Policy optimization for dynamic spatiotemporal systems

Brian Reich
Joint work with

John Drake, UGA

Eric Laber, NCSU

Nick Meyer, NCSU

Krishna Pacifici, NCSU
Motivating example: White nose syndrome

- White-nose syndrome (WNS) is an emerging disease affecting hibernating bats.

- The first case was reported in 2007 in NY, and it has since spread rapidly throughout the Eastern US.

- It is estimated that the bat population has declined by 80% in affected areas.

- This directly affects agriculture, as bats help control insects.

- This pest control is extremely valuable.
Spread of white nose syndrome
Some potential actions being considered and applied

Both active and preventative treatments are being considered.

- Close infected caves to human access.

- Close uninfected caves to human access (currently being implemented by the Rocky Mountain region of USFWS).

- Biocontrol measures, such as introducing beneficial bacteria into caves.

- Create structures for bats that replace caves and that can be monitored and cleaned.
Statement of the general problem

- Our objective is to develop a procedure for identifying the **optimal policy** for allocating treatments in space and time.

- We define a specific reward criteria and try to find the policy which maximizes the expected reward.

- The policy must balance short-term and long-term rewards.

- We need the method to be interpretable and applicable in very high dimensions.
Other applications of this general methodology

- **Conservation**: select lands to protect or designate as easements.

- **Wildlife management**: allocating catch sizes in each county.

- **Forest fires**: deciding where and when to do controlled burns.

- **Crime prevention**: assigning police officers to prevent crime.

- **Computer network**: which Facebook users to target with ads.
Conservation design in Puerto Rico

Where and when to provide agricultural incentives?
Spread of ebola in West Africa

Where and when to close borders, institute quarantines, etc?

Day Infected

Brian Reich, NCSU
Notation

- **State:** \( S_t = \{ \text{all information collected by time } t \} \), including status of each county, covariates, spatial information, etc.

- **Action:** \( A_t = \{ \text{the locations to treat at time } t \} \).

- **Policy:** \( \pi(S|\gamma) \) (\( \gamma \) are parameters) maps states to actions

\[
A_t = \pi(S_t|\gamma).
\]

- **Reward:** \( R \), e.g., the number of uninfected counties in 2050.

- **Value:** \( V_\pi = E_\pi(R) \) is the expected reward if we adhere to \( \pi \).
There are many algorithms from RL (Sutton and Barto, 1998) that dynamically update the policy as data (state/action/reward) accrue, including:

- Stochastic dynamic programming
- Q-learning
- SARSA
- Policy search

**Canonical problem**: devise a strategy for a computer to play checkers.
How is white nose syndrome different than checkers?

- We only get to play the game once, and not for long.
- We have some time to make recommendations.
- Our state and action spaces are spatial and huge ($2^n$ states)!
- We want to exploit the scientific knowledge of our collaborators.
- We want a policy with a simple form to share with stakeholders.
- We need a policy that evolves over time and accounts for uncertainty.

Our contribution is to combine methods from RL, ecology, and statistics to accomplish these objectives within a unified framework.
Proposed policy search method

We select a simple parametric form for the policy in terms of features constructed from the current state and potential actions:

\[ A_t = \pi(S_t; \gamma) = \arg \max_{A \in A_t} \sum_{l=1}^{p} F_l(S_t, A) \gamma^l. \]

- \( A_t \) is the recommended action, for example, the set of \( m \) counties to be treated.

- \( A_t \) is the set of possible actions at time \( t \), for example, all combinations of \( m \) counties of uninfected counties.
We select a simple parametric form for the policy in terms of features constructed from the current state and potential actions:

$$A_t = \pi(S_t; \gamma) = \arg \max_{A \in A_t} \sum_{l=1}^{p} F_l(S_t, A) \gamma_l.$$ 

- $F_l$ are the features and can be any function of the states (all information available at time $t$) and actions.

- Feature selection has the usual regression tradeoffs:
  - Selecting many features gives a flexible class of policies
  - Selecting only a few features favors transparency.

- $\gamma$ weights the features and must be optimized.
We select a simple parametric form for the policy in terms of features constructed from the current state and potential actions:

\[ A_t = \pi(S_t; \gamma) = \arg \max_{A \in A_t} \sum_{l=1}^{p} F_l(S_t, A) \gamma_l. \]

- The “myopic policy” would take \( p = 1 \) and \( F_1(S_t, A) \) to be the posterior predictive mean number of sites infected at time \( t + 1 \) if we treated sites \( A \).

- Computing this feature requires fitting a spatiotemporal model.

- This is optimal if we are only concerned with the next time step, but we can do better by considering future time steps.
Feature selection

Which features will help us beat the myopic policy? We use information about medium-term predictions and network structure.

- Average transmission probability from treated counties to other counties.
- Subgraph connectivity of the treated sites.
- Betweenness connectivity of the treated sites.
- Interactions.
Outline of the computational algorithm

1. Fit a spatiotemporal model using data available at time $t$.
   - Gravity model (Maher et al., 2012) with time-evolving parameters.

2. Obtain feature weights $\gamma$ for use at time $t$.
   - The value of the policy indexed by a candidate $\gamma$ is estimated using Monte Carlo sampling from the posterior predictive distribution.
   - The optimal weight is found using stochastic gradient descent.

3. Apply recommended treatments.

4. Repeat steps 1-3 each year until the end of the study.
1. Spatiotemporal gravity model

- Let $Y_{jt} = 1$ if county $j$ is infected in year $t$ and $Y_{jt} = 0$ otherwise.

- Given the history up to time $t$ ($H_t$), the probability of an infection in county $j$ at time $t$ is

$$
\text{Prob}(Y_{jt} = 1|H_t) = \begin{cases} 
1 & Y_{jt-1} = 1 \\
1 - \prod_{l \in I_{t-1}} (1 - p_{jlt}) & Y_{jt-1} = 0
\end{cases}
$$

- $I_t$ is the set of indices of counties that are infected in year $t$.

- $p_{jlt}$ is the probability of a spread from county $l$ to county $j$ in year $t$. 
1. Spatiotemporal gravity model

The infection probability is modeled as:

\[
\text{logit}(p_{jlt}) = X_{jt}^T \beta_t + \alpha_t A_{jt} - \rho_t \frac{d_{jl}}{(m_j m_l)^\nu}.
\]

- \(X_{jt}\) are covariates and \(\beta_t\) are dynamic coefficients.
- \(A_{jt}\) indicates treatment, and \(\alpha_t\) is the treatment effect.
- The final term controls spatial dependence:
  - \(d_{jl}\) is the distance between counties,
  - \(m_j\) is the number of caves in county \(j\),
  - \(\rho_t\) and \(\nu\) are unknown parameters.
2. Obtain feature weights $\gamma$ for use at time $t$

- The value of policy $\pi(S|\gamma)$, denoted $V_\gamma$, is the expected reward if we adhere to $\pi(S|\gamma)$.

- To approximate the value, we sample many realizations of the process conditional on the current state with actions taken following $\pi(S|\gamma)$.

- These posterior predictive draws account for uncertainty in the model parameters.

- These simulations could also account for any other sources of uncertainty that can be encoded in the statistical model, such as changes in the climate or land use.
2. Obtain feature weights $\gamma$ for use at time $t$

Optimization is done through stochastic gradient descent:

0. Set some initial value $\gamma_0$ and compute its value $V_0$.

1. Generate $\gamma^* \sim N(\gamma_t, cI_p)$ and compute its value $V^*$.

2. Set $\gamma_{t+1} = \gamma_t + a_t(V_t - V^*)(\gamma^* - \gamma_t)$.

3. Repeat 1-2 until convergence.

$a_t$ and $c$ are tuning parameters.
3. Apply recommended treatments

- Given a model and policy (which give the features and $\gamma_l$), the recommended action is

$$A_t = \pi(S_t; \gamma) = \arg \max_{A \in A_t} \sum_{l=1}^{p} F_l(S_t, A) \gamma_l.$$  

- This is a challenging discrete optimization problem.

- It must be done efficiently because it is also used in Monte Carlo simulations.

- We begin by adding sites sequentially one at a time.

- Once a full set is proposed, we cycle through the proposed sites and optimize them one at a time with the others fixed until convergence.
Results: Model comparisons

We compared six variants of the gravity model

$$\text{logit}(p_{jlt}) = X_{jt}^T \beta_t + \rho_t \frac{d_{jl}}{(m_j m_l)^\nu}.$$ 

1. Static ($\beta_t \equiv \beta$), non-spatial ($\rho_t \equiv 0$)

2. Static, diffusion ($\nu \equiv 0$)

3. Static, gravity

4. Dynamic, non-spatial

5. Dynamic, diffusion

6. Dynamic, gravity
The dynamic gravity model has the smallest (best) DIC.

<table>
<thead>
<tr>
<th>Spatial model</th>
<th>Coefficient model</th>
<th>DIC</th>
<th>( \bar{D} )</th>
<th>( p_D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-spatial</td>
<td>Static</td>
<td>1307</td>
<td>1302</td>
<td>4.9</td>
</tr>
<tr>
<td>Diffusion</td>
<td>Static</td>
<td>1003</td>
<td>997</td>
<td>5.8</td>
</tr>
<tr>
<td>Gravity</td>
<td>Static</td>
<td>939</td>
<td>933</td>
<td>6.4</td>
</tr>
<tr>
<td>Non-spatial</td>
<td>Dynamic</td>
<td>1223</td>
<td>1206</td>
<td>16.7</td>
</tr>
<tr>
<td>Diffusion</td>
<td>Dynamic</td>
<td>911</td>
<td>893</td>
<td>17.7</td>
</tr>
<tr>
<td>Gravity</td>
<td>Dynamic</td>
<td>878</td>
<td>860</td>
<td>18.0</td>
</tr>
</tbody>
</table>
In addition to comparing a few models, we want to show the dynamic gravity model is adequate.

This is crucial, since our entire strategy relies on having a good predictive model.

We use the Bayesian p-value:

- Let $D_0$ be some summary statistic of the dataset, e.g., the number of infected counties in 2010.

- We sample $N = 10,000$ draws from the process, and compute $D_i$ for simulated dataset $i$.

- The p-value is then $p = \frac{1}{N} \sum_{i=1}^{N} I(D_0 > D_i)$. 

### Results: Bayes p-values for the dynamic gravity model

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Observed value</th>
<th>Predictive distribution</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n inf</td>
<td>148</td>
<td>149 ( 60 , 325 )</td>
<td>0.49</td>
</tr>
<tr>
<td>n inf 2007</td>
<td>1</td>
<td>1 ( 0 , 6 )</td>
<td>0.35</td>
</tr>
<tr>
<td>n inf 2008</td>
<td>14</td>
<td>10 ( 3 , 27 )</td>
<td>0.70</td>
</tr>
<tr>
<td>n inf 2009</td>
<td>26</td>
<td>26 ( 9 , 67 )</td>
<td>0.50</td>
</tr>
<tr>
<td>n inf 2010</td>
<td>35</td>
<td>36 ( 13 , 97 )</td>
<td>0.46</td>
</tr>
<tr>
<td>n inf 2011</td>
<td>38</td>
<td>37 ( 13 , 82 )</td>
<td>0.52</td>
</tr>
<tr>
<td>n inf 2012</td>
<td>33</td>
<td>36 ( 13 , 68 )</td>
<td>0.40</td>
</tr>
<tr>
<td>mean year</td>
<td>2010</td>
<td>2010 ( 2010 , 2011 )</td>
<td>0.37</td>
</tr>
<tr>
<td>mean long</td>
<td>1399</td>
<td>1347 ( 879 , 1605 )</td>
<td>0.64</td>
</tr>
<tr>
<td>mean lat</td>
<td>393</td>
<td>388 ( 225 , 595 )</td>
<td>0.52</td>
</tr>
<tr>
<td>min long</td>
<td>-290</td>
<td>-97 ( -1816 , 817 )</td>
<td>0.47</td>
</tr>
<tr>
<td>min lat</td>
<td>-258</td>
<td>-335 ( -826 , -187 )</td>
<td>0.80</td>
</tr>
<tr>
<td>max long</td>
<td>2156</td>
<td>2051 ( 1950 , 2201 )</td>
<td>0.82</td>
</tr>
<tr>
<td>max lat</td>
<td>1118</td>
<td>1118 ( 1016 , 1325 )</td>
<td>0.49</td>
</tr>
</tbody>
</table>
Results: Bayes p-values for the static gravity model

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Observed value</th>
<th>Predictive distribution</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n inf</td>
<td>148</td>
<td>97 (25, 222)</td>
<td>0.81</td>
</tr>
<tr>
<td>n inf 2007</td>
<td>1</td>
<td>2 (0, 6)</td>
<td>0.11</td>
</tr>
<tr>
<td>n inf 2008</td>
<td>14</td>
<td>4 (0, 11)</td>
<td>0.99</td>
</tr>
<tr>
<td>n inf 2009</td>
<td>26</td>
<td>8 (1, 22)</td>
<td>0.99</td>
</tr>
<tr>
<td>n inf 2010</td>
<td>35</td>
<td>15 (3, 40)</td>
<td>0.95</td>
</tr>
<tr>
<td>n inf 2011</td>
<td>38</td>
<td>26 (6, 65)</td>
<td>0.76</td>
</tr>
<tr>
<td>n inf 2012</td>
<td>33</td>
<td>40 (10, 91)</td>
<td>0.35</td>
</tr>
<tr>
<td>mean year</td>
<td>2010</td>
<td>2011 (2010, 2011)</td>
<td>0.01</td>
</tr>
<tr>
<td>mean long</td>
<td>1399</td>
<td>1415 (738, 1730)</td>
<td>0.47</td>
</tr>
<tr>
<td>mean lat</td>
<td>393</td>
<td>447 (264, 742)</td>
<td>0.32</td>
</tr>
<tr>
<td>min long</td>
<td>-290</td>
<td>192 (-1766, 1346)</td>
<td>0.43</td>
</tr>
<tr>
<td>min lat</td>
<td>-258</td>
<td>-265 (-764, 162)</td>
<td>0.51</td>
</tr>
<tr>
<td>max long</td>
<td>2156</td>
<td>2014 (1900, 2156)</td>
<td>0.88</td>
</tr>
<tr>
<td>max lat</td>
<td>1118</td>
<td>1106 (965, 1325)</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Results: DLM coefficients $\beta_t$

**Intercept**

<table>
<thead>
<tr>
<th>Year</th>
<th>$\beta_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>-1.00</td>
</tr>
<tr>
<td>2009</td>
<td>0.00</td>
</tr>
<tr>
<td>2011</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Number of caves**

<table>
<thead>
<tr>
<th>Year</th>
<th>$\beta_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>0.00</td>
</tr>
<tr>
<td>2009</td>
<td>1.50</td>
</tr>
<tr>
<td>2011</td>
<td>3.00</td>
</tr>
</tbody>
</table>

**Freezing days**

<table>
<thead>
<tr>
<th>Year</th>
<th>$\beta_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>0.00</td>
</tr>
<tr>
<td>2009</td>
<td>0.15</td>
</tr>
<tr>
<td>2011</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**Spatial range**

<table>
<thead>
<tr>
<th>Year</th>
<th>$\rho_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>0.00</td>
</tr>
<tr>
<td>2009</td>
<td>0.15</td>
</tr>
<tr>
<td>2011</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Simulation study

- To illustrate the proposed method, we estimate the value of several policies for data similar to the WNS data.

- We generate data from the gravity model with parameters set to the posterior mean.

- We consider both active and preventative treatments, with effect size that roughly halves the infection probabilities.

- Data are generated for 2007-2022 with treatments starting in 2012.

- The reward is the number of uninfected counties in 2022.

- We are allowed 30 treatments of each type each year.
We compare 3 policies:

▶ No treatments.

▶ Nearest neighbor: Treat the counties nearest to an infected county, and infected counties nearest an uninfected county.

▶ The proposed policy search method.

We generate 100 datasets and record the number of uninfected counties for each method.
Proportion of infected counties
Proportion of infected counties for the ebola example
A semi-parametric alternative

▶ The proposed policy search method is entirely model based.

▶ If the spatiotemporal gravity model is wrong then there is no reason to trust value estimates.

▶ We could of course build a richer model, but it would be nice to automate this process.

▶ For this, we are working on a space-time version of Q-learning.

▶ We expect that this will be more robust than policy-search, and thus preferred for large datasets.
Definition of the Q-function

- In Q-learning, the policy $\pi$ for the current state $S$ is to select the action $A$ that maximizes the Q-function $Q_\pi(S, A)$.

- Q-learning targets a discounted reward $\sum_t \phi^t R_t$.

- The Q-function is the expected discounted reward if we start in state $S$, take action $A$, and then follow policy $\pi$.

- **Bellman’s equation**: Let $V(S) = \max_{A \in A} Q_\pi(S, A)$ be the discounted value if we begin in state $S$, then

  $$V(S) = E_\pi [R_{t+1} - \phi V_\pi(S_{t+1})].$$

- This leads to an estimating equation for $Q$. 
To adapt Q-learning to the spatial setting, we restrict $Q$ to the sub-class of functions that are:

1. **Local**: we assume additivity over the $n$ spatial locations

   $$Q_\pi(S, A) = \sum_{i=1}^{n} Q_i(S, A).$$

2. **Linear**: we assume the local $Q_i$ are linear in features $F_{ij}$,

   $$Q_i(S, A) = \sum_{j=1}^{p} F_{ij}(S, A) \gamma_{ij}.$$
Spatially explicit Q-functions

▶ We must now estimate the spatially-varying coefficients $\gamma_{ij}$.

▶ We use Bellman’s equation to motivate an objective function.

▶ To improve stability, we add a penalty (graphical lasso) to smooth the $\gamma_{ij}$ over space.

▶ This requires that we observe many state-action pairs, which we have not for the examples considered so far.

▶ We are also exploring ways to include scientific knowledge into this formulation.
We have proposed a general framework for policy optimization for spatial decision problems. The key features are that we attain an interpretable policy and can handle non-stationarity and high dimensions. Our simulations illustrate the potential of this approach.

Limitations:
- Entirely model-based.
- Doesn’t find the globally optimal policy.
- Not applicable in real time.
- Practical constraints?

Thanks to the US Fish & Wildlife Service for financial support.