DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706
Phone: (608) 262-2598
Fax: (608) 262-0032

TECHNICAL REPORT NO. 1138

March 19, 2007

Visualizing Abnormal Climate Changes in Central America from 1995 to 2000

Sang-Hoon Cho [1]
Department of Statistics, University of Wisconsin, Madison, WI

Hyonho Chun [2]
Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI

Key words: Data Imputation, Smoothing Spline ANOVA, El Nino and La Nina, Ozone Depletion Areas, Cloud Effect on Temperature

[1]Supported in part by NIH Grant NCI CA075142.
[2]Supported in part by NSF Grant DMS0505636 and DMS0604572 and ONR Grant N00014-06-1-0095.

Visualizing Abnormal Climate Changes in Central America from 1995 to 2000

Sang-Hoon Cho[1] and Hyonho Chun[2]

[1] Department of Statistics, University of Wisconsin - Madison, 1300 University Ave., Madison, WI 53706, U.S.A. cho@stat.wisc.edu
[2] Department of Statistics, University of Wisconsin - Madison, 1300 University Ave., Maidson, WI 53706, U.S.A. chun@stat.wisc.edu

Summary. This article elaborates on the statistical analysis that led to our main findings visually summarized in the poster[3] at the Data Exposition 2006[4] (This poster won the first prize). The statistical methods and procedures to impute missing data and to uncover the natural phenomena, El Niño and La Niña, the ozone depletion areas and the cloud effect on temperature are discussed.

Key words: Data Imputation, Smoothing Spline ANOVA, El Nino and La Nina, Ozone Depletion Areas, Cloud Effect on Temperature

1 Introduction

Data Exposition 2006 was a contest sponsored by the sections on Statistical Graphics, Statistical Computing, and Statistics and the Environment at the Joint Statistical Meeting of 2006 in Seattle, Washington. The most attractive and challenging part of the Data Exposition 2006 was a great deal of freedom to explore a number of potential aspects in the data. In pursuing the objective of the Data Exposition, we decided to focus on one interesting aspect of the data at a time, considering appropriate statistical methods for each different interest. In this article, we elaborate on the statistical analysis that led to our main findings visually illustrated in the poster for the Data Exposition.

This paper is organized as following. In Section 2, the imputation procedure for missing data will be discussed. In Section 3 to 5, the statistical

[3] The poster is available at the following websites:
http://www.amstat-online.org/sections/graphics/dataexpo/2006entries.php,
http://www.stat.wisc.edu/∼cho, or
http://www.stat.wisc.edu/∼chun.
[4] See the following website for more information on the Data Exposition 2006:
http://www.amstat-online.org/sections/graphics/dataexpo/2006.php.

approaches used to uncover our major results including El Niño and La Niña, ozone depletion areas and cloud effects on temperature, will be explained. In Section 6, a brief conclusion will be followed.

2 Data Missing and Imputation

In this section, we discuss the imputation procedure to handle the missing observations in the data. The initial data exploration identified several missing values in **cloud low** where their specific locations and time periods are summarized in Figure 1. It was necessary to impute the missing observations for further statistical analysis.

2.1 Statistical method and procedure

A nonparametric spatial model was considered to impute the missing values. The reasons for our choice of the specific model are two folds. First, there was relatively a small proportion of missing ($0.002\ \% \sim 0.003\%$) in the study area at a given time whereas there was a great proportion of missing ($21\% \sim 44\%$) at a given location over time. Second, a visual exploration found that a parametric model would not be flexible enough to capture the spatial pattern of **cloud low**.

The nonparametric spatial model used for imputation is specified as below. For a given time,

$$\begin{aligned}\text{Cloud Low}_{ij} = c + f_1(\text{lati}_i) + f_2(\text{long}_j) + f_{1,2}(\text{lati}_i, \text{long}_j) + f_3(\text{elevation}_{ij}) \\ + \text{Error}_{ij}, \quad i, j = 1, \ldots, 24, \end{aligned} \tag{1}$$

where i, j denote a row and column number on a 24 by 24 grid, respectively, c an overall mean, f_1, f_2, f_3 nonparametric functions for main effects and $f_{1,2}$ for an interaction, respectively, and $\text{Error}_{ij} \sim N(0, \sigma^2)$. The model (1) is called a "smoothing spline ANOVA". As implied from its name, this model has an analogy to a classical analysis of variance (ANOVA) decomposition. The main difference is that its components, $f_1, f_2, f_3, f_{1,2}$, are nonparametric functions and these functions are estimated by smoothing spline. The R routines for the smoothing spline ANOVA model are implemented in the package **gss**. See (Wahba, 1990) and (Gu, 2000) for general discussions on smoothing spline, smoothing spline ANOVA and the **gss** package.

There are a couple of lemmas in (Wahba and Luo, 1995) and (Luo, 1996) which provide theoretical justifications for our imputation approach using the smoothing spline ANOVA: Leaving-Out-K lemma and Imputation lemma. In our analysis, Imputation lemma was preferred whereas both lemmas should lead to an identical result. It was because neither of lemmas is implemented in R but Imputation lemma could be followed by using the **gss** package with a slight modification. See the appendix for details on this computational issue.

Here are steps we followed to impute missing data. For a given month, if there is any missing location,

1. Fit a smoothing spline ANOVA excluding missing values.
2. Predict the missing value using the fitted model.
3. Fit a smoothing spline ANOVA again including the imputed values.
4. Update the imputed values using the fitted model.
5. Repeat 3 and 4 step until the imputed values converge.

This procedure is contained in (Luo, 1996). As seen in the last step above, Imputation lemma requires an iteration. The predicted values from the initial fitted model serve as a good initial guess for the missing data.

2.2 Results

To validate the imputation result, we selected nearby locations showing temporal trends similar to those at the imputed locations during the observed time period. As shown in Figure 1, the imputed values are close to the values at the selected nearby locations during the missing time period. In the mean time of the imputation, as a byproduct, we found an unusual feature of the data that all variables except for **elevation** have identical values in Location 2 and Location 3 over the entire time period.

3 El Niño and La Niña

The main purpose of this section is to describe our approach to confirm the well-known natural phenomena, El Niño and La Niña, which create severe abnormal weather worldwide. El Niño and La Niña refer to warmer and cooler than normal sea surface temperatures (SST) in the Pacific Ocean, respectively. Since the study area of the Data Exposition covers the western Pacific Ocean, we wondered whether we could identify these climate events from the given data set.

3.1 Statistical methods and procedure

As our general approach, the mean sea surface temperature over the entire time period from 1995 to 2000 was first obtained at each grid point in the study area. By using a dynamic plot, we narrowed down to three interesting areas indicating the most and least fluctuation in the temporal trends of sea surface temperature. The mean sea surface temperatures were compared to the temporal patterns of sea surface temperature in the selected areas.

The mean sea surface temperatures, which served as normal temperature levels in our comparison, were obtained as following. The observed values were first adjusted for seasonality by subtracting the centered monthly means. The

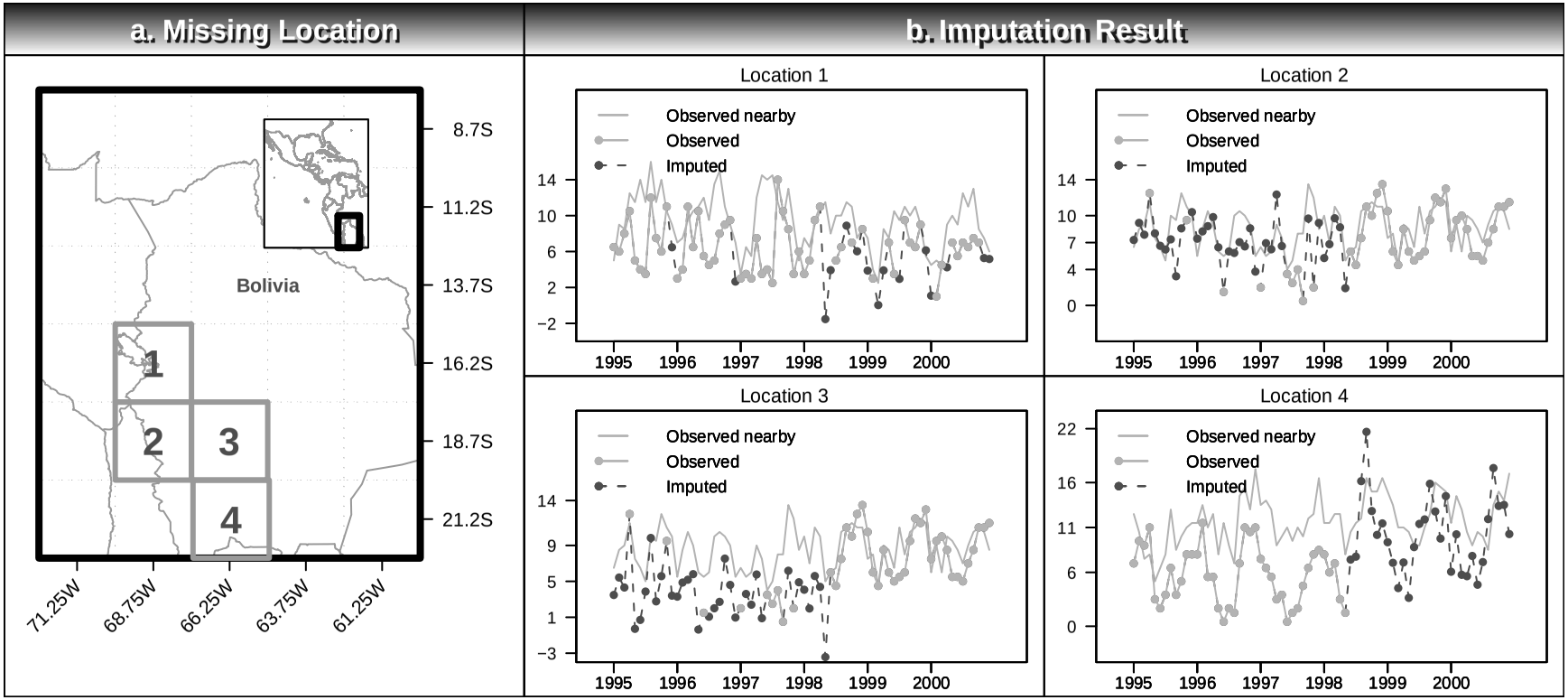

Fig. 1. The specific locations and time periods for missing values are summarized in **a** and **b**. The imputed values are compared with the observed values at nearby locations in **b**. Note that Location 2 and Location 3 share identical values during the observed time period.

adjusted observations were then averaged over the entire time period at each grid location. Finally, the averaged values in the study area were smoothed by the smoothing spline ANOVA

$$\text{Mean SST}_{ij} = c + f_1(\text{lati}_i) + f_2(\text{long}_j) + f_{1,2}(\text{lati}_i, \text{long}_j) + \text{Error}_{ij}$$

where i, j denote a row and column number on a 24 by 24 grid, respectively, c an overall mean, f_1, f_2 main effects and $f_{1,2}$ an interaction, respectively, $f_1, f_2, f_{1,2}$ are nonparametric functions and $\text{Error}_{ij} \sim N(0, \sigma^2)$.

The seasonal decomposition of time series by loess (STL) proposed by (Cleveland *et al.*, 1990) was utilized to obtain a temporal pattern of sea surface temperature (SST) at a given location. By the STL process, the sea surface temperature was decomposed as

$$\text{SST}_t = \text{Trend}_t + \text{Seasonal}_t + \text{Error}_t \tag{2}$$

where $t = 1, 2, \cdots, 72$, denotes each month from 1995 to 2000 in a sequence. From the decomposition in the equation (2), the trend component was extracted as a temporal pattern of sea surface temperature. The STL process is implemented in R function, **stl**, by B.D. Ripley.

It would be most exhaustive to compare a temporal pattern of sea surface temperature on each grid point to its mean sea surface temperature. We instead chose a less exhaustive but more informative comparison as following. By using a dynamic plot, we first narrowed down our attention to three interesting local areas including the equator, the upper coast of Peru and the coast of Chile. These locations are denoted as 1, 2, 3, respectively, in Figure 2. Location 1 and location 2 exhibited the most variation and location 3 the least in terms of temporal patterns. In the selected locations, we averaged the values along the latitude and then fitted a STL model to the averaged values at each longitude. The trends of sea surface temperature at several different longitudes were plotted over time. In Figure 2, the temporal patterns at different longitudes can be compared simultaneously.

3.2 Results

The most exciting result was that we could identify El Niño and La Niña in all the selected locations. As shown in Figure 2, the time periods at the highest and lowest surface temperatures at all three locations correspond with El Niño (1997-98) and La Niña (1995 - 96, 1998-99) periods, respectively. Although the location 3 showed the very constant temporal trend, the effects of El Niño and La Niña are certainly observable. The location 3 showing the lowest mean surface temperature corresponds with so-called Cold-Water Up-Welling areas along the coast of Chile. As an additional minor observation, not directly related to the climate events, El Niño and La Niña, we realized that the mean sea surface temperature in Figure 2 is highest around 10°N rather than the equator.

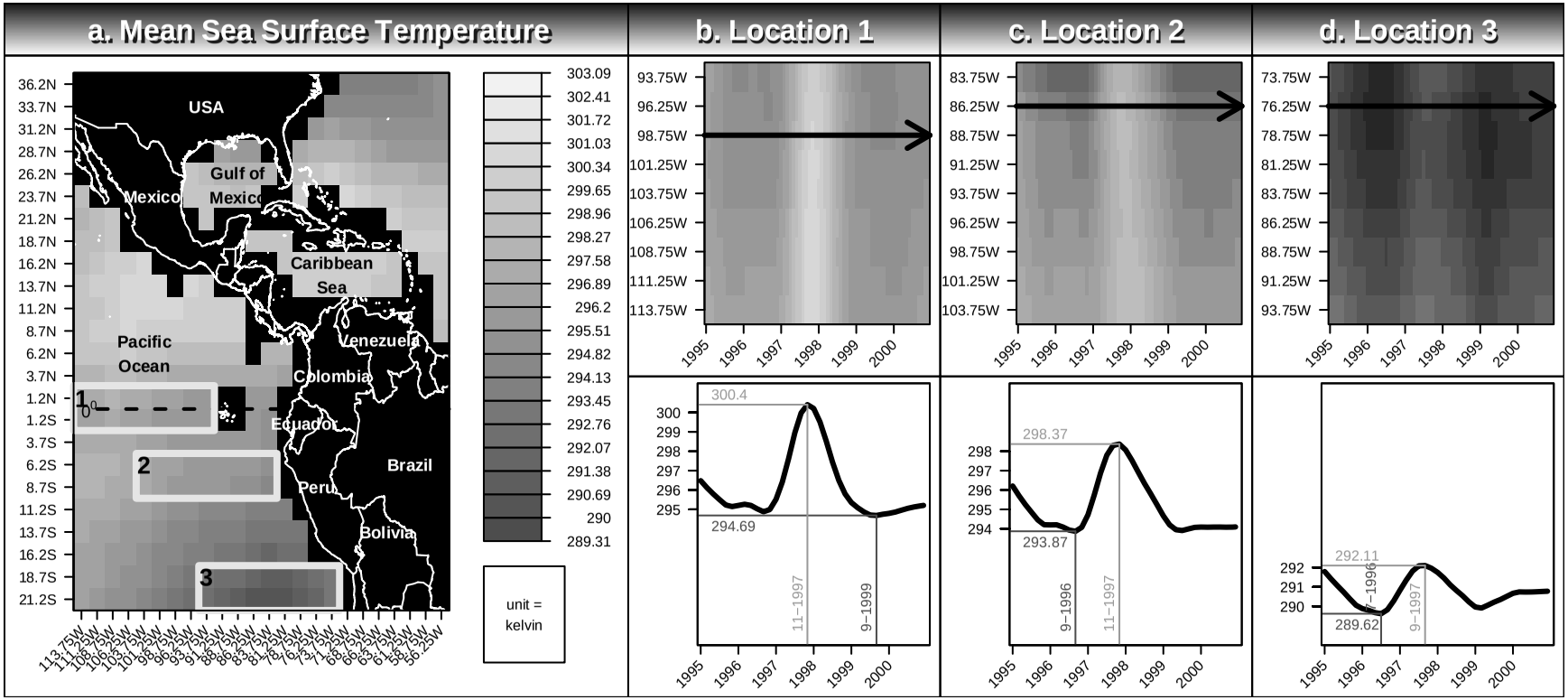

Fig. 2. Mean sea surface temperatures(SST) are plotted together with three locations of our interest in **a**. The lowest mean SST is observed at Location 3, and the highest mean SST around 10°N. The temporal trends of SST in the selected locations are plotted at several different longitudes in **b**, **c**, and **d**. All three locations show the highest and lowest SST in 1997-98 and 1995-6,1998-99, respectively.

4 Ozone Depletion Areas

In this section, we explain our approach to discover a couple of interesting locations indicating ozone depletion. The reduction in ozone levels is a serious global environmental issue, increasing the risk of harmful effects such as skin cancer, cataract rates and crop damage. It was our interest to find some locations showing abnormal ozone trends in the study area.

4.1 Statistical methods and procedure

As a simple approach, a linear model was used to obtain an increasing or decreasing pattern at each grid location. The seasonal component was first removed from the observations at a given location, which consist of 72 values from each month between 1995 and 2000. The residuals of the simple linear model fitted to the adjusted observations implied that the errors were autocorrelated with a lag 1. As such, the linear model with AR(1) error was considered as

$$\text{Adjusted Ozone} = \beta_0 + \beta_1 \text{Time} + \text{Error}$$

where Error $\sim$ AR(1). At each grid location, the estimated regression coefficient, $\hat{\beta}_1$, was plotted to show a linear ozone trend over time.

The linear ozone trends may be related to other variables. In a similar fashion, linear models were fitted to the observations of all the variables except for **pressure** and **elevation**. The linear trends at each grid location were compared among all the variables considered.

As a more general approach, the seasonal decomposition of time series by loess was used to obtain nonlinear ozone trends over time at each grid location as

$$\text{Ozone}_t = \text{Trend}_t + \text{Seasonal}_t + \text{Error}_t \tag{3}$$

where $t = 1, 2, \cdots, 72$, denotes each month from 1995 to 2000 in a sequence. The nonlinear ozone trends were considered since a linear model would be unable to identify other than linear patterns. The nonlinear ozone trends obtained by the decomposition in (3) were then classified into groups using K-means clustering algorithm. In clustering, we used the location information of latitude and longitude as additional inputs to K-means algorithm. As a heuristic approach, the hidden number of groups was determined by minimizing within-cluster dissimilarity. More specifically, the number of groups was first plotted over its corresponding within-cluster dissimilarity, and from the plot a number was then chosen when increasing the number of groups no longer resulted in a significant reduction in the within-cluster dissimilarity (See (Hastie *et al.*, 2001)). Since K-means clustering algorithm is vulnerable to local minima and is sensitive to initial values, we repeated the clustering algorithm 100 times using different random initial values and then chose the clustering result occurring more than 90 times out of 100 repetitions.

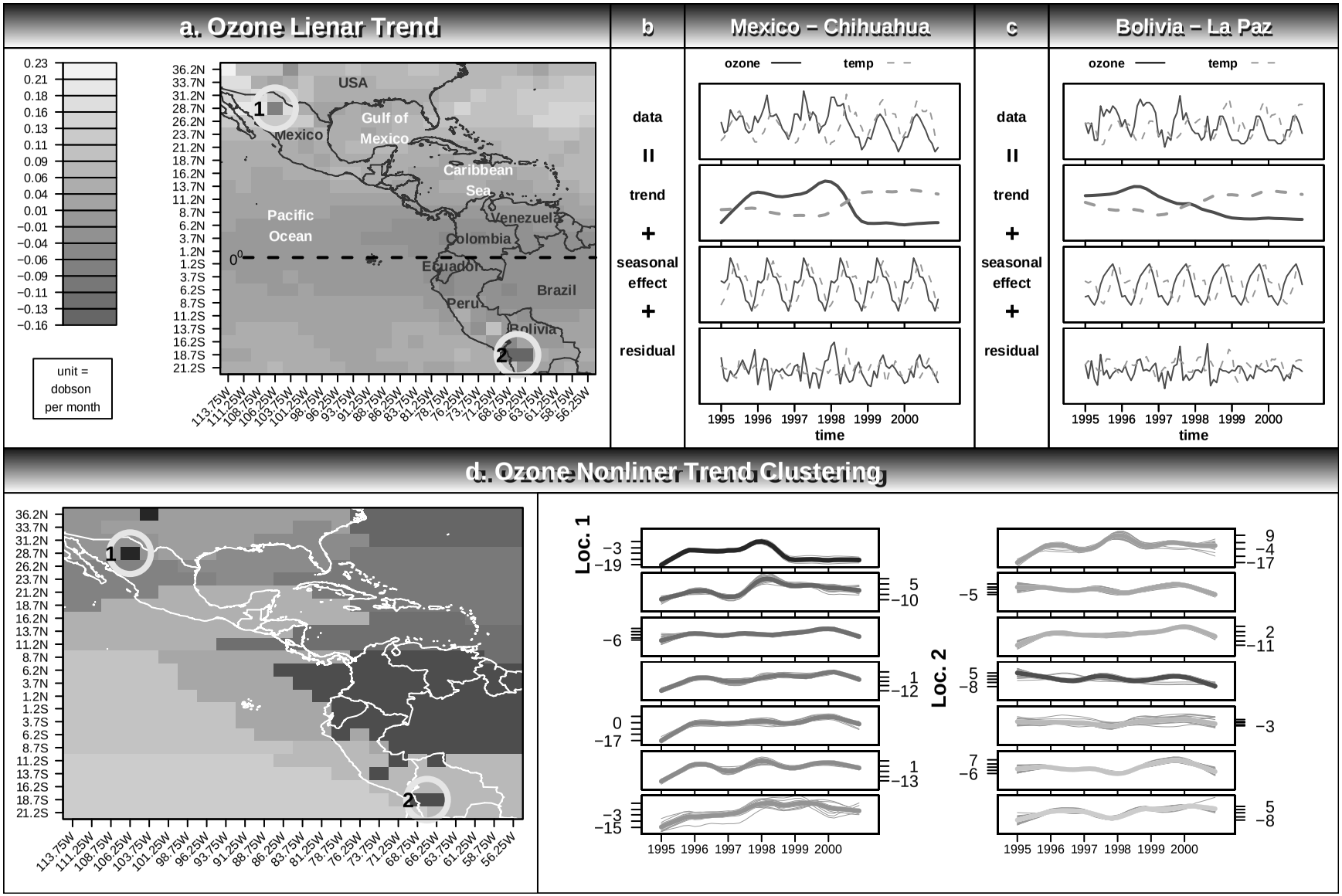

Fig. 3. The linear trends of ozone are shown in **a** and two interesting locations are circled. The STL decomposition of ozone and temperature at Location 1 and 2, respectively, are illustrated in **b** and —**c**. Ozone and temperature show opposite trends. The clustering results of nonlinear ozone trends using K-means clustering are present in **d**.

4.2 Results

By the linear ozone trends, we identified two interesting locations denoted as 1 and 2 in Figure 3. Location 1 showed a linear decreasing trend whereas all other its nearby locations showed linear increasing trends. Location 2 showed the fastest linear decreasing trend among all the locations in the study area. By further investigation, we found that there is a city, Chihuahua, in Location 1 where air and water pollution have been environmental issues and that there is a city, La Paz, in Location 2 where an ozone hole has been reported. Location 1 and 2 were consistently identified as distinct groups in clustering the nonlinear ozone trends. Another interesting observation was that ozone and surface temperature showed opposite linear trends in Location 1 and 2 as shown in Figure 3.

5 Cloud Effect on Temperature

This section describes our approach to find cloud effects on **temperature**. According to the description of the variables in the data, there is a subtle difference between **surface temperature** and **temperature**, i.e. **surface temperature** is the energy emitted from the surface of the Earth under clear sky whereas **temperature** refers to the air temperature near the surface of the Earth. Based on our conjecture that the difference between temperature and surface temperature may result from cloud effects, we considered a linear model to identify any relationship among the five variables including temperature, surface temperature and clouds (high, mid, low).

5.1 Statistical methods and procedure

Our interest here was to find cloud effects on **temperature**. In pursuing the interest, we first tried to identify areas with the marginal homogeneity of each variable in terms of its temporal trend as following. For each variable, the seasonal decomposition of time series by loess was utilized to extract nonlinear temporal trends and then K-means clustering algorithm to classify them into groups. In clustering, the location information was considered as an additional input to K-means algorithm for spatial smoothness.

As a result, we identified two regions consistently clustered as common groups in all the five variables considered. It was interesting that these regions, denoted as Region 1 and Region 2 in Figure 4, are most and least influenced by El Niño and La Niña, respectively.

In order to find relations among the five variables in the selected two regions, we proceeded as following. For each variable, all the observed values belonging to Region 1 and 2 were first averaged in each month from 1995 to 2000, respectively. The averaged values were then adjusted for seasonality by subtracting monthly averages in Region 1 and 2, respectively. The residuals

obtained by the initial fitted linear model were autocorrelated with a lag 1 and thus we considered a linear model with AR(1) error.

5.2 Result

The fitted model in Region 1 around the equator is given as

$$\text{Adj. Temp} = 0.218\ \text{Adj. Surf.Temp} - 0.037\ \text{Adj. Cloud Low} + \text{Error} \tag{4}$$

where Error $\sim$ AR(1) with its estimated autocorrelation, 0.883. The fitted model implies that the adjusted temperature and surface temperature are positively associated (p-value: 0.0002) whereas the adjusted temperature and cloud low are negatively associated (p-value: 0.0032). The negative association between **temperature** and **cloud low** was somewhat reasonable to expect since **cloud low** reflects the energy from the sun resulting in lowering the air temperature.

In Region 2 around the coast of Peru, we found no statistically significant relationship among the variables. As seen from the clustering results in Figure 4, **cloud mid** and **cloud low** showed opposite temporal trends in both regions. The constant trend of **temperature** in Region 2 may be due to the opposite relationship between **cloud mid** and **cloud low**, and it is possible that there exists confounding among covariates.

6 Conclusion

The primary objective of this article was to describe our statistical approaches used to find major results presented in the poster at the Data Exposition 2006. In Section 2, we described the imputation procedure for missing data using the smoothing spline ANOVA. In Section 3, the statistical methods and procedure using nonparametric time series and spatial models were explained in identifying the climate events, El Niño and La Niña. In Section 4, the linear and nonlinear time-series models to discover ozone depletion areas were followed. In Section 5, our approaches to find cloud effects on temperature were discussed.

Computational aspects

All the plots shown in this paper were drawn merely using the"grid" package in R. This powerful and flexible graphical package enabled us to produce very customized plots effectively presenting our main results in a way that we imagined. See (Murrell, 2006) for details on R graphics.

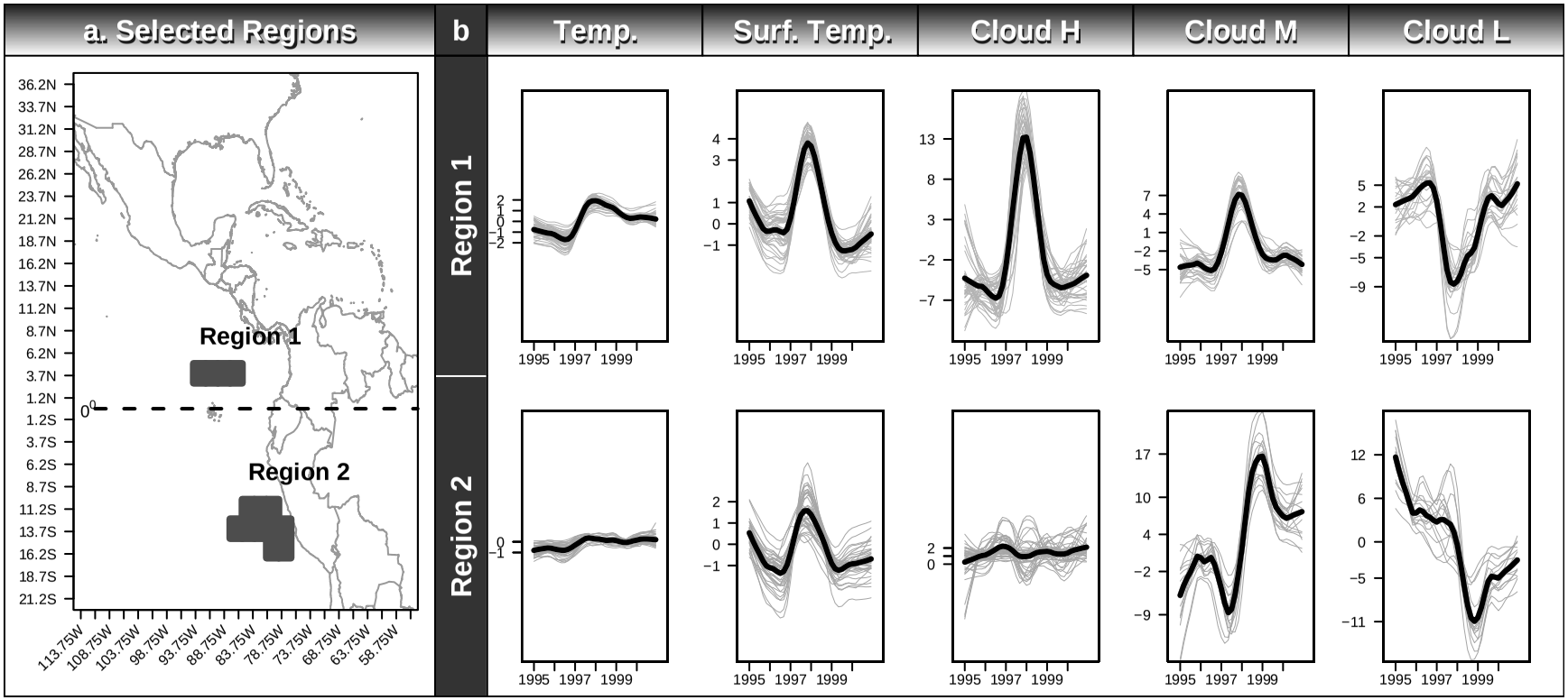

Fig. 4. Two regions sharing homogeneous temporal trends in all five variables are identified in **a**. The temporal trends of the five variables in the selected regions are plotted in **b**.

Acknowledgments

The authors are grateful to Michael R. Kosorok and Grace Wahba for their generous financial support over this project. We would like to express our special thanks to Paul Murrell for his guidance in preparing this manuscript. The authors ought to acknowledge valuable feedbacks on the poster from many anonymous attendees at Data Exposition 2006, especially, regarding with the imputation procedure. This research was supported in part by NIH Grant NCI CA075142, NSF Grants DMS0505636 and DMS0604572, and ONR Grant N00014-06-1-0095.

Correction

Please note that Figure 2 in the poster for Data Exposition 2006 needs a correction and there is a slight difference in the imputed values compared to the part **b** in Figure 1. The difference is due to the fact that smoothing parameters were not fixed in implementing Imputation lemma. See the appendix for more details.

Bibliography

[Cleveland *et al.*, 1990]Cleveland,R.B., Cleveland,W.S., McRae,J. and Terpenning,I. (1990) Stl: a seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics,* **6**, 373.

[Gu, 2000]Gu,C. (2000) *Smoothing Spline ANOVA Models*. Springer series in statistics.

[Hastie *et al.*, 2001]Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learing; data mining, inference, and prediction*. Springer series in statistics.

[Luo, 1996]Luo,Z. (1996). Backfitting in smoothing spline anova, with application to historical global temperature. Ph. d. thesis University of Wisconsion - Madison.

[Murrell, 2006]Murrell,P. (2006) *R Graphics*. Chapman & Hall/CRC.

[Wahba, 1990]Wahba,G. (1990) *Spline models for observational data*. SIAM.

[Wahba and Luo, 1995]Wahba,G. and Luo,Z. (1995). Smoothing spline anova fits for very large, nearly regular data sets, with application to historical global climate data. Technical report University of Wisconsion - Madison.

Appendix: Leaving-Out-K and Imputation Lemma

Here we elaborate on the computational issue as the imputation procedure was discussed in Section 2. As mentioned earlier, the Leaving-Out-K lemma and the Imputation lemma both can be used for the imputation purpose using the smoothing spline ANOVA. Just in case, we include both lemmas at the end of the appendix.

The Imputation lemma proposed by (Wahba and Luo, 1995) was originally intended for incomplete data. The implementation of the lemma requires an iteration, but with a good initial guess the convergence would not be an issue. In Section 2, we used as the initial guess the predicted values obtained from the initial fitted model in step 1. The imputation procedure completed in just one iteration. As a technical comment, the predicted values here should not be confused with the imputed values which can be obtained by the Leaving-Out-K lemma.

As addressed earlier in Section 2, both lemmas are currently not implemented in **gss** package. However, the Imputation lemma can be used with a minor modification of **ssanova** function in **gss** package. Note that the results of the Imputation lemma hold for fixed smoothing parameters. The **ssanova** function determines optimal smoothing parameters as default and does not provide an optional argument to use fixed smoothing parameters. By modifying the **ssanova** function, we imputed the missing values as summarized in Section 2.

From some helpful feedback at JSM, we later realized that the Leaving-Out-K lemma could also have been used. In this purpose, the function **ssanova1** can be used with a minor modification. This alternative approach has a benefit since it does not require an iteration step.

As a technical comment, the detailed justification using the **ssanova1** for the Leaving-Out-K lemma is given in the following equation.

$$\min_f \left(\frac{1}{n} \sum_{i=1, i \notin S_K}^{n} (y_i - f(t(i)))^2 + \lambda \|P^\star f\|^2 \right) \tag{5}$$

$$\propto \min_f \left(\frac{1}{n-K} \sum_{i=1, i \notin S_K}^{n} (y_i - f(t(i)))^2 + \frac{n}{n-K} \lambda \|P^\star f\|^2 \right). \tag{6}$$

(5) is the objective function of the Leaving-Out-K lemma and (6) is its equivalent objective function of the smoothing spline ANOVA. Both objective functions do not include missing data. The connection above is due to the fact that both objective functions rely on subsets of basis functions, which do not have to be chosen from the basis functions corresponding with missing data. By using a fixed but modified smoothing parameter, $\lambda^* = \frac{n}{n-K}\lambda$, we can impute at one step following the Leaving-Out-K lemma.

Lemma 1 *Leaving-Out-K Lemma, (Wahba and Luo, 1995)*
Let an reproducing kernel Hilbert space $\mathcal{H}$ be decomposed as $\oplus_{\beta=0}^{p}\mathcal{H}_\beta$, and for $f \in \mathcal{H}$ let $\|Pf\|^2 = \sum_{\beta=1}^{p}\theta_\beta^{-1}\|P^\beta f\|^2$, where P^β is a projection onto $\mathcal{H}_\beta$ and θ_β are smoothing parameters . Let $f^{[K]}$ be the solution to the variational problem:

$$\min_{f\in\mathcal{H}} \frac{1}{n} \sum_{\substack{i\,=\,1\\ i\,\notin\,\mathcal{S}_K}}^{n} (y_i - f(t(i)))^2 + \lambda\|Pf\|^2$$

where $\mathcal{S}_K = i_1, \ldots i_K$ is a subset of $1, \ldots, n$ with the property that above functional has a unique minimizer, and let $y_i^\star, i \in \mathcal{S}_K$ be imputed values for the missing data imputed as $y_i^\star = f^{[K]}(t(i)), i \in \mathcal{S}_K$. Then the solution to the problem:

$$\min_{f\in\mathcal{H}} \frac{1}{n} \left(\sum_{\substack{i\,=\,1\\ i\,\notin\,\mathcal{S}_K}}^{n} (y_i - f(t(i)))^2 + \sum_{i\in\mathcal{S}_K} (y_i^\star - f(t(i)))^2 \right) + \lambda\|Pf\|^2$$

is $f^{[K]}$.

Lemma 2 *Imputation Lemma, (Wahba and Luo, 1995)*
Let $\hat{f}$ be the minimizer of the variational problem:

$$\min_{f\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(t(i)))^2 + \lambda\|Pf\|^2$$

and write $\hat{f}$ as $A(\lambda)y$ by using influence matrix $A(\lambda)$.

Denote $A(\lambda) = \begin{pmatrix} A_{11} & \vdots & A_{12} \\ \ldots & \ldots & \ldots \\ A_{21} & \vdots & A_{22} \end{pmatrix}$. Let $g_{(0)}^2$ be a K-vector of initial values for an impuation of $(f^{[K]}(t(i_1)), \ldots, f^{[K]}(t(i_K)))^T$, and suppose $0 \prec (I - A_{22})$. Let successive imputations $g_{(l)}^2$ for $l = 1, 2, \ldots$ be obtained via

$$\begin{pmatrix} g_{(l)}^1 \\ \ldots \\ g_{(l)}^2 \end{pmatrix} = A(\lambda) \begin{pmatrix} y^1 \\ \ldots \\ g_{(l-1)}^2 \end{pmatrix}.$$

Then

$$\lim_{l\to\infty} \begin{pmatrix} g_{(l)}^1 \\ \ldots \\ g_{(l)}^2 \end{pmatrix} = \begin{pmatrix} f^{[K]}(t(1)) \\ \vdots \\ f^{[K]}(t(n)) \end{pmatrix}.$$