Spatial change-of-support and misalignment problems

Peter F. Craigmile

http://www.stat.osu.edu/~pfc/

Pan-American Advanced Study Institute on Spatio-Temporal Statistics

Búzios, RJ, Brazil

Thursday, June 26 2014

Spatial statistics

- Remember, a **spatial statistical analysis** involves studying and modeling the dependence of data, collected in space.
  - Depending on the form of the spatial process, the spatial data are typically **indexed** by locations/points or regions.
  - The locations themselves may be random.
- We imagine the spatial data to be a realisation of a **stochastic process**, a family of random variables,

$$\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\},$$

  indexed by $\boldsymbol{s} \in D$, defined on a probability space ('$\boldsymbol{s}$' could be a region).
- Clearly the **choice of** D determines the choice of spatial scale.

The choice of spatial scale – some questions

1. Which **spatial scale** is correct?
2. What about if there is **spatial misalignment** – we collected the data on one scale, but need to make inferences on a different scale.
3. How do we **change** from one spatial scale to another?
4. What if we have different spatial datasets that come to us on different spatial scales? How do we **combine** data sources?

We need to be careful not to be **misled** in our inferences.

Some useful references

- Gotway and Young [2002]
- Gelfand et al. [2001] and Banerjee et al. [2004, Chapter 6]

The modifiable areal unit problem (MAUP)

[The term is due to Openshaw and Taylor, 1979, but the problem is much older]

- Some spatial processes are interpretable only on **areal** scales.
- Famous example: crop yield, which is defined in terms of the amount of crop grown (e.g., weight) in a specific area.
  - The choice of the area is user-defined or **modifiable**.
  - An open question: how do we select the areal unit?

Two problems with MAUP

1. Scale effect or aggregation effect: we obtain potentially different inferences as we aggregate to larger regions.

2. Grouping effect or the zoning effect: inferences can change depend on how we choose to aggregate (for the same size of area).

These issues have some relationship to problems found in ecological inference [e.g., Cressie, 1996].

Ecological inference

- Robinson [1950] pointed out that often inference about individuals are made based on group-level data – an **ecological inference**.
- Using data from the 1930 US Census he related the illiteracy rate and the proportion of the population born outside the US.
  - At the state (regional) level: Corr $= -0.53$
  - At individual level: Corr $= 0.12$.
- The contradiction is due to a strong spatial effect.

  (Immigrants tended to settle in states where the native population was more literate.)

The ecological fallacy

- We obtain an **ecological fallacy** when analyses based on group data lead to different conclusions than we would obtain from the individual level data (also called an **ecological bias**).
- Two effects [Morgenstern, 1982]:
  1. Aggregation bias: the effect of the spatial aggregation itself.
  2. Specification bias: the distribution of confounding variables is different under aggregation.
- See, e.g., Wakefield and Lyons [2010].

Example: Pediatric lymphoma in Brazil

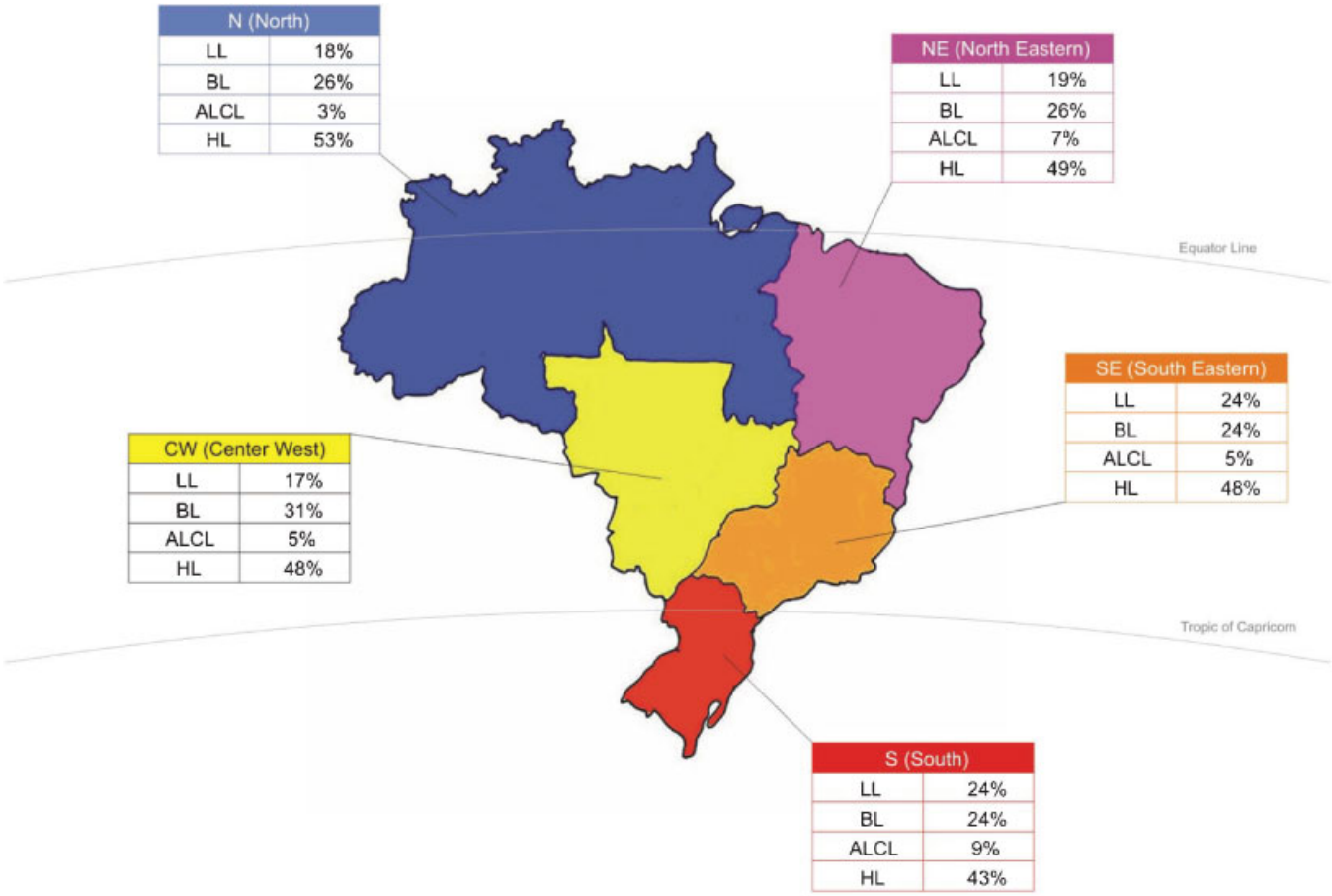

Figure 8 - Overall distribution of the most frequent pediatric lymphomas according to main geographic regions in Brazil.
Footnote: LL: lymphoblastic lymphoma; BL: Burkitt lymphoma; ALCL: anaplastic large cell lymphoma; HL: Hodgkin lymphoma.

[Gualco et al., 2010]

The union of twenty-seven federal units in Brazil

(http://en.wikipedia.org/wiki/States_of_Brazil)

A solution

- The solution is to **try** to build our statistical model for the process of interest on the finest spatial scale possible.
  - We can then aggregate to produce inferences over coarser scales.
- Problems:
  - The spatial data for the process of interest may not always be available at the finest scale.
  - Important covariates may collected at different scales.

The change of support problem (COSP)

- The **spatial support** is the volume, shape, size, and orientation associated with each spatial measurement.
- A change of support is an example of data transformation.
- Thus, the change of support problem (COSP) involves studying the statistical properties of a spatial process as we change the spatial support.

Example changes of support

[Adapted from Gotway and Young, 2002]

Observed at	Inference at	Examples
Point	Point	Kriging
Point	Line	Contouring (upscaling)
Point	Area	Block kriging (upscaling)
Point	Surface	Environmental monitoring
Area	Point	Ecological inference (downscaling)
Area	Area	MAUP, misaligned regions (up- or downscaling)

An example

- Consider a Gaussian geostatistical process

$$Z = \{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$$

  defined on the domain $D \subset \mathbb{R}^2$ (easily generalizes to other domains).

- Suppose that the process has mean function $E(Z(\boldsymbol{s})) = \mu(\boldsymbol{s})$ and stationary covariance function $\text{cov}(Z(\boldsymbol{s}), Z(\boldsymbol{s}')) = C(\boldsymbol{s} - \boldsymbol{s}')$.

- Now let B denote a block defined in D, and let $Z(B)$ denote the **block average** of the process Z in B:

$$Z(B) \;=\; |B|^{-1} \int_B Z(\boldsymbol{s}) d\boldsymbol{s},$$

  where $\int_B$ is shorthand for $\int_{\{\boldsymbol{s} \in B\}}$ and $|B| = \int_B 1 d\boldsymbol{s}$ is the area of B.

Properties of the block average

- The process $Z(B)$ is a random integral.
- We have that the block mean is

$$\mu(B) \;=\; E(Z(B)) \;=\; |B|^{-1} \int_B \mu(\boldsymbol{s}) d\boldsymbol{s}.$$

- For two blocks B and B'

$$\begin{aligned} C(B, B') \;=\; \mathrm{cov}(Z(B), Z(B')) \;&=\; |B|^{-1}|B'|^{-1} \int_B \int_{B'} \mathrm{cov}(Z(\boldsymbol{s}), Z(\boldsymbol{s}')) d\boldsymbol{s} d\boldsymbol{s}' \\ &=\; |B|^{-1}|B'|^{-1} \int_B \int_{B'} C(\boldsymbol{s} - \boldsymbol{s}') d\boldsymbol{s} d\boldsymbol{s}'. \end{aligned}$$

- We also have

$$\begin{aligned} C(\boldsymbol{s}, B) \;=\; \mathrm{cov}(Z(\boldsymbol{s}), Z(B)) \;&=\; |B|^{-1} \int_B \mathrm{cov}(Z(\boldsymbol{s}), Z(\boldsymbol{s}')) d\boldsymbol{s}' \\ &=\; |B|^{-1} \int_B C(\boldsymbol{s} - \boldsymbol{s}') d\boldsymbol{s}'. \end{aligned}$$

Exercises

1. If the mean function is constant over a block what is the block mean?

2. Suppose the mean function depends on a spatially-varying covariate $\{x(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ through the linear function $\mu(\boldsymbol{s}) = \beta_0 + \beta_1 x(\boldsymbol{s})$.

   What happens to the mean function over blocks?

   (This should tell you something about a grouping effect).

3. Let $A_1, \ldots, A_m$ denote a partition of D; i.e., $\cup_{k=1}^{m} A_k = D$ and $A_i \cap A_j = \emptyset$ for any i and j. Suppose that for any points $\boldsymbol{s}_i \in A_i$ and $\boldsymbol{s}_j \in A_j$ $(i \neq j)$ that $\text{cov}(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)) = 0$.

   Describe the statistical properties of $\{Z(A_i) : i = 1, \ldots, m\}$.

Approximating the integrals

- In practice the integrals may not be available in a closed-form.
- Gelfand et al. [2001] shows that we can approximate the integrals using points from B.
- For example letting $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{L_B}$ denote L_B points sampled uniformly from B,

$$\mu(B) \approx L_B^{-1} \sum_{k=1}^{L_B} \mu(\boldsymbol{s}_k).$$

- The sum converges in probability to the LHS as long as the spatial process Z is mean squared continuous [e.g. Stein, 1999];

  i.e., $\lim_{h \to 0} E[Z(\boldsymbol{s} + \boldsymbol{h}) - Z(\boldsymbol{s})]^2 = 0$.

  If Z is stationary, only need the covariance function to be continuous at $\mathbf{0}$.

Standard kriging

- Standard kriging is an example of a COSP.
  - Based on $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))^T$ we predict the geostatistical process Z at a new location $\boldsymbol{s}^*$.
- The best linear predictor of $Z(\boldsymbol{s}^*)$ given $\boldsymbol{Z}$ is

$$E(Z(\boldsymbol{s}^*)|\boldsymbol{Z}) \;=\; \mu(\boldsymbol{s}^*) + \boldsymbol{c}_s(\boldsymbol{s}^*)^T \boldsymbol{\Sigma}_s^{-1}(\boldsymbol{Z} - \boldsymbol{\mu}_s),$$

  where $\boldsymbol{\mu}_s$ is the mean vector of $\boldsymbol{Z}$, $\boldsymbol{\Sigma}_s$ is the $n \times n$ covariance matrix for $\boldsymbol{Z}$, and $\boldsymbol{c}_s(\boldsymbol{s})$ is a n-vector with ith element $\mathrm{cov}(Z(\boldsymbol{s}^*), Z(\boldsymbol{s}_i))$.

- The kriging variance is

$$\mathrm{var}(Z(\boldsymbol{s}^*)|\boldsymbol{Z}) \;=\; C(\boldsymbol{0}) - \boldsymbol{c}_s(\boldsymbol{s}^*)^T \boldsymbol{\Sigma}_s^{-1} \boldsymbol{c}_s(\boldsymbol{s}^*).$$

Predicting blocks

- Now based on $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))^T$ we predict $Z(B)$ at block B.
- The best linear predictor of $Z(B)$ given $\boldsymbol{Z}$ is

$$E(Z(B)|\boldsymbol{Z}) \;=\; \mu(B) + \boldsymbol{c}_{s,b}(B)^T \boldsymbol{\Sigma}_s^{-1}(\boldsymbol{Z} - \boldsymbol{\mu}_Z),$$

  where $\boldsymbol{\mu}_Z$ is the mean vector of $\boldsymbol{Z}$, $\boldsymbol{\Sigma}_s$ is the $n \times n$ covariance matrix for $\boldsymbol{Z}$, $\mu(B)$ is the mean of $Z(B)$, and $\boldsymbol{C}_{s,b}(B)$ is a n-vector with ith element $\mathrm{cov}(Z(\boldsymbol{s}_i), Z(B))$.
- The kriging variance is

$$\mathrm{var}(Z(B)|\boldsymbol{Z}) \;=\; \mathrm{var}[Z(B)] - \boldsymbol{c}_{s,b}(B)^T \boldsymbol{\Sigma}_s^{-1} \boldsymbol{c}_{s,b}(B).$$

Exercise: Block kriging

(This is the solution to the MAUP.)

- Now suppose we observe $Z(B_1), \ldots Z(B_n)$ and wish to predict $Z(B^*)$ at a new block B^* different from B_i.

  1. Derive the kriging mean and variance for this predictor.
  2. Show that if $\mu(\boldsymbol{s}) = \mu_i$ for each $\boldsymbol{s} \in B_i$ then

$$\mu(B^*) \approx \frac{\sum_{i=1}^{n} |B_i \cap B^*| \, Y(B_i)}{|B^*|},$$

     the areally-weighted average.

- The solution to downsampling (predicting $Z(\boldsymbol{s}^*)$, say, based on block averages) follows in a similar fashion.

Practical remarks

- In practice we need to estimate the mean and covariance function of the Gaussian process Z, and plug the estimates into the relevant kriging equation.

  (As discussed previously we may also need to approximate the integrals.)

- Common estimation approaches include [see, e.g., Banerjee et al., 2004, Cressie and Wikle, 2011]:

  1. Naive approaches (e.g., weighted least squares).
  2. Maximum (reduced) likelihood.
  3. Bayesian methods.

Spatial misalignment and the use of hierarchical modeling

- Often variables come to us at different spatial scales.
- This makes the statistical inference problem harder.
- Some advice:
  - Which **scale** do you want to carry out inference on?
  - It helps to write down **hierarchical models** to relate the different variables – if possible, try to relate them on a **common** spatial scale.
  - But, remember the problem of the **ecological fallacy**. It helps to build models on the finest scale possible.

Pediatric lymphoma revisited

- Suppose that we wish to predict the rate of pediatric Hodgkins lymphoma (HL) in the federal units level in Brazil, based on the following information.

Region	N	NE	CW	S	SE
HL	57	161	43	71	214
Total	114	373	105	209	500

- What is a naive estimate of the rate of HL in each federal unit?

The map (again)

(http://en.wikipedia.org/wiki/States_of_Brazil)

Building a hierarchical model

[See, e.g., Flowerdew and Green, 1989, 1993, 1994, Mugglin and Carlin, 1998]

Consider the following simplified example from Mugglin and Carlin [1998]:

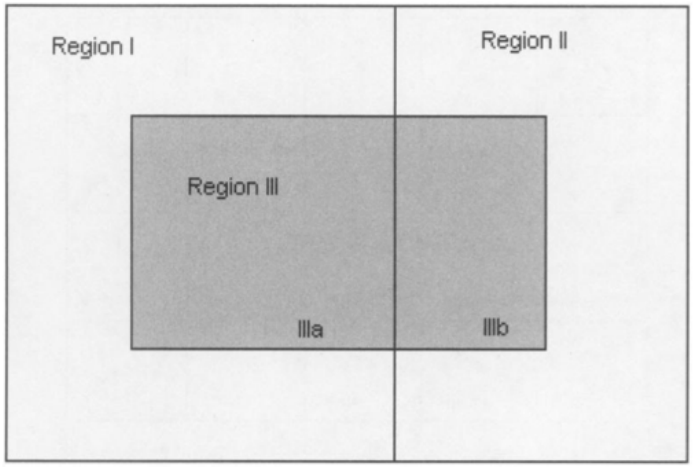

Figure 1. Regional Map For Motivating Example.

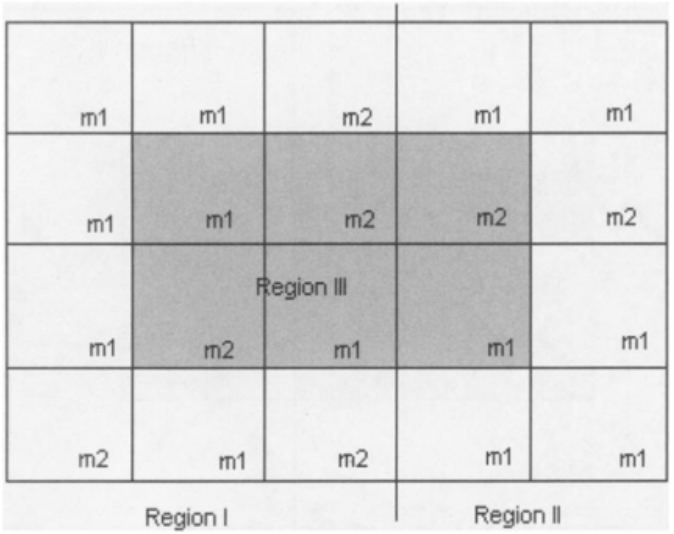

Figure 2. Subregional Map for Motivating Example.

In the right hand figure, $m1$ and $m2$ are two different means.

Multiscale models

- There is considerable interest in building statistical models over different spatial scales.
- For early climatology examples see Berliner et al. [1999], Wikle et al. [2001], Nychka et al. [2002].
- See Calder et al. [2009] for a multiscale method for predicting pollution in soils. See Craigmile et al. [2009] for a simpler model for modeling metal concentrations in public water systems in Arizona, USA.
- For tree models see, e.g., Basseville et al. [1992], Chou et al. [1994], Huang and Cressie [2000] – Johannesson et al. [2007] has a space-time extension.

Spatio-temporal extensions

- Spatio-temporal extensions follow naturally.
- See Wei [2005] for formal aggregation results for time series processes.
- For methods for spatio-temporal processes see Gelfand et al. [2001].
- Read Dominici et al. [2010] for the challenges of modeling multiple pollutants collected at different spatial resolutions. All see Choi et al. [2009].
- In context of modeling temperature, Peter Guttorp will discuss Craigmile and Guttorp [2011] this afternoon.

References

S. Banerjee, B. P. Carlin, and A. Gelfand. *Hierarchical modeling and analysis for spatial data.* Chapman and Hall, Boca Raton, FL., 2004.

M. Basseville, A. Benveniste, K. C. Chou, S. A. Golden, R. Nikoukhah, and A. S. Willsky. Modeling and estimation of multiresolution stochastic processes. *IEEE Transactions on Information Theory*, 38:766–784, 1992.

L. M. Berliner, C. K. Wikle, and R. F. Milliff. Multiresolution Wavelet Analyses in Hierarchical Bayesian Turbulence Models. In *Bayesian Inference in Wavelet-based Models*, pages 341–359. Springer, New York, NY, 1999.

C. Calder, P. Craigmile, and J. Zhang. Regional spatial modeling of topsoil geochemistry. *Biometrics*, 65:206–215, 2009.

J. Choi, M. Fuentes, and B. J. Reich. Spatial-temporal association between fine particulate matter and daily mortality. *Computational Statistics and Data Analysis*, 53:2989–3000, 2009.

K. C. Chou, A. S. Willsky, and R. Nikoukhah. Multiscale systems, kalman filters, and riccati equations. *IIEEE Transactions on Automatic Control*, 39:479–492, 1994.

P. F. Craigmile and P. Guttorp. Space-time modelling of trends in temperature series. *Journal of Time Series Analysis*, 32:378–395, 2011.

P. F. Craigmile, C. A. Calder, H. Li, R. Paul, N. Cressie, et al. Hierarchical model building, fitting, and checking: a behind-the-scenes look at a bayesian analysis of arsenic exposure pathways. *Bayesian Analysis*, 4:1–35, 2009.

N. Cressie and C. K. Wikle. *Statistics for spatio-temporal data.* John Wiley & Sons, New York, NY, 2011.

N. A. Cressie. Change of support and the modifiable areal unit problem. *Geographical Systems*, 3:159–180, 1996.

F. Dominici, R. D. Peng, C. D. Barr, and M. L. Bell. Protecting human health from air pollution. *Epidemiology*, 21:187–194, 2010.

R. Flowerdew and M. Green. Statistical methods for inference between incompatible zonal systems. *Accuracy of spatial databases*, pages 239–247, 1989.

R. Flowerdew and M. Green. Developments in areal interpolation methods and gis. In *Geographic Information Systems, Spatial Modelling and Policy Evaluation*, pages 73–84. Springer, 1993.

R. Flowerdew and M. Green. Areal interpolation and types of data. *Spatial analysis and GIS*, page 145, 1994.

A. E. Gelfand, L. Zhu, and B. P. Carlin. On the change of support problem for spatio-temporal data. *Biostatistics*, 2:31–45, 2001.

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

G. Gualco, C. E. Klumb, G. N. Barber, L. M. Weiss, and C. E. Bacchi. Pediatric lymphomas in Brazil. *Clinics*, 65:1267–1277, 2010.

H.-C. Huang and N. Cressie. Multiscale Graphical Modeling in Space: Applications to Command and Control. In *Proceedings of the Spatial Statistics Workshop*. Springer-Verlag, Moore, New York, 2000.

G. Johannesson, N. Cressie, and H.-C. Huang. Dynamic multi-resolution spatial models. *Environmental and Ecological Statistics*, 14: 5–25, 2007.

H. Morgenstern. Uses of ecologic analysis in epidemiologic research. *American Journal of Public Health*, 72:1336–1344, 1982.

A. S. Mugglin and B. P. Carlin. Hierarchical modeling in geographic information systems: Population interpolation over incompatible zones. *Journal of Agricultural, Biological, and Environmental Statistics*, 3:111–130, 1998.

D. Nychka, C. K. Wikle, and J. A. Royle. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2:315–331, 2002.

S. Openshaw and P. J. Taylor. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Statistical applications in the spatial sciences*, 21:127–144, 1979.

W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351, 1950.

M. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York, NY, 1999.

J. Wakefield and H. Lyons. Spatial aggregation and the ecological fallacy. In *Handbook of Spatial Statistics*. CRC press, 2010.

W. Wei. *Time series analysis (2nd edition)*. Pearson, 2005.

C. K. Wikle, R. Milliff, D. Nychka, and L. M. Berliner. Spatiotemporal Hierarchical Bayesian Modeling: Tropical Ocean Surface Winds. *Journal Of The American Statistical Association*, 96, 2001.