Climate, Health, and Statistics

Bo Li

December 5, 2019

University of Illinois at Urbana-Champaign

Data Science Week, Department of Mathematical Sciences
Purdue University Fort Wayne, IN

Acknowledgement

Former/current students:

- Luis Barboza, *University of Costa Rica*
- Lyndsay Shand, *Sandia National Lab*
- Sooin Yun, *UIUC*

Collaborators:

- Dolores Albarracín, *UIUC*
- Caspar Ammann, *NCAR*
- Julien Emile-Geay, *University of Southern California*
- Trevor Park, *UIUC*
- Doug Nychka, *School of Colorado Mines*
- Jason Smerdon, *Columbia University*
- Frederi Viens, *Michigan State University*
- Xianyang Zhang, *Texas A&M University*

Partial support from NSF-1602845, NSF-1830312, NIH-R56, NIH R01MH114847

Overview of my research

- Statistics for climate studies
    - Paleoclimate reconstruction
    - Characterization of spatiotemporal pattern of climate fields
- Environmental health
    - HIV diagnosis prediction
    - West Nile Virus infection and environmental variables
- Theory and methodology in spatial statistics
    - Model teleconnection between climate variables
    - Nonparametric models for spatial and spatio-temporal random fields
    - Nonstationary models for spatio-temporal random processes
    - Comparing two spatio-temporal random fields

Why care about the PAST climate?

- Accurate and precise reconstructions of past climate help to characterize *natural climate variability* on longer time scales.
- Spatially wide-spread instrumental temperature observations extend back to only about 1850.
- Validate climate models - Atmosphere/Ocean General Circulation Model (AOGCM)

How to recover past climate?

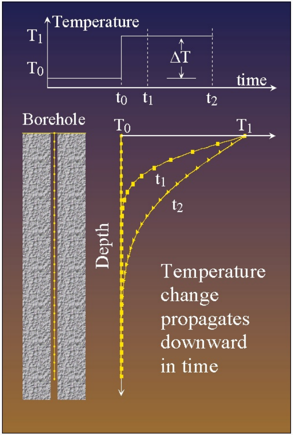
- Earth's climate history written in ice, wood and stone.
- Reconstruct the past temperature from indirect observations (proxies) such as
  - Tree-ring width and densities
  - Pollen
  - Borehole
  - Speleothems (cave deposits)
  - Coral records, etc.
- Radiative Forcings: Solar, Volcanic eruption and Greenhouse gases.

Tree-ring and Pollen

Utescher, 1997

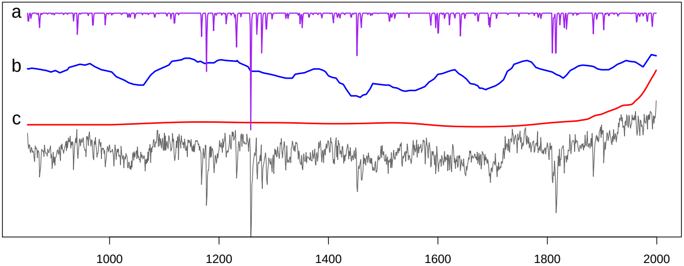
Climate indicators: Tree ring width and density; Pollen assemblage

Data - Borehole



Footprint of temperature revolution: Borehole depth profile

Forcings



a: Volcanism (contains substantial noise)

b: Solar irradiance

c: Green house gases

How to integrate different data sources

Skill of each proxy and forcings

- Tree ring (Dendrochronology): annual to decadal
- Pollen: bi-decadal to semi-centennial
- Borehole: centennial and onward
- Forcings: external drivers

Goal: Reconstruct the 850-1849 temperature by all proxies, forcings and the 1850-1999 temperature

Bayesian Hierarchical Model (BHM) to integrate all proxies, forcings and temperatures and get inference of past temperatures

Bayesian Hierarchical Model (BHM)

Distribution rule:

$$[P, T, \boldsymbol{\theta}] = [P|T, \boldsymbol{\theta}][T|\boldsymbol{\theta}][\boldsymbol{\theta}]$$

Three hierarchies:

- Data Stage: [Proxies|Temperature, Parameters]
  Likelihood of Proxies given temperatures
- Process Stage: [Temperature|Parameters]
  Physical model of temperature process
- Parameter Stage: [Parameters]
  Specify the prior of parameters

BHM

- **D**, **P** and **B**: tree-ring (Dendrochronology), Pollen and Borehole.
- $\mathbf{M}_D$, $\mathbf{M}_P$ and $\mathbf{M}_B$: *transformation matrices* in forward models to relate temperature to proxies.
- $\mathbf{T}_1$: Unknown temperatures requiring reconstruction
- $\mathbf{T}_2$: the observed instrumental temperatures

(i) Data stage:

$$
\begin{aligned}
\mathbf{D}|(\mathbf{T}_1', \mathbf{T}_2')' &= \mu_D + \beta_D \mathbf{M}_D (\mathbf{T}_1', \mathbf{T}_2')' + \boldsymbol{\epsilon}_D, \\
\mathbf{P}|(\mathbf{T}_1', \mathbf{T}_2')' &= \mu_P + \beta_P \mathbf{M}_P (\mathbf{T}_1', \mathbf{T}_2')' + \boldsymbol{\epsilon}_P, \\
\mathbf{B}|(\mathbf{T}_1', \mathbf{T}_2')' &= \mathbf{M}_B \{\mu_B + \beta_B (\mathbf{T}_1', \mathbf{T}_2')' + \boldsymbol{\epsilon}_B\}, \\
\mathbf{V}|\mathbf{V}_0 &= (1 + \epsilon_V)\mathbf{V}_0;
\end{aligned}
$$

with

$$
\begin{aligned}
&\boldsymbol{\epsilon}_D \sim \mathrm{AR}(2)(\sigma_D^2, \phi_{1D}, \phi_{2D}) \quad \boldsymbol{\epsilon}_B \sim \text{iid } N(0, \sigma_B^2) \\
&\boldsymbol{\epsilon}_P \sim \mathrm{AR}(2)(\sigma_P^2, \phi_{1P}, \phi_{2P}) \quad \epsilon_V \sim \text{iid } N(0, 1/64)
\end{aligned}
$$

BHM

- $\mathbf{S}$, $\mathbf{V}_0$, and $\mathbf{C}$: the time series vectors of solar irradiance, volcanism and greenhouse gases
- $\mathbf{V}$: the volcanic series with error.
- $\mathbf{T}_1$: Unknown temperatures requiring reconstruction
- $\mathbf{T}_2$: the observed instrumental temperatures

(ii) Process stage:

$$(\mathbf{T}_1', \mathbf{T}_2')'|(\mathbf{S}, \mathbf{V}_0, \mathbf{C}) = \beta_0 + \beta_1\mathbf{S} + \beta_2\mathbf{V}_0 + \beta_3\mathbf{C} + \boldsymbol{\epsilon}_T,$$

$$\boldsymbol{\epsilon}_T \sim \mathrm{AR}(2)(\sigma_T^2, \phi_{1T}, \phi_{2T})$$

Main results

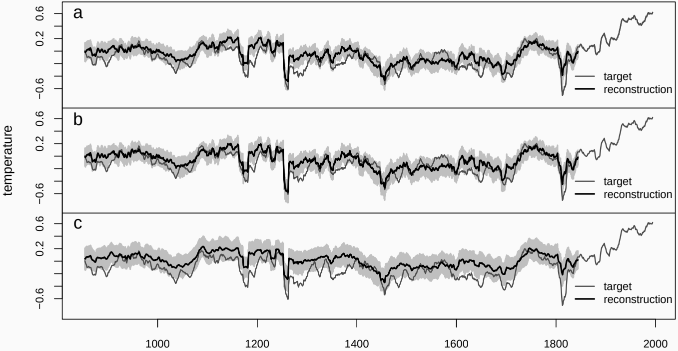


Figure 1: The reconstructions using tree-rings and pollen together with forcings in three scenarios. **a**: modeling $\mathbf{T}$ and without noise; **b**: modeling $\mathbf{T}_1$ and without noise; **c**: modeling $\mathbf{T}$ and with noise.

Main results

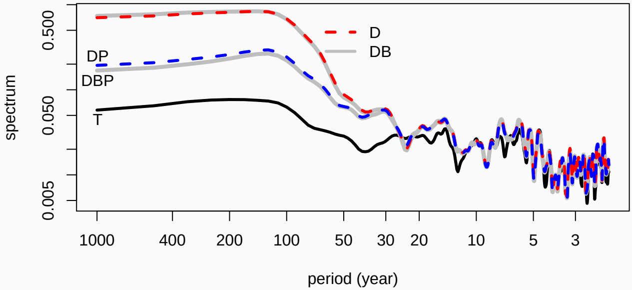


Figure 2: Using smoothed spectrum of reconstruction residuals from the five data models to illustrate the frequency band at which proxies capture the variation of the temperature process (*Li, Nychka and Ammann, 2010)*.

Error structure

- Basic hierarchical Bayesian models:

  Data: $Proxy|Climate = \alpha_0 + \alpha_1 f(Climate) + error$

  Process: $Climate|Forcings = \beta_0 + \beta_1 Forcings + error$

  or $Climate = stochastic\ process$
- Precise uncertainty quantification depends on appropriate modeling of errors.
- Errors are usually assumed to be either short (AR(1) or AR(2)) or no memory (white noise) in the reconstruction.
- Is short or no memory error structure sufficient?
- Is there long-range correlation?
  (*Barboza, Li, Tingly and Viens, 2014*)

Error structure - Long memory

A stochastic process is said to have long-memory if its autocovariance function $\rho(t)$ satisfies:

$$\lim_{t\to\infty}\frac{\rho(t)}{ct^{-M}} = 1$$

for some constant c and $M \in (0, 1)$.

Or through Hurst parameter H,

$$\rho(t) \propto t^{2H-2} \text{ for } H \in (0.5, 1) \text{ for large } t.$$

Error structure - Data

- Temperature Anomalies (Celsius degrees): collected since 1850 over a worldwide grid of climatological stations.
  - HadCRUT3v (HAD): combined land air- and sea-surface temperatures.
  - CRUTEM3v (CRU): land air surface temperatures.
- 1209 biological proxies (*Mann et al., 2008*) collected over different regions and different time horizons.

Assessment of different error structures

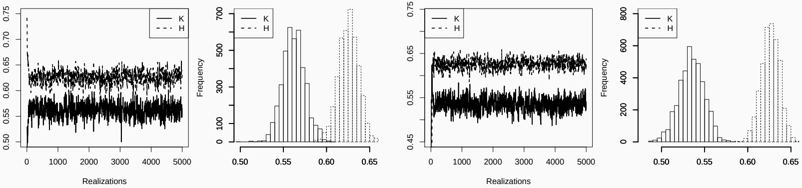
Scenarios	Memory Length		Forcing
	$P\|T$	$T\|F$ or T	
A	long - fGn(H)	long - fGn(K)	✓
B	short - AR(1)	short - AR(1)	✓
C	no memory	no memory	✓
D	long - fGn(H)	long - fGn(K)	X
E	short - AR(1)	short - AR(1)	X
F	no memory	no memory	X
G	short - AR(1)	long - fGn(K)	✓
H	long - fGn(H)	short - AR(1)	✓

Possible long memory (H and K not fixed)

No memory ($H = K = \frac{1}{2}$)

No external forcings: $\beta_i = 0, i = 1, 2, 3$

Hurst parameter estimation



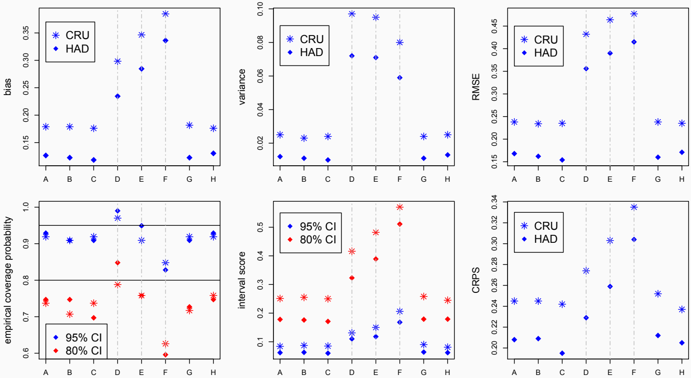
(a) HAD **(b)** HAD **(c)** CRU **(d)** CRU

– Parameter estimates for Scenario A (allow long memory on both models and with forcing in the process model).

– H: Hurst parameter in $P|T$; K: Hurst parameter in $T|F$.

– All significantly larger than 0.5!

Assessment of different error structures



- When forcings are included
  - The prediction is not sensitive to the error structure, but the long memory seems to improve the uncertainty quantification.
- When forcings are not included
  - The long memory model is obviously the best choice.

Current study in Barboza et al. (2019)

- More complete state-of-the-art proxy data (Pages2k data),
- Thorough exploration of data reduction methods
- Integrated Nested Laplace Approximations (INLA)
- Are we living in extraordinary times?

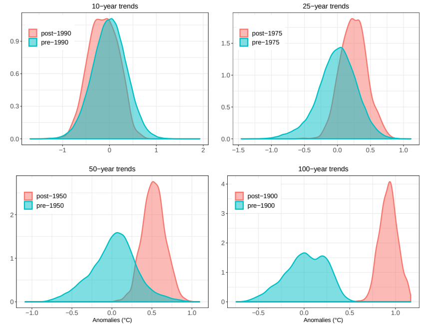


Figure 3: Comparison of the distribution of trends of reconstructed anomalies for different time horizons

Knowledge about HIV

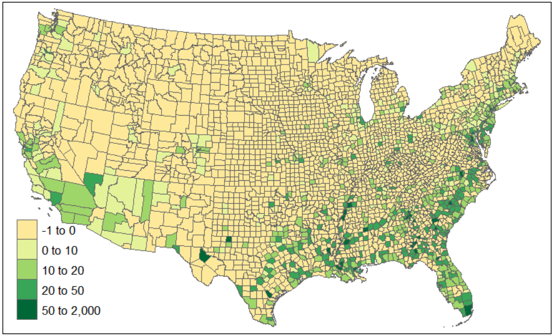
- Human immunodeficiency virus (HIV) can lead to acquired immunodeficiency disease (AIDS)
- Nationally the number of newly diagnosed HIV cases has declined by 19% in the last decade
- Progress has been *uneven* across demographic groups and geographic regions.
  - e.g., Slower declines if any are seen among African Americans and in the south of US

HIV Data

- Annual new HIV diagnosis data from 2008-2014 at county level across the United States (https://aidsvu.org)
- HIV rates are reported as the number of cases per 100,000 people for a given county
- HIV rates are suppressed in any of the following situations:
    - A county has very few cases (< 5) or has a small population size (< 100)
    - The state health department requested not to release its data to AIDSVu due to re-release agreements with the CDC
    - There are no counties in the state such as in Alaska, DC and Puerto Rico.
- Due to the rareness of the disease and the confidentiality constraints, only 25% of all possible county-time observations across the United States have new diagnoses available in the given time frame.

HIV Data

- Highest rates of diagnoses are in the South, West and Northeast.
- Negative rates indicate missing values.



HIV new diagnosis rates in cases per 100,000 during 2012 (*aidsvu.org*).

Motivation

- The National HIV/AIDS Strategy identifies a key goal of intensifying efforts in the communities with the greatest concentration of HIV cases.
- Timely public health insights about the disease will aid in effective public health action, particularly at the local level.
- Regional prediction of disease is central to orchestrating appropriate public health responses.
- Developing models to predict future diagnoses should allow health departments to *intervene before the surge* in new diagnoses occurs.

Formulate the problem

Goal: Make county-level one year ahead prediction of new HIV diagnosis rates using publicly available data

Major sources of variability in the new HIV diagnosis rates:

- Social and economic demographics and other STDs
  (a) General demographic annual summaries by county
  – *United States Census Bureau*
  (b) Social and Economic variable
  – *American Community Survey*
  (c) Prevalence of other STDs such as chlamydia and syphilis
  – *healthindicators.gov*
- Space-time dependency

Particulars of our problem

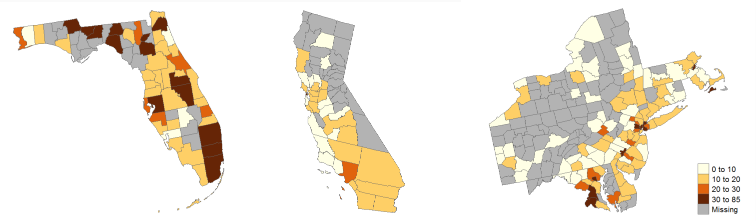
- Rarity of disease leads to few to no incidents in many regions
- Limited reporting time span of only 7 years (2008-2014)
    - A linear or quadratic temporal trend may impose too strong an assumption
    - Choose an autoregressive model with order 1 (AR(1)) to model the temporal correlation
    - Statistically unreliable to fit individual AR(1) model for each county due to so few observations
    - Evolution rates of neighboring counties are likely similar

A promising solution: Borrow strength from neighboring regions in both space and time by taking correlation into account

HIV Data

Focus on three concentrated areas of the US

- Florida – 67 counties, 75% of county-time observations available
- California – 58 counties, 59% of county-time observations available
- New England states: Connecticut, Delaware, Maryland, Massachusetts, New Jersey, New York, and Pennsylvania – 199 counties, 74% county-time observations available collectively.



Data expoloration

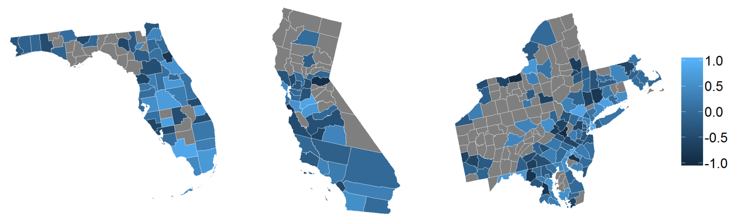


Figure 4: Spatial maps of independent ρ estimates using maximum likelihood estimation for counties in Florida, California and New England.

	Florida		California		New England	
	Statistic	p-value	Statistic	p-value	Statistic	p-value
Moran's I	0.0343	0.3216	0.1072	0.1389	0.2598	0.0003
Geary's C	0.9610	0.3758	0.8098	0.0665	0.7384	0.0005

Table 1: Test Statistics and p-values for Moran's I and Geary's C on testing the null hypothesis of no spatial correlation among ρ_i versas the one-sided alternative hypothesis of positive spatial correlation.

Model development

- Spatially varying autoregressive (SVAR) models (Nobre et al. 2011) seems a good choice.
- Modeling the county specific ρ as a spatially dependent random process using the conditional autoregressive (CAR) model developed in Leroux et al. (1999),
  - Allows for flexible county-specific autoregressive coefficients
  - Makes the estimation of an AR(1) model for each county reliable by borrowing strength from neighbors
  - Reduces the rank of model, easing the potential for overfitting
- Compare them to (1) six competing models in the generalized mixed effects modeling framework (Knorr-Held, 2000) and (2) two spatially invariant autoregressive models.

A Bayesian hierarchical SVAR model

Let $Y_{i,t}$ denote HIV rate for county i at year t and $Z_{i,t} = \log(Y_{i,t})$.

The variance of $Y_{i,t}$ depends on its corresponding population size, i.e., $n_{i,t}$ of county i at time t, so does $Z_{i,t}$.

Level I.

$$Z_{i,t} = \eta_{i,t}(\boldsymbol{\beta}, \boldsymbol{\theta}) + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim N\left(0, \sigma^2 q_{it}\right)$$

- $\eta_{i,t}$: a spatio-temporal random process
- $q_{it} = \frac{c}{n_{i,t} Y_{i,t}}$ for $c = 100,000$.

A Bayesian hierarchical SVAR model

Level II. (Basic form of SVAR model)

$$\eta_{i,t}(\boldsymbol{\beta}, \boldsymbol{\theta}) = X_{i,t-1}^{T}\boldsymbol{\beta} + \psi_{i,t}\rho_i(Z_{i,t-1} - X_{i,t-2}^{T}\boldsymbol{\beta}) \tag{1}$$

- $X_{i,t-k}^{T}\boldsymbol{\beta}$: the linear effects of the previous year's covariates
- $\rho_i \in (-1, 1),\ i = 1, \ldots n$: spatially varying AR(1) coefficients
- $\psi_{i,t} = \sqrt{q_{it}/q_{i(t-1)}}$: ensure that ρ_i measures the correlation between two random components, because the variance of $Z_{i,t}$ is proportional to q_{it}

A Bayesian hierarchical SVAR model

We may add additional random effects

Variations of Level II model

$$\eta_{i,t}(\boldsymbol{\beta},\boldsymbol{\theta}) = X_{i,t-1}^{T}\boldsymbol{\beta} + \psi_{i,t}\rho_i(Z_{i,t-1} - X_{i,t-2}^{T}\boldsymbol{\beta}) + \phi_i \tag{2}$$

$$\eta_{i,t}(\boldsymbol{\beta},\boldsymbol{\theta}) = X_{i,t-1}^{T}\boldsymbol{\beta} + \psi_{i,t}\rho_i(Z_{i,t-1} - X_{i,t-2}^{T}\boldsymbol{\beta}) + \phi_i + \delta_{i,t}$$

- $\boldsymbol{\phi} \sim N(0, \Sigma_\phi)$: Σ_ϕ follows Leroux et al.(1999):

$$\Sigma_\phi = \tau_\phi^2[(1-\lambda_\phi)I + \lambda_\phi R]^{-1}$$

- $\delta_{i,t}$: spatio-temporal interaction random effect

Prior for ρ via a copula approach

A typical Gaussian prior is not applicable to model the spatial dependency in $\boldsymbol{\rho}$ due to its truncated range.

⇒ We propose a copula approach, enabling the modeling of the marginal distribution of ρ_i separately from its spatial dependency structure.

Use a Gaussian copula and assume Uniform(-1,1) marginal distribution

Joint density function of $\boldsymbol{\rho}$:

$$\pi(\rho_1, \cdots, \rho_n) = \frac{\phi_\Omega\left(\Phi^{-1}(\frac{\rho_1+1}{2}), \ldots, \Phi^{-1}(\frac{\rho_n+1}{2})\right)}{\prod_{i=1}^{n} \frac{1}{2}\phi\left(\Phi^{-1}(\frac{\rho_i+1}{2})\right)}$$

- $\boldsymbol{\Phi}_\Omega(\cdot)$: joint CDF of a MV Gaussian distribution with zero mean and Leroux et al. (1999) covariance structure.
- $\Phi^{-1}(\cdot)$: inverse CDF of a standard normal random variable.

Models for Comparison

Below lists eight different forms of $\eta_{i,t}$ (1-6 are similar to those in Knorr-Held, 1999) that will be compared to our SVAR model:

1. $X_{i,t-1}^T\boldsymbol{\beta}$,
2. $X_{i,t-1}^T\boldsymbol{\beta} + \phi_i$,
3. $X_{i,t-1}^T\boldsymbol{\beta} + \alpha_t$,
4. $X_{i,t-1}^T\boldsymbol{\beta} + \delta_{i,t}$,
5. $X_{i,t-1}^T\boldsymbol{\beta} + \alpha_t + \phi_i$,
6. $X_{i,t-1}^T\boldsymbol{\beta} + \alpha_t + \phi_i + \delta_{i,t}$,

7.* $X_{i,t-1}^T\boldsymbol{\beta} + \psi_{i,t}\rho(Z_{i,t-1} - X_{i,t-2}^T\boldsymbol{\beta})$,

8.* $X_{i,t-1}^T\boldsymbol{\beta} + \psi_{i,t}\rho(Z_{i,t-1} - X_{i,t-2}^T\boldsymbol{\beta}) + \phi_i$,

*typical autoregressive models where ρ is assumed fixed for all counties.

Prediction and Model assessment

Prediction: Obtain $\widehat{Z}_{i,t+1}$ through forward sampling, then take exponential and use the posterior median as the prediction.

– Hold 2014 data as testing data and make prediction for that year.

– Model assessment was made based on the following measures:

  - Mean squared prediction error (MSPE)
  - Predictive model choice criterion (PMCC) (Gelfand and Gosh, 1998; Gneiting and Raftery, 2007)
  - Continuous rank probability score (CRPS) (Gneiting and Raftery, 2007)
  - Empirical coverage probability (ECP) at a 95% nominal level
  - Observation of the previous time point as the baseline prediction

Model comparison

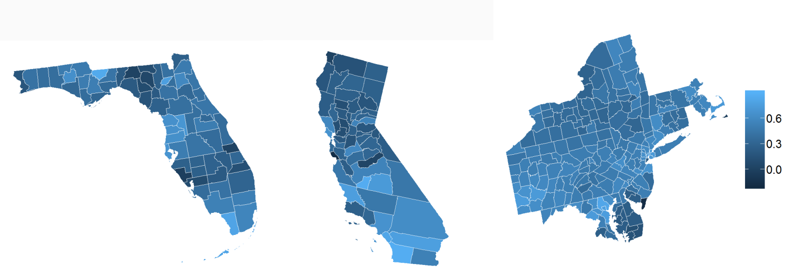
	Florida				California				New England			
Model	MSPE	PMCC	CRPS	ECP	MSPE	PMCC	CRPS	ECP	MSPE	PMCC	CRPS	ECP
(1)	75.30	**221.8**	**3.905**	0.9574	**7.231**	**94.84**	**1.456**	0.9697	**45.88**	**393.7**	**2.557**	0.9196
(2)	**56.89**	223.0	3.906	0.9574	9.715	98.11	1.621	0.9697	66.43	399.3	2.949	0.9107
1	78.26	242.9	4.387	0.9787	10.42	113.9	1.843	0.9697	67.47	442.7	3.041	0.9821
2	58.04	242.0	4.152	1.0000	8.397	138.5	2.092	1.0000	61.41	416.3	2.751	0.9643
3	80.29	245.1	4.367	0.9362	10.28	106.4	1.748	0.9394	65.11	442.2	2.983	0.9018
4	74.91	275.2	4.650	1.0000	9.457	125.6	1.919	0.9697	64.09	455.9	2.958	0.9821
5	60.62	231.7	3.990	0.9787	7.951	116.2	1.801	1.0000	55.10	385.3	2.605	0.9018
6	67.73	240.5	4.223	1.0000	8.031	107.2	1.685	0.9697	63.31	411.1	2.722	0.929
7	76.86	230.8	4.106	0.9362	9.297	107.8	1.704	0.9697	58.65	436.5	2.838	0.9821
8	65.67	226.8	3.922	0.9574	8.289	107.9	1.646	0.9697	57.78	398.4	2.717	0.9286
Y^*_{t-1}	61.51	–	–	–	10.03	–	–	–	62.58	–	–	–

Bold are the smallest numbers of their respective columns.

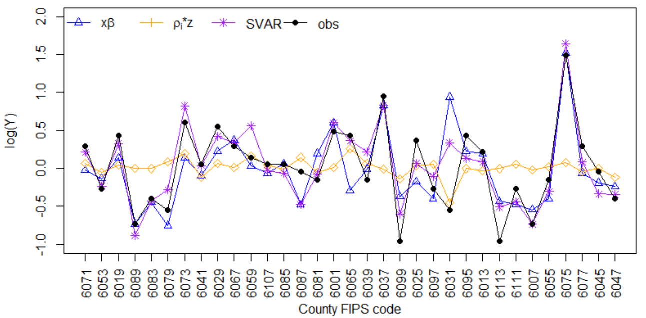
- Model (1): Basic SVAR model
  - Appears to provide the best prediction for California and New England
- Model (2): Basic SVAR model + spatial random effects
  - Provides the best prediction for Florida
  - FL has the most insignificant spatial correlation in Moran's I test, therefore modeling spatial correlation only through ρ_i is insufficient

Model comparison

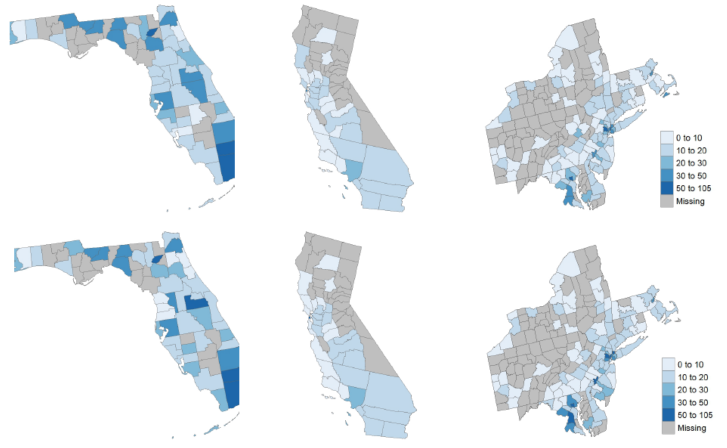
- In model 8, the estimates for ρ are around 0.23, 0.18, and 0.17 for FL, CA, and New England, respectively.
- The estimates for ρ_i in Models (1) and (2) are also significantly different from zero at many counties



Spatial maps of posterior means of ρ_i from Model (1) for FL, CA and NE.



Break down of the contribution from each term in Model (1) to the prediction of 2014 new HIV diagnosis rates in California, where $\rho_i^* Z$ indicates the contribution from $\rho_i(Z_{i,t-1} - X_{i,t-2}\beta)$ and SVAR indicates the overall model prediction.



Observed (top) and predicted (bottom) new HIV diagnosis rates in 2014 for Florida, California and New England using Model 2, 1, and 1, respectively.

Background

The study of climate field reconstruction requires to compare two space-time random fields (Li and Smerdon, 2012; Fremdt et al., 2013; Horváth et al., 2013; Li et al., 2016)

- Mean
- Covariance
- Mean + Covariance
- Trend
- ......

Background

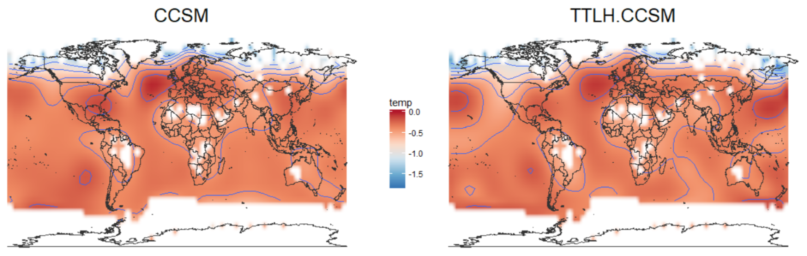


Figure 5: Mean comparison between CCSM and its TTLH reconstruction.

Background

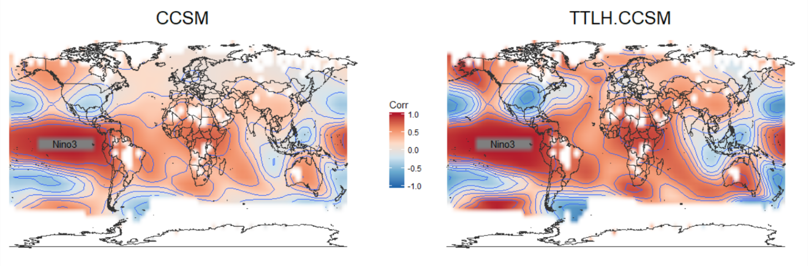


Figure 6: ENSO teleconnection comparison between CCSM and its TTLH reconstruction.

Background

- The previous methods offer only an overview of the difference between two spatiotemporal random fields as a whole
- Detection of local discrepancies between two random fields can be more informative
  - For example, if the two mean functions are different, where are the differences located?
- We propose to compare the characteristics of two spatiotemporal random fields at each location and then adjust the multiplicity due to multiple comparison.

Multiple Testing

- Let $X(\boldsymbol{s};t)$ and $Y(\boldsymbol{s};t)$ be two spatiotemporal random fields observed over spatial locations, $\boldsymbol{s} \in D$, and time t.
- Assume $X(\boldsymbol{s};t)$ and $Y(\boldsymbol{s};t)$ are stationary in time.
- Denote $\theta_X(\boldsymbol{s})$, $\theta_Y(\boldsymbol{s})$ to be certain characteristics of the distribution of $X(\boldsymbol{s};t)$ and $Y(\boldsymbol{s};t)$.
- We are interested in testing

$$H_{0,\boldsymbol{s}} : \theta_X(\boldsymbol{s}) = \theta_Y(\boldsymbol{s}) \text{ vs. } H_{0,\boldsymbol{s}} : \theta_X(\boldsymbol{s}) \neq \theta_Y(\boldsymbol{s})$$

simultaneously for all $\boldsymbol{s} \in D$.

Multiple Testing

- The family-wise error rate (FWER): the probability of making at least one false rejection, e.g., Bonferroni method.
- The power of a FWER controlling procedure is greatly reduced as the number of tests increases.
- False discovery rate (FDR): the expected proportion of false rejections among all rejections (Benjamini and Hochberg, 1995), is preferred in such case
- An FDR procedure is
  - Valid: if it controls the FDR at the nominal level
  - Optimal: if it has the smallest false negative rate (FNR)

Our FDR control – Mirror procedure

- Assume a two-component mixture model for location wise p-values
- Derive optimal rejection region under the mixture model
- Derive cutoff value based on mirror property of f_0
- Develop EM-algorithm to estimate the mixture model with either a nonparametric or a semiparametric alternative density function

Model for $p_{\boldsymbol{s}}$

Let $p_{\boldsymbol{s}}$ be the p-value for testing $H_{0,\boldsymbol{s}}$ against $H_{a,\boldsymbol{s}}$.

We assume that $p_{\boldsymbol{s}}$ follow the two-component mixture model:

$$f(p_{\boldsymbol{s}}; \boldsymbol{s}) = \pi(\boldsymbol{s}) f_0(p_{\boldsymbol{s}}) + (1 - \pi(\boldsymbol{s})) f_1(p_{\boldsymbol{s}}; \boldsymbol{s}),$$

where

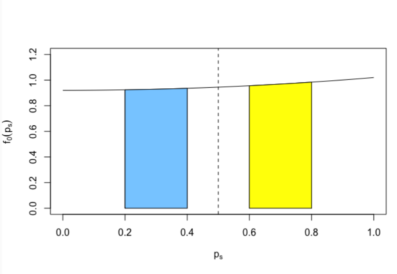
- $\pi(\boldsymbol{s}) \in [0, 1]$ is the probability that the p-value is from the null
- f_0 and f_1 are the null and alternative distributions, respectively

Model for $p_{\boldsymbol{s}}$

$$f(p_{\boldsymbol{s}};\boldsymbol{s}) = \pi(\boldsymbol{s}) f_0(p_{\boldsymbol{s}}) + (1-\pi(\boldsymbol{s})) f_1(p_{\boldsymbol{s}};\boldsymbol{s}),$$

- We assume that f_0 is mirror conservative (Lei and Fithian, 2018)

$$\int_{a_1}^{a_2} f_0(p)dp \leq \int_{1-a_2}^{1-a_1} f_0(p)dp, \quad 0 \leq a_1 \leq a_2 \leq 0.5.$$



- Assume $f_1(p_{\boldsymbol{s}})$ is monotonically decreasing
- $\pi(\boldsymbol{s})$ and $f_1(p_{\boldsymbol{s}})$ are unknown and need to be estimated

Our FDR control – Mirror procedure

- Derive optimal rejection region under the mixture model

$$\mathrm{LFDR}_{\boldsymbol{s}}(p_{\boldsymbol{s}}) = \frac{\pi(\boldsymbol{s}) f_0(p_{\boldsymbol{s}})}{\pi(\boldsymbol{s}) f_0(p_{\boldsymbol{s}}) + (1 - \pi(\boldsymbol{s})) f_1(p_{\boldsymbol{s}}; \boldsymbol{s})} \leq t.$$

- Derive cutoff value based on mirror property of f_0

$$\mathrm{FDP}(t) \leq \frac{\sum_{\boldsymbol{s} \in D} \mathbf{1}\left\{\mathrm{LFDR}_{\boldsymbol{s}}(1 - p_{\boldsymbol{s}}) \leq t\right\}}{1 \vee \sum_{\boldsymbol{s} \in D} \mathbf{1}\left\{\mathrm{LFDR}_{\boldsymbol{s}}(p_{\boldsymbol{s}}) \leq t\right\}} := \mathrm{FDP}_{\mathrm{up}}(t),$$

  Set $t^* = \max\{t \in [0,1] : \mathrm{FDP}_{\mathrm{up}}(t) \leq \alpha\}$. We then reject $H_{0,\boldsymbol{s}}$ if

$$\mathrm{LFDR}_{\boldsymbol{s}}(p_{\boldsymbol{s}}) \leq t^*.$$

- Develop EM-algorithm to estimate the mixture model with either a nonparametric or a semiparametric density function

Application with climate data

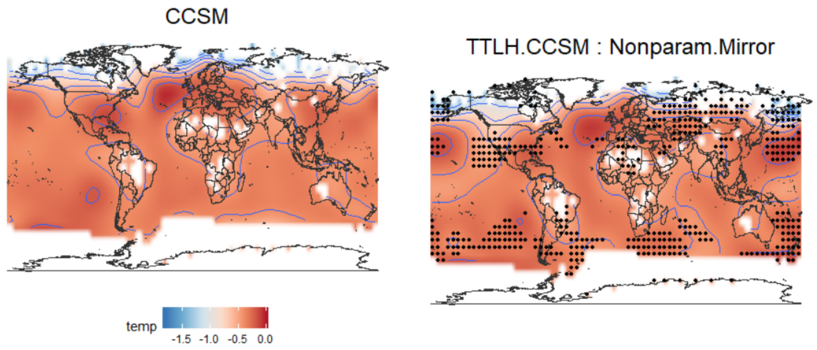


Figure 7: Mean comparison between CCSM and its reconstruction using TTLH. The left plot shows the spatial mean of CCSM climate field and the right plot shows the spatial mean of TTLH reconstruction together with the testing results of nonparametric mirror. The black dots indicate where the mean is different.

Application with climate data

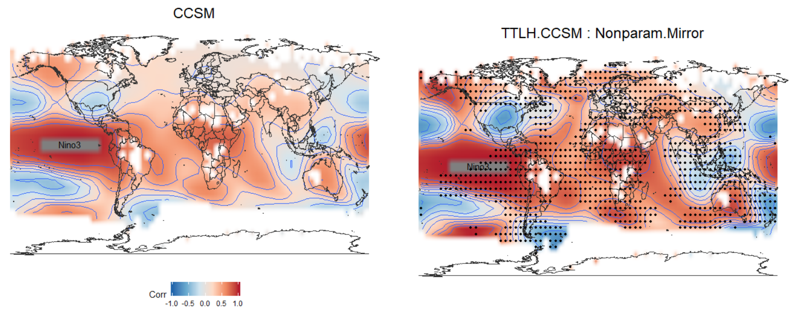


Figure 8: ENSO teleconnection comparison between CCSM and its TTLH reconstruction. The plots show the correlation between the averaged temperature over Nino3 region and local temperatures at each other location with CCSM (left) and TTLH (right). The black dots indicate where the teleconnection strength is different.

Summary and Discussion

Briefly introduce some of my work in climate, public health and methodology for spatio-temporal data

- Studies in paleoclimate reconstruction using BHM
- Prediction of new HIV diagnosis rate at county level
- Comparison between two spatio-temporal random fields