Space, time, and Society: Three Spatio-temporal modeling paradigms with applications

Shyam Ranganathan

Virginia Tech

January 8, 2019

Acknowledgements

Xinwei Deng, Statistics

Julia Gohlke, Population Health Sciences

Leigh-Anne Krometis, Biological Systems Engineering

Korine Kolivras, Geography

Linsey Marr, Civil and Environmental Engineering

Scotland Leman, Statistics

James Hawdon, Sociology

Peter Hauck, Computer Science

Graduate Students:

Zhihao Hu, Statistics

Christopher Grubb, Statistics

Lauren Butting, Population Health Sciences

Michael Marston, Geography

Ethan Smith, Biological Systems Engineering

Shane Bookhultz, Statistics

Nathan Wycoff, Statistics

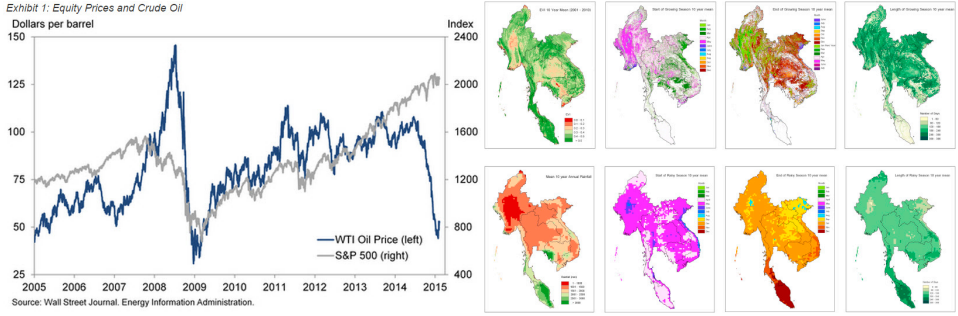
Funding support:

Hume Center for National Security and Technology.

Outline

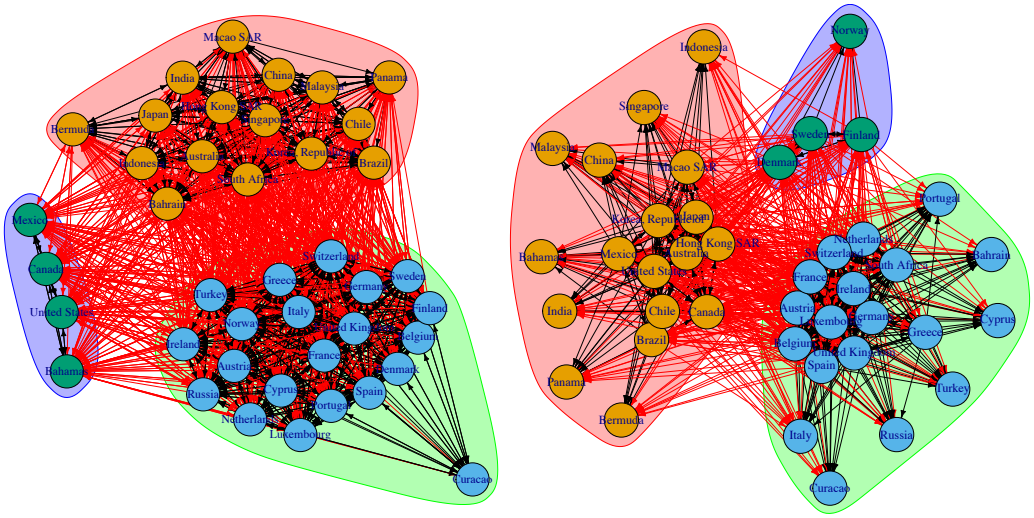
1. Three Spatio-temporal problems and three different approaches to solving them
2. Neighborhood VAR
3. Hierarchical Modeling
4. Spatio-temporal LDA

Multiple dependent time series

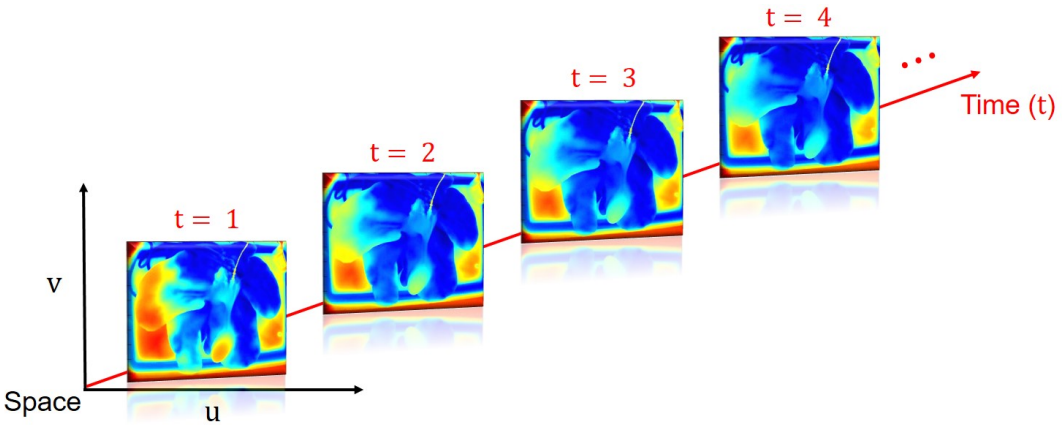


Motivation - High-dimensional Examples

Dependent trade and economic networks:



Motivation - High-dimensional Examples



Reference: Lan, Q., Sun, H., Robertson, J.L., Deng, X., & Jin, R. (2017). Non-invasive Assessment of Liver Quality in Transplantation based on Thermal Imagining Analysis. Manuscript submitted for publication, Grado Department of Industrial & Systems Engineering, Virginia Tech, Virginia.

Three problems

1. Multiple dependent timeseries with some form of "spatial" dependence between the different timeseries
2. Multiple dependent timeseries with "spatial" embedding of the data and a dependence relation between the timeseries
3. Multiple dependent timeseries with "spatial" diffusion processes the object of study

I. Neighborhood Vector Auto Regression

Vector Autoregression (VAR)

A q^{th} order VAR with p dependent time series is given by

$$\mathbf{y}_t = A_1\mathbf{y}_{t-1} + A_2\mathbf{y}_{t-2} + \ldots + A_q\mathbf{y}_{t-q} + e_t, t = 1, 2, \ldots, T$$

where $\mathbf{y}$ is the p-dimensional time series given by $\mathbf{y}_t = [y_{1,t}, y_{2,t}, \ldots, y_{p,t}], t = 1, 2, \ldots, T$

A_i are the $p \times p$ coefficient matrices that capture the relationships between the different time series

e_t is the noise, typically assumed to be zero mean and with no serial correlation. T is the length of the time series.

High-dimensional VAR

Complexity in the estimation of the VAR model is due to lag order q and the dimension p, but the complexity is only linear in q whereas it is quadratic in p. In the high-dimensional case where p is comparable to T, this leads to serious problems in estimation.

Hence VAR is not preferred in many high-dimensional problems.

e.g.: sensor fusion problems, spatio-temporal problems, image processing problems etc., each developing their own solutions!

Sparsity

One way to handle high-dimensional data is to impose some sparsity constraints.

Instead of estimating all p^2 coefficients within a coefficient matrix, we assume that only a small number of these are significant.

Multiple approaches are possible here:

1. Davis et al. (2016) define a "Sparse VAR" algorithm by using partial spectral coherence to estimate which AR coefficients should be non-zero.
2. Basu and Michailidis (2015) use regularized estimation for sparse estimation. Tan et al. (2016) use another shrinkage-based estimator.
3. Guo et al. (2016) use a banded VAR approach similar to the banding assumptions used in estimating covariance matrices (Bickel and Levina (2008) among others). We will expand on this approach

Banded VAR

In Banded VAR, the assumption is that the coefficient matrices are "banded", i.e., the time series show dependence only among adjacent timeseries. The following figure is from Guo et al. (2016) in Biometrika.

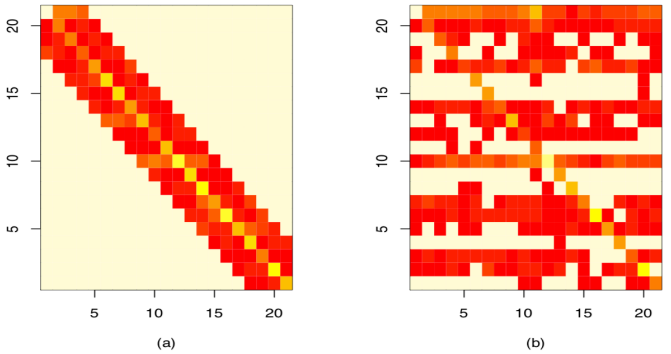


Fig. 3. Example 2: (a) Estimated banded coefficient matrix $\widehat{A}$ for the model based on the ordering using distances to Heilongjiang, and (b) estimated sparse coefficient matrix $\widetilde{A}$ by lasso. The larger the absolute value of a coeffcient is, the darker the colour is.

Banded VAR

The VAR(q) model in this case can be written as

$$\mathbf{y}_t = A_1\mathbf{y}_{t-1} + \ldots + A_q\mathbf{y}_{t-q} + \epsilon_t, t = 1, \ldots T,$$

where $\mathbf{y}$ is p-dimensional as before.

The sparsity condition for banded VAR is specified as $a^m_{i,j} = 0, |i-j| > k_0, m = 1, \ldots q$ so that only the k_0 adjacent timeseries (on either side) affect estimates of the focal timeseries.

k_0 is called the bandwidth and it needs to be estimated from the data. Guo et al. (2016) use the marginal BIC to estimate the bandwidth

Given a particular bandwidth, the coefficient matrices can be estimated using the OLS estimator for each timeseries separately

Neighborhood VAR

We make the spatial dependence notion in banded VAR explicit - a few "neighbors" contain most of the information about the focal timeseries.

We formalize this notion of "neighborhood" using a $p \times p$ distance matrix $\mathbf{D}$, where the element $D_{i,j}$ contains the distance between the timeseries indexed by i and j.

The banded VAR is a special case where the timeseries are always assumed to be arranged in a one-dimensional lattice with the focal timeseries at the origin and all the successive timeseries arranged equally away from it, according to their presence in the ordering $[y_1, \ldots y_p]$

In a 2-D problem, e.g. a problem from image processing, we can define the distance to be based on a Manhattan-type distance, where we obtain a generalization of the banded VAR into a block-banded VAR structure as we consider distances along both dimensions of the matrix

Neighborhood VAR - Definitions

We assume that the timeseries $y_1, \ldots, y_p$ come from sources $s_1, \ldots, s_p$ in some space, say $\Re^m$ with well-defined distances between them $d(i, j)$. We define the $d - neighborhood$ of s_i, as

$\mathcal{N}_i^d = \{j : d(s_i, s_j) \leq d\}$. Note that every timeseries is a neighbor of itself for every value of d and at every time instant t

For every timeseries $y_i, i = 1, \ldots, p$, we define the neighborhood VAR regression by the equation $y_i(t) = \sum_{j \in \mathcal{N}_i^d} A(i, j) y_j(t-1), t = 2, \ldots, T$

Neighborhood VAR - Algorithm

Algorithm Neighborhood VAR Estimation

Input: $[\mathbf{y_1}, \dots \mathbf{y_T}]$, Lag order: q, $\mathbf{D}$ (We assume the distance matrix is given)
Output: Coefficient matrices: $\hat{\mathbf{A}}_\mathbf{1}, \dots \hat{\mathbf{A}}_\mathbf{q}$
for $d \ in \ 1 : d_{max}$ **do**
  for $i \ in \ 1 : p$ **do**
    Find the 'd-neighborhood' $\mathcal{N}_i^d$ of the i^{th} timeseries
    Perform regression for the i^{th} timeseries on $\mathcal{N}_i^d$ and compute coefficients $\beta^{d,i}$
    Compute the marginal BIC as $BIC(d,i) = \log(RSS(d,i)) + \frac{1}{n} d^m C_n \log(p \vee n)$, m - dimension of space
  end for
end for
Find $\hat{d} = \max_{1 \le i \le p}(\operatorname{argmin}_{1 \le d \le d_{max}} BIC^{d,i})$
Optional: Use BIC within the set given by $\mathcal{N}_i^d$ to choose a smaller subset of predictors for each timeseries

Banded VAR - Asymptotic properties

Summary of Conditions:

1. Strict stationarity of the coefficient matrices
2. Identifiability of coefficients
3. Positive definite autocovariance matrix for the process y_t
4. Innovation process ϵ_t is iid with zero mean and covariance Σ_ϵ with finite moments

Theorem

$Pr(\hat{d} = d_0) \to 1, as\ T \to \infty$

Theorem

$\|\hat{A}_j - A_j\|_F = O_P(p/T)^{1/2}, \|\hat{A}_j - A_j\|_2 = O_P(log\ p/T)^{1/2}, as\ T \to \infty$

Theorem

If Σ_ϵ is banded with bandwidth s_0 and has finite L1-norm, for any integers, $r, j \geq 0$, there exists a banded matrix $\Sigma_j^{(r)}$ with bandwidth $2(2r+j)d_0 + s_0 + 1$ such that
$\|\Sigma_j^{(r)} - \Sigma_j\|_2 \leq C_1 \delta^{2(r+j)+1}, \|\Sigma_j^{(r)} - \Sigma_j\|_1 \leq C_2 \delta^{2(r+j)+1}$ *for constants C_1, C_2 independent of r, p and $\delta \in (0,1)$*

Neighborhood VAR - Asymptotic properties

The asymptotic properties of Neighborhood VAR are the same as those for Banded VAR, in that:

1. The correct distance is selected as $T \to \infty$.
2. The norm of the error in coefficients matrix goes down with increasing T.
3. The autocovariance matrix formed using a Neighborhood VAR and the same approximation as in the Banded VAR paper Guo et al. (2016) will converge to the true autocovariance matrix.

Neighborhood VAR - Simulation Results

The data is generated from a model with 1-D spatial decay:
$y_i(t) = \beta_0 y_i(t-1) + \sum_{j \in \mathcal{N}^{d_0}, j \neq i} (\beta_0 \exp(-0.5 d(i,j)) + \epsilon_j(t)) y_j(t-1) + e_t$
The distance is computed as $d(i,j) = |i-j|$, where the i^{th} timeseries is assumed to be located at point i on the 1-D lattice. This is similar to the banded VAR assumption but we make a realistic assumption of decaying contribution of any timeseries as a function of their distance from the focal timeseries.

The bandwidth for the model is fixed at k_0 so that there is no correlation between timeseries at a distance more than k_0 apart.

Both banded VAR and neighborhood VAR were used to estimate the coefficient matrices and the bandwidth.

Neighborhood VAR - Simulation Results

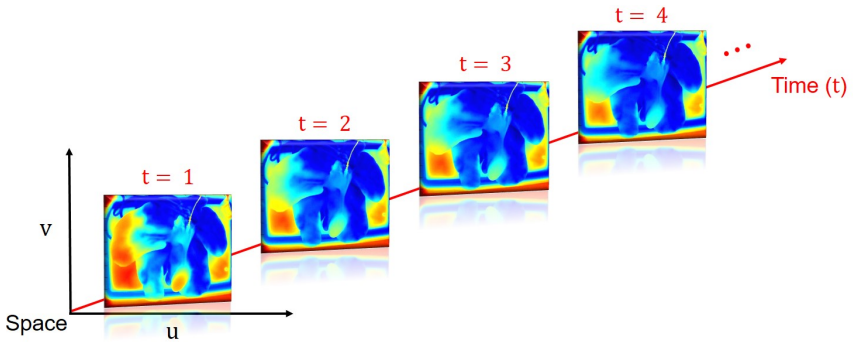
p, k_0	Estimated bandwidth BVAR $\hat{k_{BVAR}}$				Estimated bandwidth NVAR $\hat{k_{NVAR}}$						
	0	1	2	3	0	1	2	3	4	5	6
$p = 100, k_0 = 1$	14	478	8	0	0	230	244	25	1	0	0
$p = 100, k_0 = 2$	74	134	292	0	0	29	318	131	19	2	1
$p = 100, k_0 = 3$	125	155	143	77	0	29	113	308	45	5	0
$p = 100, k_0 = 4$	149	230	99	22	1	70	99	131	186	12	1
$p = 100, k_0 = 5$	196	227	73	4	4	87	120	147	95	45	2
$p = 100, k_0 = 6$	235	215	45	5	4	97	150	129	85	28	7

p, k_0	Mean error norm (BVAR)	SD error norm (BVAR)	Mean error norm (NVAR)	SD error norm (NVAR)
$p = 100, k_0 = 1$	0.28	0.03	0.32	0.05
$p = 100, k_0 = 2$	0.34	0.03	0.36	0.05
$p = 100, k_0 = 3$	0.39	0.05	0.39	0.04
$p = 100, k_0 = 4$	0.43	0.08	0.41	0.05
$p = 100, k_0 = 5$	0.46	0.09	0.42	0.06
$p = 100, k_0 = 6$	0.49	0.11	0.42	0.06

Neighborhood VAR - Applications

Liver imaging data - 145 time instants of Thermal Imaging of the liver. The last several time instants of this image series is stationary so can be modeled using VAR models. We look at each pixel series as a timeseries, with clear dependences across multiple timeseries.

Reference: Lan, Q., Sun, H., Robertson, J.L., Deng, X., & Jin, R. (2017). Non-invasive Assessment of Liver Quality in Transplantation based on Thermal Imagining Analysis. Manuscript submitted for publication, Grado Department of Industrial & Systems Engineering, Virginia Tech, Virginia.



Neighborhood VAR - Applications

We apply the Neighborhood VAR algorithm to the low-resolution ($p = 7,304$ timeseries!) liver imaging data. We estimate the mean square prediction error of the algorithm on a rolling basis.

1. We fix $i : i + 100$ time instants as training data for i ranging from 1:44 so that we have 100 training observations always.
2. We estimate the coefficient matrix based on the training data.
3. We compute the mean square prediction error for the out-of-sample holdout data.

Based on our implementation, we find that the mean square prediction error is in the range of 0.09-0.16. For reference, the actual data is in the range of 5-25. So we get really small prediction errors.

The estimated neighborhood distance is in the range of 0-2 (for different values of i).

Neighborhood VAR - Applications

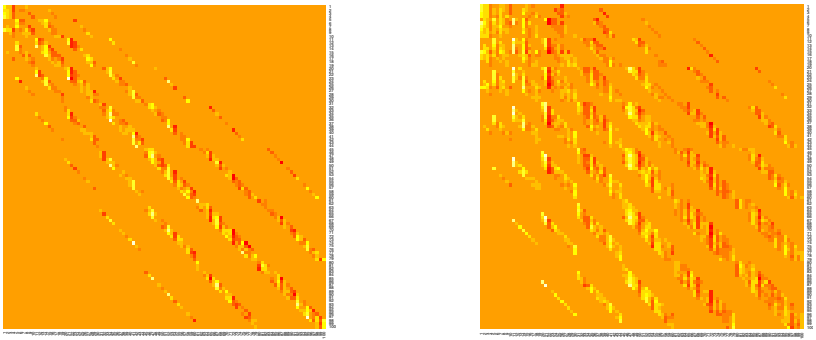


Figure 1: Two illustrative heatmaps showing a visualization of the coefficient matrices for the VAR(1) model of the liver imaging data with only 100 timeseries. The first coefficient matrix suggests a distance of 3, while the second suggests a neighborhood distance of 5 is required. Clearly the Banded VAR cannot capture these patterns.

Future Directions

1. Extend to more general cases - the networks example - where "distance" is not obviously given - latent space approaches.
2. Estimating the distance matrix $\mathbf{D}$ when it is not given, from the data itself needs to be done efficiently, and this will need to be worked into the problem.
3. The formulation lends itself to a Bayesian approach in estimation. We will consider this in future implementations.

References

1. Guo, S., Wang, Y., & Yao, Q. (2016). High-dimensional and banded vector autoregressions. Biometrika, asw046.Chicago
2. Davis, R. A., Zang, P., & Zheng, T. (2016). Sparse vector autoregressive modeling. Journal of Computational and Graphical Statistics, 25(4), 1077-1096.
3. Basu, S., & Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. The Annals of Statistics, 43(4), 1535-1567.
4. Tan, H. R., Ting, C. M., Salleh, S. H., Kamarulafizam, I., & Noor, A. M. (2016, December). Shrinkage estimation of high-dimensional vector autoregressions for effective connectivity in fMRI. In Biomedical Engineering and Sciences (IECBES), 2016 IEEE EMBS Conference on (pp. 121-126). IEEE.
5. Bickel, P. J., & Levina, E. (2008). Regularized estimation of large covariance matrices. The Annals of Statistics, 199-227.

II. Modeling health outcomes using a hierarchical model

Problem Statement

- *Hypothesis:* Exposure to heat during pregnancy leads to adverse birth outcomes
- *Definitions:*
  - *Heat exposure:* A function of the heat/temperature 'felt' by individual – physical data on temperature, but actual 'exposure' moderated by occupation, socio-economic status, infrastructure etc – spatio-temporal modeling for both heat and for exposure
  - *Adverse birth outcomes:* Pre-term births, low birth weight etc. indicative of adverse birth outcomes – modeling can be at individual level, or at an aggregated level – dichotomous, continuous or count data

Prior Work on heat exposure

- Non-accidental mortality increases during heat waves in cities (Anderson and Bell 2011; Peng et al. 2011).
- Previously identified important covariates include SES, age, chronic disease status, minority status, and geography (greater risk in northern 'less hot' cities).
- Basu et al. (2011) reported a positive association between preterm birth and heat waves in California.
- Others have detected positive associations outside of the U.S. (Strand et al. (2012) in Brisbane, Australia, Schifano et al. (2013) in Rome, Italy

Prior Work on exposure effects modeling

Table 1
Characteristics of the included studies on ambient temperature and preterm birth/gestational age.

Study	Location	Study design	Sample	Exposure measurement	Covariates adjusted for	Statistical method and result			Study qualityscore (0–12)
						Statistical method	Statistic	Estimate	
Europe									
Lee et al. (2008)	London, UK	Ecological *	482,568 singleton live births, 1988–2000	Daily max and min temperature at the time of birth	Long-term trend, seasonality, day of the week, public holiday	Time-series logistic regression	Risk change per 1 °C increase	Max temperature: OR = 1.00 (95%CI: 0.99–1.00, p > 0.05) Min temperature: OR = 1.00 (95%CI: 1.00–1.00, p > 0.05)	9
Flouris et al. (2009)	Greece	Ecological	516,874 live births, 1999–2003	Mean temperature during the birth month	No	Correlation analysis	Correlation coefficient between temperature and gestational age	Both sex: $r = -0.210$ ($p < 0.001$) Males: $r = -0.208$ ($p < 0.001$) Females: $r = -0.211$ ($p < 0.001$)	8
Dadvand et al. (2011)	Barcelona, Spain	Retrospective cohort	7585 singleton births spontaneous labour, 2002–2005	Heat–humidity index	Maternal demographic and clinical characteristics, and infant sex	Linear regression model	Gestational age change after high heat index exposure on the day before delivery	5.3-day reduction (95% CI: −10.1 to −0.05, p = 0.03)	12
Wolf and Armstrong (2012)	Two German States	Ecological *	All reported hospital singleton births from Brandenburg (2002–2010) and Saxony (2005–2009)	Daily mean temperature	Long-term trend, seasonality, day of the week	Logistic time-series regression combined with constrained distributed lag model	Temperature effect (ORs) as a linear and a categorical variable	No clear evidence for an association between temperature and PTB was found ($p > 0.05$)	9
Schifano et al. (2013)	Rome, Italy	Ecological *	All singleton live births by natural delivery, 2001–2010	Maximum apparent temperature (MAT) and heat waves in the month preceding delivery	Long-term trend, seasonality, days of holiday, and air pollution	Poisson generalized additive model combined with distributed lag model	Percent change during heat waves and per 1 °C increase in MAT	During heat waves: +19% increase (95% CI: 7.91–31.69) Per 1 °C increase in MAT: 1.9% (95%CI: 0.86–2.87)	10
Vicedo-Cabrera et al. (2014)	Valencia, Spain	Ecological *	20,148 singleton natural births during the warm season (May–September), 2006–2010	MAT and daily minimum temperature	Long-term trend, seasonality, day of the week, public holiday, and relative humidity	Quasi-Poisson generalized additive models combined with distributed lag non-linear models	Percent change in risk relative to median temperature	20% increase when MAT;the 90th percentile two days before delivery 5% increase when minimum temperature ≥90th percentile in the last week	10
Vicedo-Cabrera et al. (2015)	Stockholm, Sweden	Ecological *	All singleton spontaneous births collected from the Swedish Medical Birth Register, 1998–2006 (gestational age ≥22weeks)	Daily mean temperature during the last month of gestation	Long-term trend, seasonality, day of the week, public holiday, and relative humidity	Quasi-Poisson generalized additive models combined with distributed lag non-linear models	Cumulative risk ratio relative to median temperature	Mean temperature = 75th percentile: RR = 2.50 (95% CI: 1.02–6.15)	10
Arroyo et al. (2016b)	Madrid, Spain	Ecological *	298,705 live singleton births, 2001–2009	Daily maximum temperature	Linear trends, seasonality, and the autoregressive nature of the series, day of the week	Autoregressive over-dispersed Poisson regression models	Relative risks(RRs) for interquartile increase in temperature	RR = 1.055 (95%CI: 1.018–1.092)	10
Cox et al. (2016)	Flanders, Belgium	Ecological *	807,835 live-born singleton births with a gestational age	Daily minimum and maximum air temperature	Long-term trend, seasonality, day of the week, public	Quasi-Poisson generalized additive models combined	Percent increase in risk relative to median temperature	Extreme heat (99th vs. 50th percentile): 15.6% (95% CI: 4.8	11

(*continued on next page*)

Issues in prior work

Definitional issues:

- 'Heat waves' are taken to be indicators of exposure. In addition, a variety of heat wave definitions exist.
- Infrastructure/systematic effects not accounted for.

Modeling issues:

- Exposure taken to be the same as temperature at a spatial location.
- Exposure at a particular point in pregnancy alone considered.

General issues:

- Typically, huge amounts of missing data.
- Small effect sizes – need some kind of modeled causal inference to avoid over-fitting.

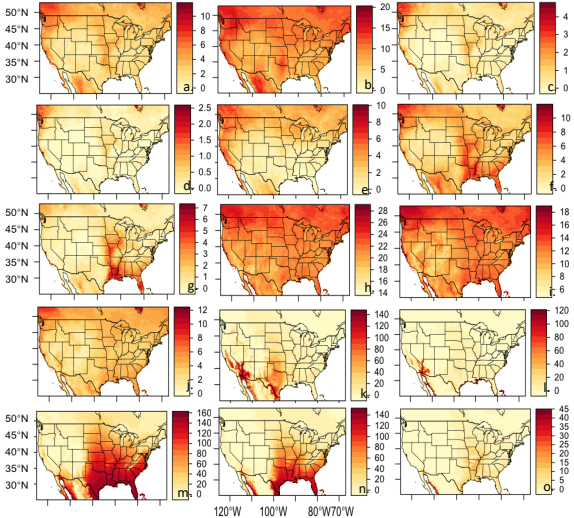
Heat wave definitions

Smith et al. 2013 Climatic Change 118:811-825

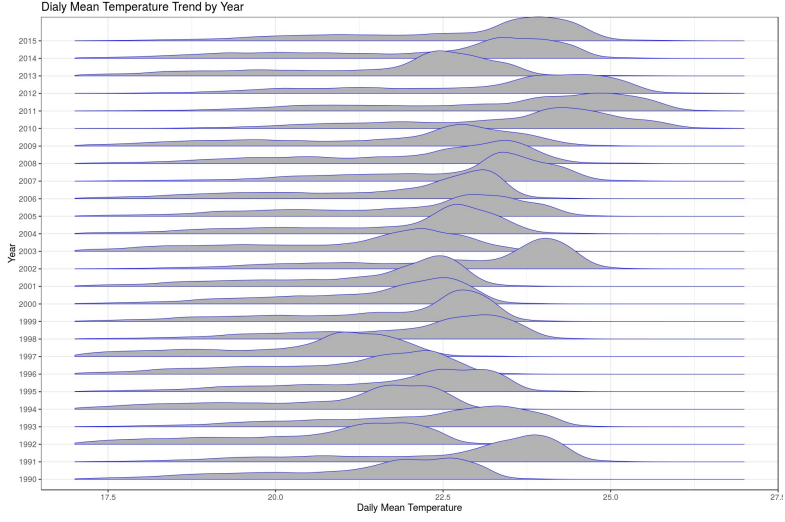
HI	Definition	Reference
HI01	Mean daily temperature > 95th percentile for ≥ 2 consecutive days	Anderson and Bell (2011)
HI02	Mean daily temperature > 90th percentile for ≥ 2 consecutive days	Anderson and Bell (2011)
HI03	Mean daily temperature > 98th percentile for ≥ 2 consecutive days	Anderson and Bell (2011)
HI04	Mean daily temperature > 99th percentile for ≥ 2 consecutive days	Anderson and Bell (2011)
HI05	Minimum daily temperature > 95th percentile for ≥ 2 consecutive days	Anderson and Bell (2011)
HI06	Maximum daily temperature > 95th percentile for ≥ 2 consecutive days	Anderson and Bell (2011)
HI07	Maximum daily temperature ≥ 81st percentile every day, ≥ 97.5th percentile for ≥ 3 nonconsecutive days, and consecutive day average ≥ 97.5th percentile	Peng et al. (2011)
HI08	Maximum daily apparent temperature[b] > 85th percentile for ≥ 1 day	Hattis et al (2012); Steadman (1984)
HI09	Maximum daily apparent temperature[b] > 90th percentile for ≥ 1 day	Hattis et al (2012); Steadman (1984)
HI10	Maximum daily apparent temperature[b] > 95th percentile for ≥ 1 day	Hattis et al (2012); Steadman (1984)
HI11	Maximum daily temperature >35°C (95°F) for ≥ 1 day	Tan et al. (2007)
HI12	Minimum daily temperature > 26.7°C (80.1°F) or maximum daily temperature > 40.6°C (105.1°F) for ≥ 2 consecutive days	(Robinson 2001)
HI13	Maximum daily heat index[c] > 80°F for ≥ 1 day	Rothfusz (1990); Steadman (1979)
HI14	Maximum daily heat index[c] > 90°F for ≥ 1 day	Rothfusz (1990); Steadman (1979)
HI15	Maximum daily heat index[c] > 105°F for ≥ 1 day	Rothfusz (1990); Steadman (1979)
HI16	Maximum daily heat index[c] > 130°F for ≥ 1 day	Rothfusz (1990); Steadman (1979)

Heat wave definitions

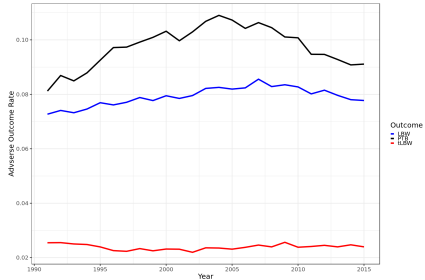
Smith et al. 2013 Climatic Change 118:811-825

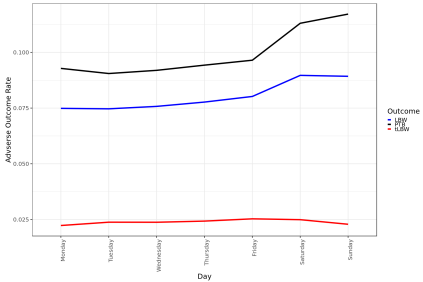


Temporal Variation



Temporal Variation





Data

- Address-level birth records were obtained through a Data Sharing Agreement with Virginia Department of Health and was approved under VDH and VT IRB protocols.
- Addresses from a total of 2,203,198 (86.7%) of the $2,542,519$ original birth records were successfully geocoded (highest geocoding rates in later years).
- Singleton births, $\geq$ 22 weeks gestation are included in the analysis.
- Responses: Preterm Birth ($<$ 37 weeks clinical estimate of gestational age) (N=239,311) Low birth weight ($<$2500 g, but greater than 200 g) (N=200,398) Term low birth weight ($<$2500 g, but $\geq$ 37 weeks gestation)
- Individual-level covariates: Payment method, maternal education, maternal age, birth order, marital status, race, ethnicity

Data

- Heat data obtained from Phase 2 of the North America Land Data Assimilation System (NLDAS-2) on 13.75 kilometre grid (hourly data).
- Supplementary heat data on a 1 kilometer grid with a Moderate Resolution Imaging Spectroradiometer (MODIS) (once in 8 days data).
- Rural-Urban Commuting Area Codes (RUCA), version 2.0 provides a measure of rurality with 3 categories – "urban focused", "large rural city/town (micropolitan) focused", and "small rural and isolated town focused".
- CDC's Social Vulnerability Index uses 15 U.S. census variables at tract level to help local officials identify communities that may need support in preparing for hazards; or recovering from disaster.

Data Structure

Variable	Type	Range	Level	Notes
County	String		County	95 Counties, but there used to be more
DOB Year	Int	[1990, 2015]	Individual	
DOB Season	String		Individual	
DOB Day	String		Individual	
DOB Holiday	Logical		Individual	Indicator of whether DOB was a major holiday
Gestation	Int	[22, 55]	Individual	Estimated gestation length, thrown out if < 22
Plurality	Int	[1,]	Individual	Thrown out if > 1
Birth Order	Int	[1,]	Individual	
Weight	Int	[200, 8500]	Individual	
Mother Age	String		Individual	Aggregated into $<$18, 18-35, $>$35
Mother Race	String		Individual	Black or Other
Mother Ethnicity	String		Individual	Hispanic or Other
Mother Education	String		Individual	
Marital Status	Logical		Individual	
SVI	String		Census Tract	Aggregated into Categories
RUCA	String		Census Tract	Aggregated into Percentile groups, 0-25, ..., 75-100
PTB	Logical		Individual	Gestation $<$ 37 weeks
LBW	Logical		Individual	Weight $<$ 2500g
tLBW	Logical		Individual	Weight $<$ 2500g, Gestation $>$ 36 weeks
HIXX	Logical		Individual	Calculated from closest Lat/Long to address

Modeling solution

A general spatio-temporal modeling solution to address the multiple issues can be formulated in a hierarchical manner. A rough caricature is:

$g(y_i|x, s, z) \sim N(.,.)$ — Observation/Data model
$x|f \sim N(.,.)$; $f \sim GP()$ — Spatial Process model
$s_t|s_{t-1} \sim N(.,.)$ — Temporal model

Into this framework, we can introduce missingness mechanisms, causal inference etc.

But, first we work on separate aspects of this problem by breaking things down into their elements

Spatial Effect – Systematic/Infrastructural effects

Apart from the spatial correlation in the temperature variable, there are systematic effects fue to policy or infrastructure.

We model this using a hierarchical model (equivalently, a varying slopes model with interactions) that captures systematic variations.

Here, the heat exposure effect on pre-term birth is modeled as a random effect, with the regression coefficients themselves predicted by county-level aggregate variables for "Social Vulnerability Index (SVI)", which captures socio-economic characteristics, and "Rural-Urban Commuting Area Codes (RUCA)", which captures the rural/urban divide in infrastructure.

- $\mathbf{y} = \mu + \mathbf{h}\beta_c + \mathbf{C}\gamma + \epsilon$
- $\beta_c = \gamma_0 + \gamma_1 \mathbf{c_{SVI}} + \gamma_2 \mathbf{c_{RUCA}} + \mathbf{J}$

Modeling

M1: $\mathbf{y} = \mu + \boldsymbol{h\beta} + \boldsymbol{C\gamma} + \boldsymbol{\epsilon}$

- fixed effects for heat exposure and covariates

M2: $\mathbf{y} = \mu + \boldsymbol{h\beta} + \boldsymbol{hc_{SVI}\delta_1} + \boldsymbol{hc_{RUCA}\delta_2} + \boldsymbol{C\gamma} + \boldsymbol{\epsilon}$

- additional fixed effects for interaction between heat exposure & RUCA and heat exposure & SVI

M3: $\mathbf{y} = \mu + \boldsymbol{h\beta_r} + \boldsymbol{hc_{SVI}\delta_{1r}} + \boldsymbol{hc_{RUCA}\delta_{2r}} + \boldsymbol{C\gamma} + \boldsymbol{\epsilon}$

- The hierarchical model equivalent for M3 is:
- $\mathbf{y} = \mu + \boldsymbol{h\beta_c} + \boldsymbol{C\gamma} + \boldsymbol{\epsilon}$
- $\beta_c = \gamma_0 + \gamma_1 \mathbf{c_{SVI}} + \gamma_2 \mathbf{c_{RUCA}} + \boldsymbol{\eta}$

Heat Exposure Index Descriptive Tables

Using the full data:

HI	Definition	Reference	Births.During.HI..n.....	PTB..n.....	LBW..n.....	tLBW..n.....
HI01	Mean daily temp > 95th percentile for > 1 consecutive days	Anderson and Bell 2011	193,207 (4.34%)	15,430 (4.38%)	12,435 (4.55%)	4,298 (4.50%)
HI02	Mean daily temp > 90th percentile for > 1 consecutive days	Anderson and Bell 2011	289,025 (6.49%)	22,901 (6.50%)	18,264 (6.69%)	6,299 (6.60%)
HI03	Mean daily temp > 98th percentile for > 1 consecutive days	Anderson and Bell 2011	120,816 (2.71%)	9,672 (2.74%)	7,805 (2.86%)	2,662 (2.79%)
HI04	Mean daily temp > 99th percentile for > 1 consecutive days	Anderson and Bell 2011	86,173 (1.93%)	6,898 (1.96%)	5,605 (2.05%)	1,901 (1.99%)

Limit to 95th percentile mean daily temperature > 26:

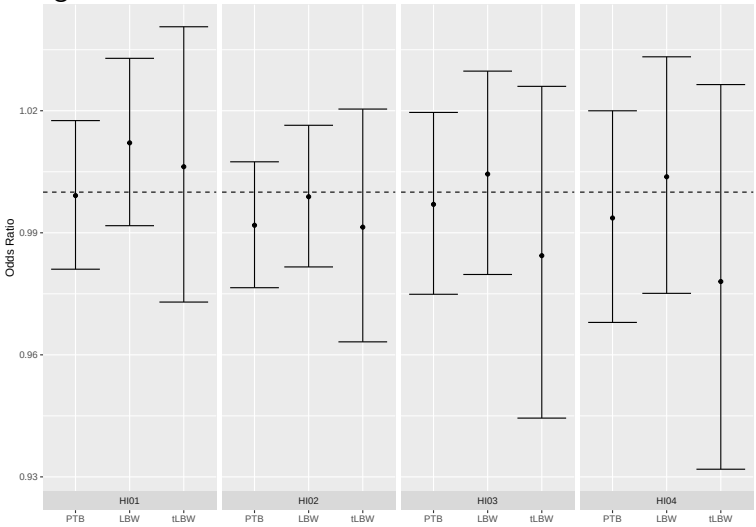
HI	Definition	Reference	Births.During.HI..n.....	PTB..n.....	LBW..n.....	tLBW..n.....
HI01	Mean daily temp > 95th percentile for > 1 consecutive days	Anderson and Bell 2011	102,371 (2.86%)	8,271 (2.91%)	6,794 (3.08%)	2,415 (3.13%)
HI02	Mean daily temp > 90th percentile for > 1 consecutive days	Anderson and Bell 2011	171,453 (4.80%)	13,768 (4.84%)	11,081 (5.02%)	3,874 (5.02%)
HI03	Mean daily temp > 98th percentile for > 1 consecutive days	Anderson and Bell 2011	54,205 (1.52%)	4,427 (1.56%)	3,677 (1.67%)	1,280 (1.66%)
HI04	Mean daily temp > 99th percentile for > 1 consecutive days	Anderson and Bell 2011	33,899 (0.948%)	2,730 (0.959%)	2,313 (1.05%)	814 (1.05%)

Limit to 95th percentile mean daily temperature > 28:

HI	Definition	Reference	Births.During.HI..n.....	PTB..n.....	LBW..n.....	tLBW..n.....
HI01	Mean daily temp > 95th percentile for > 1 consecutive days	Anderson and Bell 2011	4,172 (1.21%)	379 (1.35%)	342 (1.55%)	132 (1.68%)
HI02	Mean daily temp > 90th percentile for > 1 consecutive days	Anderson and Bell 2011	8,562 (2.49%)	740 (2.63%)	630 (2.86%)	234 (2.98%)
HI03	Mean daily temp > 98th percentile for > 1 consecutive days	Anderson and Bell 2011	1,606 (0.467%)	143 (0.509%)	141 (0.639%)	53 (0.675%)
HI04	Mean daily temp > 99th percentile for > 1 consecutive days	Anderson and Bell 2011	798 (0.232%)	73 (0.26%)	67 (0.304%)	28 (0.357%)

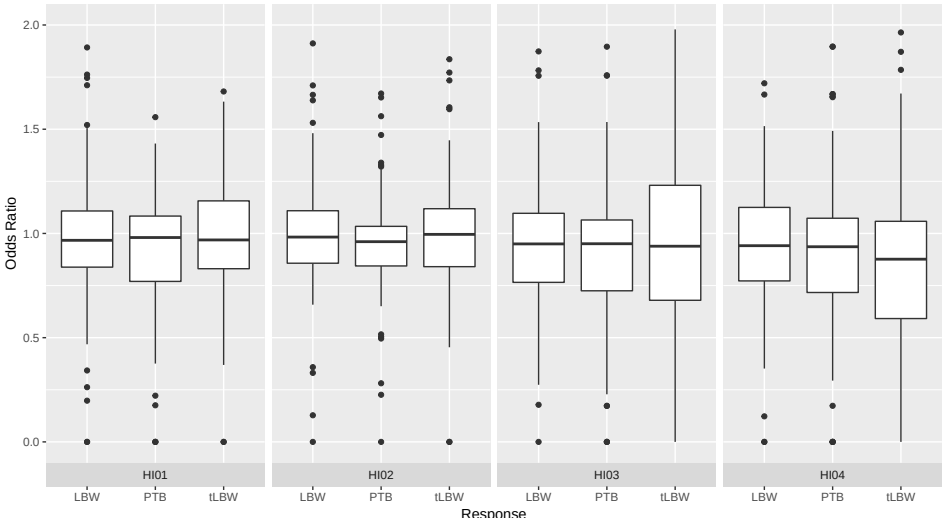
Model Coefficient Graphs

Using the full data:



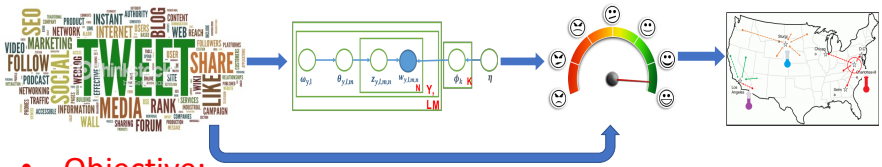
Model Coefficients (M1.1)

Here, we fit models for each county with 1000 or more observations. This shows the range of HI odds ratios for different response and predictor combinations.



III. Spatio-temporal topic flow modeling of twitter data

Topic Flow Modeling to detect polarization



- Objective:

Forecast threats due to polarization in society induced by diffusion of information on online social media

- Impact:

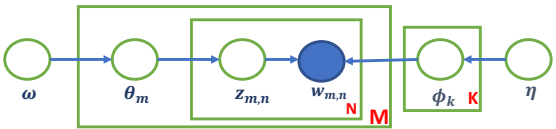
“Information warfare” has a become a real and tangible threat and can be combated only by understanding information diffusion patterns and their outcomes

- Methods:

We postulate a novel Spatio-temporal LDA model for online social media data, create a polarization measure and build an accompanying threat barometer that can help monitor/forecast spatio-temporal units that are most under threat due to polarization

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA)



M – number of documents
N – number of words
K – number of topics
$w_{m,n}$ - nth word in mth document
$z_{m,n}$ - nth topic in mth document
ϕ_k - distribution of words in topic k
θ_m - distribution of topics in document m
ω, η – parameters for word,topic distributions

Hierarchical model:

$$\boldsymbol{\phi_k} \sim \boldsymbol{Dir}(\boldsymbol{\eta}), \boldsymbol{\theta_m} \sim \boldsymbol{Dir}(\boldsymbol{\omega}), \boldsymbol{z_{m,n}} \sim \boldsymbol{Mult}(\boldsymbol{\theta_m}), \boldsymbol{w_{m,n}} \sim \boldsymbol{Mult}\left(\boldsymbol{\phi}_{\boldsymbol{z_{m,n}}}\right)$$

1. For each topic, choose the distribution of words in the dictionary ($\boldsymbol{\phi_k}$)
2. For each document, choose the distribution of topics ($\boldsymbol{\theta_m}$)
3. For each word in the document:
   a. first choose a topic the word comes from ($\boldsymbol{z_{m,n}}$)
   b. then, choose a word from the topic ($\boldsymbol{w_{m,n}}$)

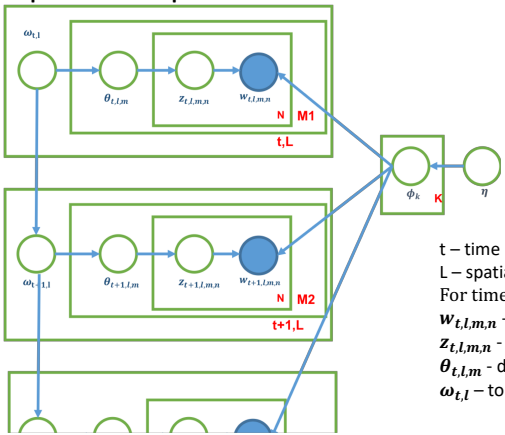
Blei, Ng, Jordan (2003), Pritchard, Stephens, Donnelly (2000)

Many LDAs

1. Dynamic LDA (Blei and Lafferty, 2006)
2. Correlated Topic Model (Blei and Lafferty, 2006, 2007)
3. Hierarchical LDA (Blei et al. 2004, Li and Perona, 2005)
4. Weighted LDA (Tang et al, 2005)
5. Spatial LDA (Wang and Grimson, 2007)
6. Twitter opinion Topic model (Lim and Buntine, 2014)

Spatio-Temporal LDA (ST-LDA)

ST-LDA: Spatio-Temporal LDA



t – time index
L – spatial locations
For time and location t, l:
$\boldsymbol{w_{t,l,m,n}}$ - nth word in mth document
$\boldsymbol{z_{t,l,m,n}}$ - nth topic in mth document
$\boldsymbol{\theta_{t,l,m}}$ - distribution of topics in document m
$\boldsymbol{\omega_{t,l}}$ – topic distribution parameter

ST-LDA

Modeling Topic Flows (Conceptual)

Temporal Diffusion

$\Omega_t = \boldsymbol{B}(\boldsymbol{\theta})\Omega_{t-1} + \epsilon_y$; For time $t = 1, \dots, T$. VAR process for time correlation

$\boldsymbol{B}(\boldsymbol{\theta})_{(l1,l2)} = \theta_2 e^{\theta_1 Dist^2(l1,l2)} + \theta_3$; Spatial locations $(l1, l2)$.

Spatial process: $\boldsymbol{\theta}$=$\{\theta_1, \theta_2, \theta_3\}$ are *scale, length-scale and nugget* parameters

The specific topic proportions for each document in a spatio-temporal unit are:

$$\theta_{t,l,m} \sim N(\omega_{t,l}, \sigma_\omega^2 I)$$

Random effects model to model document variability

Topic Flow

$w_{t,,,}|z_{t,,,}, \phi_{z_{t,,,}} \sim \text{Multinomial}(\phi_{z_{t,,,}})$ Word Frequencies for document *m* at time *t*

$\phi_{z_{t,l,m,n}} \sim \text{Dirichlet}(\eta)$

$z_{t,l,m,n}|\theta_{t,l,m} \sim \text{Multinomial}(\theta_{t,l,m})$ Topic evolution at time *t*

Note: Indexing has been abbreviated or suppressed for brevity.

Latent Dirichlet Allocation Topic Flow Model for evolving spatial-temporal topics.

Scenario 1: Low polarization/Bonding

Scenario 2: Medium polarization

Scenario 3: High polarization

Implementation

1. Implemented a pipeline for tweetbase – 11 million tweets over 2 weeks spread across the US. (this is highly selective already!)
2. Creating elasticsearch database
3. Implementing NLP – stemming, sentiment analysis etc.
4. Implemented a variational EM algorithm on subset of tweets
5. Some natural topics spring out but lots of junk too – related to non-convexity – need to add robustness
6. Inference too slow – need for speed