steps:

- *E step:* Compute

$$Q(\Psi|\Psi^{(q)}) = \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\Psi^{(q)}} \left[\log p(\boldsymbol{y}, \boldsymbol{X}|\Psi)\right]$$

- *M step:* Get

$$\Psi^{(q+1)} = \arg\max_{\Psi} Q(\Psi|\Psi^{(q)})$$

Simplification

$$
\begin{aligned}
Q(\Psi|\Psi^{(q)}) &= \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\Psi^{(q)}}\left[\log p(\boldsymbol{y},\boldsymbol{X}|\Psi)\right] \\
&= \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\Psi^{(q)}}\left[\log f(\boldsymbol{y}|\boldsymbol{x},\theta)\right] + \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\Psi^{(q)}}\left[\log p(\boldsymbol{x}|\beta)\right]
\end{aligned}
$$

The first term does not depend on β while the second does not involve θ. Therefore we will write

$$
Q(\theta|\Psi^{(q)}) = \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\Psi^{(q)}}\left[\log f(\boldsymbol{y}|\boldsymbol{x},\theta)\right]
$$

$$
Q(\beta|\Psi^{(q)}) = \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\Psi^{(q)}}\left[\log p(\boldsymbol{x}|\beta)\right]
$$

Since the number of hidden states are finite and the observed variable $\boldsymbol{Y}$ is conditionally independent given $\boldsymbol{X}$, we can simplify

$$
\begin{aligned}
Q(\theta|\Psi^{(q)}) &= \sum_{\boldsymbol{x}} \log\left[f(\boldsymbol{y}|\boldsymbol{x},\theta)\right] p(\boldsymbol{x}|\boldsymbol{y},\Psi^{(q)}) \\
&= \sum_{i\in S}\sum_{\boldsymbol{x}} \log\left[f_i(y_i|x_i,\theta)\right] p(\boldsymbol{x}|\boldsymbol{y},\Psi^{(q)}) \\
&= \sum_{i\in S}\sum_{x_i} \log\left[f_i(y_i|x_i,\theta)\right] p(x_i|\boldsymbol{y},\Psi^{(q)})
\end{aligned}
$$

$$
Q(\beta|\Psi^{(q)}) = \sum_{\boldsymbol{x}} \log\left[p(\boldsymbol{x}|\beta)\right] p(\boldsymbol{x}|\boldsymbol{y},\Psi^{(q)})
$$

But now we're stuck: we don't have expressions for $p(\boldsymbol{x}|\beta)$, $p(\boldsymbol{x}|\boldsymbol{y},\Psi^{(q)})$, nor $p(x_i|\boldsymbol{y},\Psi^{(q)})$, and summing over all $\boldsymbol{x}$ is intractable. We turn to the Gibbs formulation and model approximation.

First implication of Hammersley-Clifford

Recall H-C says that if $P_{\boldsymbol{X}}$ is a MRF, it is a Gibbs distribution, i.e., we can express the density of $\boldsymbol{X}$ as $P_G(\boldsymbol{x}|\beta) = \exp\{-\beta \sum_c V_c(\boldsymbol{x}_c)\}$.

This formulation of $P_{\boldsymbol{X}}$ allows us to easily compute the probability of X_i given its neighbors

$$P_G(x_i|\boldsymbol{x}_{N(i)}) = \frac{\exp\left\{-\beta \sum_{c \ni i} V_c(\boldsymbol{x}_c)\right\}}{\sum_{x_i} \exp\left\{-\beta \sum_{c \ni i} V_c(\boldsymbol{x}_c)\right\}}$$

where $N(i) \subset S$ are the neighbors of i. We will leverage this fact repeatedly.

Second implication of Hammersley-Clifford

Expressing the density of $\boldsymbol{X}$ in Gibbs form (i.e., $P_G(\boldsymbol{x}|\beta)$)

$$\begin{aligned} P_G(\boldsymbol{y}, \boldsymbol{x}|\theta, \beta) &= f(\boldsymbol{y}|\boldsymbol{x}, \theta) P_G(\boldsymbol{x}|\beta) \\ &= W(\beta)^{-1} \prod_{i \in S} f_i(y_i|x_i, \theta) \prod_c \exp\{-V_c(\boldsymbol{x}_c|\beta)\} \\ &= W(\beta)^{-1} \exp\left\{ \sum_{i \in S} \log f_i(y_i|x_i, \theta) - \sum_c V_c(\boldsymbol{x}_c|\beta) \right\} \\ &\propto P(\boldsymbol{x}|\boldsymbol{y}, \theta, \beta) \end{aligned}$$

reveals that $P_{\boldsymbol{X}|\boldsymbol{Y}}$ is *another* MRF since, clearly, we can group the terms in the exponent into the graph's cliques, making this a Gibbs random field.

Approach 1 – Overview

"Approach 1" estimates the quantities

$$Q(\theta|\Psi^{(q)}) = \sum_{i\in S}\sum_{x_i} \log\left[f_i(y_i|x_i,\theta)\right] P_G(x_i|\boldsymbol{y},\Psi^{(q)}) \tag{1}$$

$$Q(\beta|\Psi^{(q)}) = \sum_{\boldsymbol{x}} \log\left[p(\boldsymbol{x}|\beta)\right] p(\boldsymbol{x}|\boldsymbol{y},\Psi^{(q)})$$

$$\approx \sum_{i\in S}\sum_{\boldsymbol{x}_{\overline{N(i)}}} \log\left[P_G(x_i|\boldsymbol{x}_{N(i)},\beta)\right] P_G(\boldsymbol{x}_{\overline{N(i)}}|\boldsymbol{y},\Psi^{(q)}) \tag{2}$$

where (1) is the exact quantity but uses the Gibbs form, and (2) uses the pseudolikelihood introduced by Besag [1975]. (Note $\overline{N(i)} = N(i)\cup\{i\}$.) The conditional probabilities $P_G(x_i|\boldsymbol{y},\Psi^{(q)})$ and $P_G(\boldsymbol{x}_{\overline{N(i)}}|\boldsymbol{y},\Psi^{(q)})$ are approximated using MCMC since they cannot be computed exactly.

Pseudo-likelihood

We won't dwell on "Approach 1" (given the MCMC computational demands) beyond showing the widely-used pseudo-likelihood:

An approximation of the likelihood $P_G(\boldsymbol{x}) = W^{-1} exp(-H(\boldsymbol{x}))$ is the **pseudo-likelihood** introduced by Besag [1975] and defined as

$$\mathscr{P}\mathscr{L}(\boldsymbol{x}) = \prod_{i \in S} P_G(x_i | \boldsymbol{x}_{N(i)})$$

where $N(i)$ denotes the set of neighbors of i. Recall from our coverage of the Gibbs distribution that each term in the product is easy to compute,

$$P_G(x_i | \boldsymbol{x}_{N(i)}) = \frac{exp\left\{-\sum_{c \ni i} V_c(\boldsymbol{x}_c)\right\}}{\sum_{x_i} exp\left\{-\sum_{c \ni i} V_c(\boldsymbol{x}_c)\right\}}.$$

Approach 2 – Overview

"Approach 2" uses the **mean field approximation principle** to approximate MRF $P_{\boldsymbol{X}}$ and then $P_{\boldsymbol{X}|\boldsymbol{Y}}$ to obtain

$$Q(\theta|\Psi^{(q)}) = \sum_{i\in S}\sum_{x_i} \log\left[f_i(y_i|x_i,\theta)\right] P_G(x_i|\boldsymbol{y},\Psi^{(q)})$$

$$\approx \sum_{i\in S}\sum_{x_i} \log\left[f_i(y_i|x_i,\theta)\right] P_{\tilde{\boldsymbol{x}}^{(q)}}(x_i|y_i,\Psi^{(q)})$$

$$Q(\beta|\Psi^{(q)}) = \sum_{\boldsymbol{x}} \log\left[p(\boldsymbol{x}|\beta)\right] p(\boldsymbol{x}|\boldsymbol{y},\Psi^{(q)})$$

$$\approx \sum_{i\in S}\sum_{x_i} \log\left[P_{\tilde{\boldsymbol{x}}^{(q)}}(x_i|\beta)\right] P_{\tilde{\boldsymbol{x}}^{(q)}}(x_i|y_i,\Psi^{(q)})$$

where the $P_{\tilde{\boldsymbol{x}}^{(q)}}(x_i|\cdot)$ are approximations based on configuration $\tilde{\boldsymbol{x}}^{(q)}$ obtained either via the mean, mode, or simulated field algorithm.

The mean field approximation principle

- The **mean field approximation** is originally a method to approximate the mean of a MRF $P_{\mathbf{X}}$ and can be used to provide an approximation of its *distribution*.
- The idea: approximate the effect of all the other sites on any given site i by a *single constant effect*
- For instance, for all j different from i, we fix the X_j's to their mean value $\mathbb{E}_G[X_j]$, denoted by m_j for all $j \in S \setminus \{i\}$.
- The resulting system behaves as one composed of *independent variables* for which computation gets tractable:

$$P_G(\boldsymbol{x}) \approx \prod_{i \in S} P_i^{mf}(x_i) = \prod_{i \in S} P_G(x_i | \boldsymbol{m}_{N(i)})$$

Energy function for site i

We begin by defining a new energy function for site i. Recall

$$P_G(\boldsymbol{x}) = W^{-1} \exp\{-H(\boldsymbol{x})\} = W^{-1} \exp\left\{-\sum_c V_c(\boldsymbol{x}_c)\right\},$$

where dependence on parameter β has been made implicit. Then then let

$$\begin{aligned} H_i^{mf}(x_i) \; := \; H(\boldsymbol{x})|_{x_j = m_j,\, j \neq i} \; &= \sum_{c \ni i} V_c\left((x_i, \boldsymbol{m}_{c\setminus\{i\}})\right) + \sum_{c \not\ni i} V_c\left(\boldsymbol{m}_c\right) \\ &= H_i^{mf\,loc}(x_i) + R_i^{mf\,loc}(\boldsymbol{m}_{S\setminus\{i\}}) \end{aligned}$$

where $\boldsymbol{m} := \{m_i : i \in S\}$, and subsets $\boldsymbol{m}_{S\setminus\{i\}}$, $\boldsymbol{m}_c$, and $\boldsymbol{m}_{c\setminus\{i\}}$ are analogously defined. *Crucially*, note we can decompose $H_i^{mf}(x_i)$ into the mean field local energy at pixel i, denoted by $H_i^{mf\,loc}(x_i)$, and a term, $R_i^{mf\,loc}(\boldsymbol{m}_{S\setminus\{i\}})$, that does not depend on x_i.

Approximation of $P_G(x_i)$

The mean field theory suggests that the marginal distribution of the field at site i,

$$P_G(x_i) = W^{-1} \sum_{\boldsymbol{x}_{S\setminus\{i\}}} \exp(-H(\boldsymbol{x}))$$

can be approximated by

$$P_i^{mf}(x_i) = {W_i^{mf}}^{-1} \exp(-H_i^{mf}(x_i)) = {W_i^{mf\ loc}}^{-1} \exp(-H_i^{mf\ loc}(x_i)),$$

where

$$W_i^{mf} := \sum_{x_i} exp(-H_i^{mf}(x_i)) \quad \text{and} \quad W_i^{mf\ loc} := \sum_{x_i} exp(-H_i^{mf\ loc}(x_i)),$$

which is also the conditional probability of X_i given $\boldsymbol{X}_{N(i)} = \boldsymbol{m}_{N(i)}$, i.e.,

$$P_i^{mf}(x_i) = P_G(x_i | \boldsymbol{m}_{N(i)})$$

Calculation sidebar

Note equality

$$W_i^{mf\,-1} \exp(-H_i^{mf}(x_i)) = W_i^{mf\;loc\,-1} \exp(-H_i^{mf\;loc}(x_i))$$

from the previous slide is just simple arithmetic:

$$\begin{aligned}
& P_i^{mf}(x_i) \\
& = W_i^{mf\,-1} \exp(-H_i^{mf}(x_i)) \\
& = \left[- \sum_{x_i} exp(-H_i^{mf}(x_i)) \right] \exp(-H_i^{mf}(x_i)) \\
& = \left[\sum_{x_i} exp\left\{ H_i^{mf\;loc}(x_i) + R_i^{mf\;loc}(m_{S\setminus\{i\}}) \right\} \right] \exp\left\{ -H_i^{mf\;loc}(x_i) - R_i^{mf\;loc}(m_{S\setminus\{i\}}) \right\} \\
& = \left[\cancel{exp\left\{ R_i^{mf\;loc}(m_{S\setminus\{i\}}) \right\}} \sum_{x_i} exp\left\{ H_i^{mf\;loc}(x_i) \right\} \right] \cancel{\exp\left\{ -R_i^{mf\;loc}(m_{S\setminus\{i\}}) \right\}} \exp\left\{ -H_i^{mf\;loc}(x_i) \right\}
\end{aligned}$$

Self-consistency condition

The mean field approximation of the joint distribution $P_G(\boldsymbol{x})$ is then given by the product

$$P^{mf}(\boldsymbol{x}) = \prod_{i \in S} P_i^{mf}(x_i) = \prod_{i \in S} P_G(x_i | \boldsymbol{m}_{N(i)})$$

Note, however, that, to *compute* $P_i^{mf}(x_i)$, we need the mean values at sites j different from i. But these mean values are unknown and it is *actually the goal of the approximation* to compute them.

As we shall see, mean field approximation depends on a **self-consistency condition** which is that the mean computed based on the approximation must be equal to the mean used to define this approximation.

Setting up a fixed point equation

Replace in our previous notation, the exact mean values $m_j, j \in S$ by the *mean values in the approximation*, denoted by $\bar{x}_j, j \in S$. The same expressions hold as before and we shall not modify our notation. For example,

$$H_i^{mf}(x_i) = H(\boldsymbol{x})|_{x_j=\bar{x}_j,\, j\neq i} \quad \text{instead of} \quad H_i^{mf}(x_i) = H(\boldsymbol{x})|_{x_j=m_j,\, j\neq i}$$

and let $\mathbb{E}_i^{mf}[X_i]$ denote the expectation under the new P_i^{mf}, i.e.,

$$\bar{x}_i := \mathbb{E}_i^{mf}[X_i] = {W_i^{mf}}^{-1} \sum_{x_i} x_i \exp(-H_i^{mf}(x_i))$$
$$= {W_i^{mf\,loc}}^{-1} \sum_{x_i} x_i \exp(-H_i^{mf\,loc}(x_i)),$$

where, note, the last expression is a function of just the $\{\bar{x}_j,\ j \in N(i)\}$ we will denote $g_i(\{\bar{x}_j,\ j \in N(i)\})$.

Mean field approximation

Mean field approximation consists of solving the fixed point equation

$$\bar{\boldsymbol{x}} = g(\bar{\boldsymbol{x}}) = \begin{cases} g_1(\{\bar{x}_j,\ j \in N(1)\}) \\ \vdots \\ g_n(\{\bar{x}_j,\ j \in N(n)\}) \end{cases}$$

via fixed-point iteration. We then take

- the solution $\bar{\boldsymbol{x}} = \{\bar{x}_i;\ i \in S\}$ as an estimate of the exact mean field $\boldsymbol{m}$, and
- $P_{\bar{\boldsymbol{x}}}(\boldsymbol{x}) := \prod_{i \in S} P_{\bar{\boldsymbol{x}}}(x_i) = \prod_{i \in S} P_G(x_i | \bar{\boldsymbol{x}}_{N(i)})$ as an estimate of $P^{mf}(\boldsymbol{x})$.

Important considerations

- The mean field approximation is *optimal* (in the sense of the Kullback-Leibler divergence) among systems of independent variables (Chandler [1987])
- When a solution to the fixed point equation exists, it is usually computed *sequentially* (i.e., one $\bar{x}_i$ at a time) and *iteratively*
- More generally, we talk about **mean-field-like approximations $\tilde{x}$** when the value at site i does not depend on the values at other sites which are all set to constants (*not necessarily the means*) independently of the value at site i

Approximation of $P_G(\boldsymbol{x}|\beta)$

Suppose we create, from the observations $\boldsymbol{y}$ and some current parameter estimates $\Psi^{(q-1)}$, a configuration $\tilde{\boldsymbol{x}}^{(q)}$. For each site i, set the neighbors to $\tilde{\boldsymbol{x}}_{N(i)}^{(q)}$ and replace the marginal distribution $P_G(\boldsymbol{x}|\beta)$ by

$$P_{\tilde{\boldsymbol{x}}^{(q)}}(\boldsymbol{x}|\beta) = \prod_{i \in S} P_G(x_i|\tilde{\boldsymbol{x}}_{N(i)}^{(q)}, \beta)$$

The joint distribution $P_G(\boldsymbol{y}, \boldsymbol{x}|\Psi)$ is thus replaced by

$$\prod_{i \in S} f_i(y_i|x_i, \theta) P_G(x_i|\tilde{\boldsymbol{x}}_{N(i)}^{(q)}, \beta)$$

Approximation of $P_G(\boldsymbol{x}|\boldsymbol{y}, \Psi^{(q)})$

which corresponds to an *observed likelihood* of the form

$$
\begin{aligned}
P_{\tilde{\boldsymbol{x}}^{(q)}}(\boldsymbol{y}|\Psi) &= \sum_{\boldsymbol{x}} f(\boldsymbol{y}|\boldsymbol{x}, \theta) P_{\tilde{\boldsymbol{x}}^{(q)}}(\boldsymbol{x}|\beta) \\
&= \prod_{i \in S} \sum_{x_i} f_i(y_i|x_i, \theta) P_G(x_i|\tilde{\boldsymbol{x}}^{(q)}_{N(i)}, \beta) \\
&= \prod_{i \in S} P_G(y_i|\tilde{\boldsymbol{x}}^{(q)}_{N(i)}, \Psi)
\end{aligned}
$$

Approximation of $P_G(\boldsymbol{x}|\boldsymbol{y}, \Psi^{(q)})$ (cont.)

The approximation of $P_G(\boldsymbol{x}|\boldsymbol{y}, \Psi^{(q)})$ derives naturally from the previous two slides:

$$
\begin{aligned}
P_{\tilde{\boldsymbol{x}}^{(q)}}(\boldsymbol{x}|\boldsymbol{y}, \Psi^{(q)}) &= \frac{f(\boldsymbol{y}|\boldsymbol{x}, \theta^{(q)}) P_{\tilde{\boldsymbol{x}}^{(q)}}(\boldsymbol{x}|\beta^{(q)})}{P_{\tilde{\boldsymbol{x}}^{(q)}}(\boldsymbol{y}|\Psi^{(q)})} \\
&= \prod_{i \in S} \left\{ \frac{f_i(y_i|x_i, \theta^{(q)}) P_G(x_i|\tilde{\boldsymbol{x}}_{N(i)}^{(q)}, \beta^{(q)})}{\sum_{x_i} f_i(y_i|x_i, \theta^{(q)}) P_G(x_i|\tilde{\boldsymbol{x}}_{N(i)}^{(q)}, \beta^{(q)})} \right\} \\
&= \prod_{i \in S} P_G(x_i|y_i, \tilde{\boldsymbol{x}}_{N(i)}^{(q)}, \Psi^{(q)}) \\
&= \prod_{i \in S} P_{\tilde{\boldsymbol{x}}^{(q)}}(x_i|y_i, \Psi^{(q)})
\end{aligned}
$$

Mean-field-like approximation of $Q(\Psi|\Psi^{(q)})$

Finally, having approximated MRFs $P_G(\boldsymbol{x}|\beta)$ and $P_G(\boldsymbol{x}|\boldsymbol{y},\Psi^{(q)})$, we write

$$
\begin{aligned}
Q(\theta|\Psi^{(q)}) &= \sum_{i\in S}\sum_{\boldsymbol{x}} \log\left[f_i(y_i|x_i,\theta)\right] p(\boldsymbol{x}|\boldsymbol{y},\Psi^{(q)}) \\
&\approx \sum_{i\in S}\sum_{\boldsymbol{x}} \log\left[f_i(y_i|x_i,\theta)\right] P_{\tilde{\boldsymbol{x}}^{(q)}}(\boldsymbol{x}|\boldsymbol{y},\Psi^{(q)}) \\
&= \sum_{i\in S}\sum_{\boldsymbol{x}} \log\left[f_i(y_i|x_i,\theta)\right] \prod_{i\in S} P_{\tilde{\boldsymbol{x}}^{(q)}}(x_i|y_i,\Psi^{(q)}) \\
&= \sum_{i\in S}\sum_{x_i} \log f_i(y_i|x_i,\theta) P_{\tilde{\boldsymbol{x}}^{(q)}}(x_i|y_i,\Psi^{(q)})
\end{aligned}
$$

Mean-field-like approximation of $Q(\Psi|\Psi^{(q)})$ (cont.)

and

$$
\begin{aligned}
Q(\beta|\Psi^{(q)}) &= \sum_{\boldsymbol{x}} \log\left[p(\boldsymbol{x}|\beta)\right] p(\boldsymbol{x}|\boldsymbol{y}, \Psi^{(q)}) \\
&\approx \sum_{\boldsymbol{x}} \log\left[P_{\tilde{\boldsymbol{x}}^{(q)}}(\boldsymbol{x}|\beta)\right] P_{\tilde{\boldsymbol{x}}^{(q)}}(\boldsymbol{x}|\boldsymbol{y}, \Psi^{(q)}) \\
&= \sum_{x} \log\left[\prod_{i\in S} P_{\tilde{\boldsymbol{x}}^{(q)}}(x_i|\beta)\right] \prod_{i\in S} P_{\tilde{\boldsymbol{x}}^{(q)}}(x_i|y_i, \Psi^{(q)}) \\
&= \sum_{i\in S}\sum_{x_i} \log P_{\tilde{\boldsymbol{x}}^{(q)}}(x_i|\beta) P_{\tilde{\boldsymbol{x}}^{(q)}}(x_i|y_i, \Psi^{(q)})
\end{aligned}
$$

Choosing the values $\tilde{\boldsymbol{x}}^{(q)}$

We now have a working estimation procedure. Moreover, recall we have *options* on how to produce the $\tilde{\boldsymbol{x}}^{(q)}$, namely

- based on the marginal field distribution $P_{\tilde{\boldsymbol{x}}^{(q)}}(\boldsymbol{x}|\beta)$ or the conditional field distribution $P_{\tilde{\boldsymbol{x}}^{(q)}}(\boldsymbol{x}|\boldsymbol{y}, \Psi^{(q)})$
- using the **mean**, **mode**, or **simulated** field algorithms.

We opt for the *conditional field distribution* since it has the advantage of taking the observations directly into account and several studies (Celeux et al. [2003], Archer and Titterington [2002]) give reasons dissuading from using the mean field approximation on the marginal field.

Regarding the *mean-field-like algorithms*, we present all three since none of them consistently outperforms the others across image types.

Final algorithm

1. Produce configuration $\tilde{\boldsymbol{x}}^{(q)}$, i.e., the values of the neighbors:

   Let $\tilde{\boldsymbol{x}}^{(q-1+i/n)}$ be $(\tilde{x}_1^{(q)}, \ldots, \tilde{x}_i^{(q)}, \tilde{x}_{i+1}^{(q-1)}, \ldots, \tilde{x}_n^{(q-1)})$, the configuration updated until site i. One iteration of the procedure using sequential updating, consists respectively of,

(see next slide)

Final algorithm (cont.)

1. Produce configuration $\tilde{\mathbf{x}}^{(q)}$ (cont.):

   - Mean field algorithm

$$\tilde{x}_i^{(q)} = \frac{\sum_{x_i} x_i \exp\left\{-\beta^{(q-1)} \sum_{c \ni i} V_c\left(x_i, \tilde{\mathbf{x}}_{c\setminus\{i\}}^{(q-1+i-1/n)}\right) + \log f_i(y_i|x_i, \theta^{(q-1)})\right\}}{\sum_{x_i} \exp\left\{-\beta^{(q-1)} \sum_{c \ni i} V_c\left(x_i, \tilde{\mathbf{x}}_{c\setminus\{i\}}^{(q-1+i-1/n)}\right) + \log f_i(y_i|x_i, \theta^{(q-1)})\right\}}$$

   - Mode field algorithm

$$\tilde{x}_i^{(q)} = \arg\max_{x_i} \; f_i(y_i|x_i, \theta^{(q-1)}) P_G(x_i|\tilde{\mathbf{x}}_{N(i)}^{(q-1+(i-1)/n)}, \beta^{(q-1)})$$

   - Simulated field algorithm

     $\tilde{x}_i^{(q)}$ is simulated from $P_G(x_i|y_i, \tilde{\mathbf{x}}_{N(i)}^{(q-1+(i-1)/n)}, \Psi^{(q-1)})$, which is proportional to $f_i(y_i|x_i, \theta^{(q-1)}) P_G(x_i|\tilde{\mathbf{x}}_{N(i)}^{(q-1+(i-1)/n)}, \beta^{(q-1)})$

Final algorithm (cont.)

2. EM iteration:

   - *E step*: compute $P_{\tilde{\mathbf{x}}^{(q)}}(x_i|y_i, \Psi^{(q)})$ and $P_{\tilde{\mathbf{x}}^{(q)}}(x_i|\beta)$ for all $i \in S$
   - *M step*: set $\Psi^{(q)} = (\theta^{(q)}, \beta^{(q)})$ with

$$\theta^{(q)} = \arg\max_{\theta} \sum_{i \in S} \sum_{x_i} P_{\tilde{\mathbf{x}}^{(q)}}(x_i|y_i, \Psi^{(q)}) \log f_i(y_i|x_i, \theta)$$

and

$$\beta^{(q)} = \arg\max_{\beta} \sum_{i \in S} \sum_{x_i} P_{\tilde{\mathbf{x}}^{(q)}}(x_i|y_i, \Psi^{(q)}) \log P_{\tilde{\mathbf{x}}^{(q)}}(x_i|\beta)$$

Next steps

- Continue working on adapting RoadTracer to our needs
- Finalize iterative algorithm to produce base undirected graph G
- If RoadTracer doesn't work for us, implement some post-processing method
- Fit HMRF
- Look to land uses other than roads

Thank You

References I

G.E.B. Archer and D.M. Titterington. Parameter estimation for hidden markov chains. *Journal of Statistical Planning and Inference*, 108(1):365–390, 2002. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(02)00318-X. URL https://www.sciencedirect.com/science/article/pii/S037837580200318X. C.R. Rao 80th Birthday Felicitation Volume, Part II.

Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. Roadtracer: Automatic extraction of road networks from aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4720–4728, 2018.

Julian Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):179–195, 1975. ISSN 00390526, 14679884. URL http://www.jstor.org/stable/2987782.

Gilles Celeux, Florence Forbes, and Nathalie Peyrard. Em procedures using mean field-like approximations for markov model-based image segmentation. *Pattern Recognition*, 36(1):131–144, 2003. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(02)00027-4. URL https://www.sciencedirect.com/science/article/pii/S0031320302000274.

References II

D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, 1987.