1 Introduction

We describe a set of exploratory tools to investigate and to model spatial-temporal patterns of diseases. The essential idea is to integrate statistical and visual modelling with interactive visual representation. The disease data represents multiple spatial-temporal processes where each reported observation (number of cases in space-time region) is an aggregate quantity (averaged over the particular region). The implemented visual representations of such data include static and dynamic maps and time series plots. The modelling tools include transformations, color mappings, smoothing, aggregation, estimation of dependence structure, and other models. The toolset can be easily extended by a non-expert using the S language (Becker, Chambers, and Wilks, 1988) . We start by describing the data, then describe our approach and some features of the implementation.

2 Data

The data we use were obtained from the Centers for Disease Control and Prevention (CDC) which operates the National Notifiable Diseases Surveillance System (NNDSS) in partnership with the Council of State and Territorial Epidemiologists (CSTE). The CDC collects weekly provisional information on the occurrence of diseases that are defined as "notifiable" by CSTE. Further details concerning the NNDSS can be found, for example, in Chorba et al. (1989).

The dataset contains weekly by state reports on 57 diseases for the period between 1980 and 1994. There are 783 report weeks in this period and the reports are provided for 51 states, 3 territories and New York City. The names of the reported diseases, the total number of cases over the reported period, and the number of missing reports are in Table **??**. The table conveys a general idea of how widespread each particular disease is and how much care is taken to report the disease cases.

3 Multivariate Interactive Animation System For Map Analysis

We previously analyzed similar data on the disease mumps in Eddy and Mockus (1993). The analysis we performed produced successively smoother non-interactive dynamic maps of the incidence rates and a dynamic map of the residuals from a two way analysis of variance model. Here we generalize this approach to multiple diseases. We have designed a system to integrate dynamic and static maps, transformations, smoothing, and other techniques for spatio-temporal modelling and visualization. We refer to this system as MIASMA (for Multivariate Interactive Animation System For Map Analysis).

The system accepts observations in the form of a three dimensional array. The first dimension ranges over all regions in space (e.g., counties or states). The second dimension ranges over all time periods (e.g., days, weeks, months, years). The third

Name	# of Cases	# of Missing	Name	# of Cases	# of Missing
Asep-Mening	135327	399	GC-Mil	31221	2529
Brucellosis	1053	13469	Syphilis-(Total)	457212	341
Chickenpox	374072	36404	Syphilis-Civ	350404	3901
Diphtheria	30	16755	Syphilis-Mil	566	25303
Enceph-Prim	15346	266	Rabies-Animal	79813	220
Post-Eceph-(Total)	2666	229	Antrax	4	21747
Post-Eceph-CPox	1350	21746	Botulism	372	21747
Post-Eceph-Mump	25	21747	Cong-Rubella	68	21747
Post-Eceph-Other	468	21747	Leprosy	2745	5263
Hep-B	280586	464	Leptospirosis	388	21746
Hep-A	328964	467	Polio-Total	17	21747
Hep-unsp	66036	480	Polio-Paralytic	16	21747
Malaria	15353	290	Polio-Nonparalytic	0	21747
Measles	69714	406	Polio-Unspecified	1	21747
Mening-Inf (Total)	36703	242	Plague	65	21747
Mening-Inf-Civ	7576	25293	Pisttacosis	535	21747
Mening-Inf-Mil	10	25305	Rabies-Human	19	21747
Mumps	64535	1045	Cholera	164	21747
Pertussis	38241	403	Hep-NA-NB	40045	5942
Rubella	13980	410	Legionellosis	11379	5924
Tetanus	411	16710	Measles-Indigenous	48668	8660
Tuberculosis	327441	351	Measles-Imported	3500	8664
Tularemia	2816	226	Toxic-Shock-Syndrome	3424	8664
Typhoid-Fever	5786	293	BOT-Food	132	21747
RMSF	11067	294	BOT-Infant	198	21747
Typhus-Murine	20	38782	BOT-Other	42	21747
Trichinosis	166	21747	H.-Influenzae	10718	30233
Gonorrhea-(Total)	10522000	338	Lyme-Disease	26525	30763
GC-Civ	8874720	3901			

Table 1: The list of diseases and numbers of cases

dimension lists all the quantities of interest (e.g., population size, number of cases for a particular disease, incidence rates for a particular disease, or derived quantities). Notice, that we only consider quantities that are averages over regions in space and time. The first two dimensions of the array define a dynamic map and the last one implies that there may be multiple dynamic maps.

The input to the system can be either original reported data or the output from a statistical model (e.g. residuals and/or effects of a spatial temporal model). Those, we can successively fit models and inspect residuals until we obtain a satisfactory fit.

The system is implemented as a selection of different classes of tools. The most important classes include multiple visual representation, transformation, handling of missing values, aggregation and smoothing, superpositions of several quantities, and statistical model fitting toolsets.

We separate modelling tools into two groups: models for visual representation, and models for statistical analysis. The models for visual representation must be tightly integrated into the system to allow adequate interactive response, while statistical models can be loosely integrated via file sharing. We separate the implementation of modelling and visualization tools to simplify extensions to the system.

3.1 Display

An example display of our system is given in Figure **??**.

The system consists of the main control window and various view windows. The

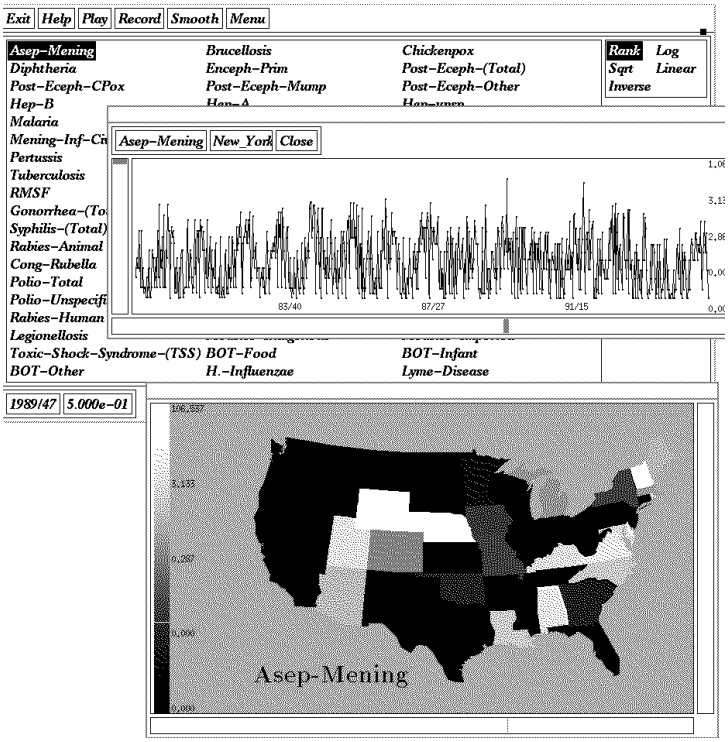

Figure 1: Control window and two views of the MIASMA

control window contains menus and selection lists. Modelling and transformation methods are controlled from the main window. In Figure **??** the data on the disease *aseptic meningitis* is selected and the rank transformation is being used. The main window also contains the current date (year and week) for the dynamic map view shown at the bottom. A time series plot of the disease incidence in Alabama is in the window overlapping the control window. The state and the disease can be selected interactively using scrollbars at the bottom and at the left of the time series plot.

3.2 Visual Models

The models/tools for visual representation include view selection, spatial smoothing, time interpolation, transformation, color mapping, display of missing values, and display of multiple quantities. The available views are time series, static maps, and dynamic maps. Each type of view is presented in a separate window and an unlimited number of windows can be created for each view.

The time series view presents the time series of the data for a particular quantity and spatial region over the available time period. Two scrollbars allow selection of the spatial region and quantity of interest.

The static map view displays a static map of the data where color of a pixel in the window encodes the value of the quantity at the particular location in time and

space. Since the quantities are aggregates over a space-time region we provide several interpolation techniques. The simplest one just shows a constant value within each region. As with a time series view the two scrollbars allow selection of the quantity to be displayed and the time moment.

The dynamic map view is an animation of a static map. As with the static map, several interpolation techniques are provided. The simplest—piecewise constant interpolation—shows a constant value within each region and within each time interval. A slightly smoother animation can be obtained by interpolating linearly between two time intervals, namely, two maps are computed for each time period as the key frames and then linearly interpolated for the intermediate frames. The number of intermediate frames can be changed to increase or decrease the speed of the animation. A spatially smooth animation can be obtained by spatial interpolation as described in Eddy and Mockus (1993). The smoothing parameter and starting location for the animation can be selected using scrollbars.

We consider spatial and temporal smoothing as a visual model since it facilitates perception of the dynamic map. Time smoothing removes jumps in time (the jumps create a distracting blinking appearance) and smoothing in space hides region boundary artifacts (which are clearly perceived by an observer but are often irrelevant to the quantity of interest). Time-space smoothing represents a simple visual model that allows qualitative (without excessive detail) display of data that was aggregated over regions in time and space. We have found that in a dynamic display it is important to limit the amount of information since it is not trivial to perceive every detail even with the high bandwidth of a human visual system. Smoothing can be viewed as fitting of a statistical model; we discuss that approach later.

Essential visual representation models are various transformations of the data into the range of available display attributes. The simplest—linear transformation—can often be inadequate due to the discrete nature and limited range of display attributes, such as, pixels, colors, patterns. We have found that different transformations emphasize different features present in the data, e.g., few large outliers can make a dynamic map look almost constant if a linear transformation is used. Rank transformation, on the other hand, ignores the outliers and make the distribution of the display attributes uniform over the available range. In addition to linear and rank transformations we have arbitrary power transformations (extended by a logarithmic transformation). In the case of dynamic and static maps we encode the value of a quantity of interest by color. The transformation (mentioned above) converts each value into an integer code in the range between 1 and 256 (code 0 is reserved for the background). Those codes are then displayed according to the colormap that maps each code into a color. The colormap can be selected interactively. For a discussion on how to select colormaps to convey quantitative information see Levkowitz and Herman (1992).

3.2.1 Display of Missing Values

By inspecting Table **??** it becomes apparent that a substantial amount of observations are missing. It is essential to address this problem in constructing a visual

representation of the missing data.

We implement a number of ways to address this problem. In the time series view we show the data as small dots. Missing data is absent from the display, although we can infer its presence from the larger horizontal gaps. We take two different approaches in the case of static and dynamic maps. The two alternatives are to leave the missing data out (use a neutral, background, transparent, or some other color that is not present in the color scale) or to fill in some color using the available data. Currently we show missing values in the background color or impute the color (by taking a median value for each time moment) and add a pattern to indicate that that value was not observed.

3.3 Statistical models

We have found several statistical models very useful to study these data. We start from the simplest but very useful models of aggregation, then consider smoothing methods, a two-way table, estimation of dependence structure, and best linear prediction (kriging).

3.3.1 Aggregation

Various aggregation methods can dramatically reduce the amount of data and simplify the inspection process. By aggregation we mean reduction of the number of observations in our dataset (represented by a 3-dimensional array (diseases, time moments, states)) by combining several cells into a single cell. The aggregations differ in which cells are selected for aggregation (neighborhoods) and in which method is used to produce a single value out of values in the neighborhood. Another operation which we call smoothing operates the same way as aggregation except it does not reduce the number of cells in the data array, i.e., for each cell a neighborhood of cells is defined and then the combine operator is applied to the values in the neighborhood of a cell to produce a single value for that cell. Since the smoothing and aggregation methods are so similar we will consider only aggregation methods. Analogous selection of neighborhoods and combine operators is available for smoothing operations too.

Definition of neighborhoods. We consider several ways to select neighborhoods by selecting a direction in the data array and by selecting the size of the neighborhood. Since the data is represented by a 3-dimensional array $x_{i,j,k}$ the aggregation direction is defined by selecting the index of the array. The first index corresponds to spatial location, the second - to time intervals, and the third - to diseases.

For example, we can aggregate over time intervals with the window size of 4 time intervals to convert weekly data to monthly data. The monthly data array

$$X_{i,J,k} = \mathrm{Agg}(x_{i,4J,k}, x_{i,4J+1,k}, x_{i,4J+2,k}, x_{i,4J+3,k}),$$

where $x_{i,j,k}$ is weekly data array, and Agg() is aggregation operator. The aggregation/smoothing operator can be selected independently of the neighborhood selection.

It can range from a simple sum ($\mathrm{Agg}(x, y, z, w) = x + y + z + w$) to an ARIMA filter, where the result represents parameter values of the ARIMA process, or the smoothed version of of the time series.

We can investigate periodic behavior in the dataset by defining appropriate neighborhoods. For example, to produce a standard yearly cycle from weekly data we can aggregate values for a particular week over all years.

Spatial neighborhoods need be treated differently from time neighborhoods. To define spatial neighborhoods we need to define adjacencies between the locations of observations because the simple ordering by time is no longer present. In our case regions A_i (states) form a partition of A (the continental US). We define two spatial regions to be adjacent (or one-adjacent) if they share a common border consisting of more than one point. If there is a region to which they both are adjacent then we call them two-adjacent. Similarly we can define k-adjacent regions. The sizes of the spatial neighborhood is the number k.

It may be of interest to aggregate over different quantities (diseases) in an attempt to capture relationships between different diseases. Any composition of aggregation/smoothing methods can be performed within MIASMA.

Aggregation operators can be divided into several classes: arithmetic, order, selection, composition, and other. The arithmetic operators include sum and variance, the order operators include various quantiles, the selection operators select a value(s) based on position within the neighborhood (section, several sections). Composition of the operators is also possible. More complicated complicated operators are described in the next section.

3.3.2 Other statistical models

Given a complicated structure of observations it seems useful to be able to inspect the model and the residuals. For example, let z_{ijk} be the reported incidence rates of disease k in state i for month j. We can use median polishing to fit a model $z_{ijk} = s_{ik} + t_{jk} + \eta_{ijk}$, where s_{ik}'s are state-disease effects and t_{jk}'s are time-disease effects. Since the model contains a large number of parameters we may inspect the parameters as well as the fit using the tools available in MIASMA. To do that we can use derived datasets $\mathrm{Effects}_{ijk} = s_{ik} + t_{jk}$ or $\mathrm{Residuals}_{ijk} = \eta_{ijk}$ instead of the original observed incidence rates. In particular, the residuals do not have seasonal and longer term trends observed in the disease reports, but can indicate unusually high incidence rates (epidemics).

Another approach is to model the disease as a space-time process. The observed data would represent the integrals of the incidence rate process over the regions A_{ij} in space and time, $x_{ijk} = \int_{A_{ij}} f_k(x)dx$. To predict the process f_k at a particular location in space and time one has to estimate (or know) the covariance function of the process f_k. For a reference on spatial prediction see Cressie (1991). Estimation of the spatial covariance function from aggregate data is described in Mockus (1994).

4 Summary

We designed and implemented a system (MIASMA) to analyze spatio-temporal patterns in a dataset containing weekly reports of 57 diseases in the United States. The system integrates statistical and visual representation tools for interactive modelling and exploratory analysis of similar datasets. The visual representation tools can be used to look at the raw data, at the fitted models, or at the residuals from the fitted models. We found that interactive model fitting and exploratory analysis is essential in dealing with large spatial dataset. Our modeling and visualization tools are geared to analyze aggregate data, when observations represent averages over space-time regions of some underlying process. In particular we implement smoothing, interpolation, aggregation, and prediction methods for aggregate space-time dataset. We separated implementation of the visual representation and statistical analysis tools to simplify extensions of our system.

5 References

Becker, R.A., Chambers, J.M., Wilks, A.R. (1988). *The New S Language.* Wadsworth and Brooks/Cole, Pacific Grove, CA.

Centers for Disease Control (1989). Mumps – United States, 1985-1988, *Morbidity and Mortality Weekly Report,* **38**(7): 101-5.

Chorba, T.L., Berkelman, R.L., Safford, S.K., Gibbs, N.P., and Hull, H.F. (1989). Mandatory Reporting of Infectious Diseases by Clinicians. *Journal of the American Medical Association*, **262**: 3018-3019.

Cressie, N.A.C. (1991). *Statistics for Spatial Data.* John Wiley and Sons, New York.

Eddy, W.F. and Mockus, A. (1993). An Example of Noninteractive Dynamic Graphics for Manufacturing Process Data. *International Statistical Review,* **61**(1): 81-95.

Levkowitz, H. and Herman, G.T. (1992). Color scales in image data, in *IEEE Computer Graphics & Applications* January, 1992. 72-80.

Mockus A. (1994). Predicting a Space-Time Process from Aggregate Data Exemplified by the Animation of Mumps Disease. PhD Thesis. Department of Statistics, Carnegie Mellon University.