A Visual Analytics Approach to High-Dimensional Logistic Regression Modeling and its Application to an Environmental Health Study

Chong Zhang, Jing Yang *
University of North Carolina at Charlotte

F. Benjamin Zhan, Xi Gong †
Texas State University

Jean D. Brender ‡
Texas A&M Health Science Center

Peter H. Langlois §
Texas Department of State Health Services

Scott Barlowe ¶
Western Carolina University

Ye Zhao ‖
Kent State University

ABSTRACT

In the domain of epidemiology, logistic regression modeling is widely used to explain the relationships among explanatory variables and dichotomous outcome variables. However, logistic regression modeling faces challenges such as overfitting, confounding, and multicollinearity when there is a large number of explanatory variables. For example, in the birth defect study presented in this paper, variable selection for building high quality models to identify risk factors from hundreds of pollutant variables is difficult. To address this problem, we propose a novel visual analytics approach to logistic regression modeling for high-dimensional datasets. It leverages the traditional modeling pipeline by providing (1) intuitive visualizations for inspecting statistical indicators and the relationships among the variables and (2) a seamless, effective dimension reduction pipeline for selecting variables for inclusion in high quality logistic regression models. A fully working prototype of this approach has been developed and successfully applied to the birth defect study, which illustrates its effectiveness and efficiency. Its application in an insurance policy study and feedback from domain experts further demonstrate its usefulness.

Keywords: High-dimensional, Logistic Regression, Dimension Reduction, Visual Analytics

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interfaces (GUI); G.3 [Probability and Statistics]: Correlation and regression analysis

1 INTRODUCTION

Birth defects are structural or chromosomal abnormalities that a baby has at birth. Despite the significant morbidity and mortality associated with these conditions, causes for an estimated 65% to 75% of birth defects remain unknown [4]. Scientists do not fully understand the specific relationships between the environment and abnormalities. To help uncover these relationships, the authors have been working on a project with the aim of revealing associations between maternal exposure to air pollutants and congenital malformation in offspring.

A large, high-dimensional environmental health dataset has been created for this study. In particular, information for 60,613 cases (births with major congenital malformations) and 244,927 controls (births without major congenital malformations) between 1996 and 2008 were retrieved from the Texas Birth Defects Registry and birth records. The dataset includes neural tube defects, heart defects, oral clefts, and limb reduction defects. Information about the release of 449 toxic chemicals in Texas during the same period was retrieved from the Toxics Release Inventory (TRI) database of the United States Environmental Protection Agency (EPA). The case and control data and the TRI data were linked using procedures developed by the research team [3, 37]. For each case and each control, maternal exposures to the 449 chemicals were calculated and recorded as numerical values [38]. Five maternal and infant characteristics, such as the mother's age group, level of education, and the gender of an infant, are recorded as categorical values.

Regression models [18] can be used to establish relationships between response variables and explanatory variables. The interpretation of regression coefficients (β parameters) is the expected change in the response variable for a one-unit change in an explanatory variable while holding other explanatory variables in the model constant. In our study, the response variable defines whether a child is born with one of the selected birth defects and the explanatory variables are the chemical exposure and the maternal and infant attributes. Because the response variable is dichotomous, we model the logit-transformed probability [14] of having the birth defect $\pi(x)$ as a linear relationship with the explanatory variables. Univariate logit models (i.e., $logit[\pi(x)] = \beta_0 + \beta_1 X$) may lead to an abundance of false-positives [31] because the effect of X might be caused by other explanatory variables that are not considered in the model. In contrast, multivariate logit models (i.e., $logit[\pi(x)] = \beta_0 + \beta_1 X1 + \beta_2 X2 + ... + \beta_p Xp$) simultaneously measure the relationship between the dichotomous outcome variable and multiple explanatory variables. This allows the model to distinguish false-positives from true risk factors that should be further confirmed/rejected through subsequent epidemiological analyses.

The high-dimensional nature of our dataset brings significant challenges to logit modeling. They include (1) **Overfitting**: A high-dimensional logit model may describe noise or random error instead of the underlying relationship between the outcome and explanatory variables [14], (2) **Confounding**: A confounder is an extraneous variable associated with both the outcome and one or more true risk factors. Along with the explanatory variable, it may explain all or part of the observed effect of the true risk factors thereby complicating and perhaps masking the true relationship between the outcome and the explanatory variables. Extreme confounding will lead to multicollinearity (see below), (3) **Multicollinearity**: If two highly correlated variables are placed into the same model, it may become unstable or over/under-estimate variable effects [5], and (4) **Weak effect**: A weak association with the outcome may not be easily found without eliminating the influence of other related explanatory variables.

Automatic approaches to selecting variables for logit model building, such as Forward Stepwise selection [7], Ridge Regression [6], LASSO [33], and the Elastic Net [39], often result in models that are unstable, non-reproducible, or have extra parameters and bias. Therefore, it is desired to allow users to participate in

*e-mail: {czhang22, jyang13}@uncc.edu
†e-mail: {zhan, gong}@txstate.edu
‡e-mail: jdbrender@sph.tamhsc.edu
§e-mail: peter.langlois@dshs.state.tx.us
¶e-mail: sabarlowe@email.wcu.edu
‖e-mail: zhao@cs.kent.edu

explanatory model building. They can identify confounders, determine variable inclusion and exclusion, and interpret weak associations.

To address this need, we propose a visual analytics approach that facilitates the building of high quality logit models for risk factor identification. To the best of our knowledge, our approach is among the first visual analytics efforts toward this purpose. The approach is motivated by the birth defect study and is general enough to be used in other application domains where high-dimensional logit modeling is needed for explanation. The main contributions of this paper include: (1) A design study where visual analytics techniques are developed for identifying birth defect risk factors from a high-dimensional environmental health dataset, (2) A novel visual analytics approach to high-dimensional logit modeling. It seamlessly integrates statistical procedures and visualization techniques to make the modeling process easy, intuitive, and more accurate than automatic approaches. A fully working prototype has been developed (see Figure 1) which received positive feedback from two epidemiologists and a statistician, (3) Case studies where the prototype was used to identify potential risk factors for limb reduction defects and characterize caravan insurance policy holders.

2 Related work

Logit model building for high-dimensional data is an important research topic and many automatic methods have been proposed. However, all existing methods have drawbacks for our project. Stepwise Forward selection [7] often exhibits high variance and severe bias [14, 33]. Ridge Regression [6] shrinks the coefficient estimates of all variables even though it performs well in terms of multicollinearity. LASSO [33] is limited when handling highly correlated variables because it tends to choose only one among a group of variables with high correlations. The Elastic Net [39] suffers from a double amount of shrinkage which introduces unnecessary bias. In addition, these methods cannot provide reliable confidence intervals [22], which is not acceptable in epidemiology. Our visual analytics approach proposes a new solution to address this problem.

Epidemiologists rely heavily on statistical software packages, such as R [25] and IBM SPSS [15], to conduct logistic regression modeling. These packages allow users to conduct univariate analysis, variable selection, and multivariate logistic regression through a menu-driven or command line interface. However, they do not provide an integrated pipeline allowing users to conduct the multiple steps necessary for high-dimensional logit model building. Compared to these packages, our prototype provides a more integrated, transparent, user-friendly, and efficient working environment to conduct high-dimensional logit modeling.

In the visualization community, variable selection has been widely studied. The Value and Relation display [35] visually conveys the correlation among the variables with textures and distances and allows users to interactively select correlated or non-correlated variables. DimStiller [16] is a visualization system with a dimension analysis and reduction workflow where users can interactively transform the data using a variety of techniques. SmartStripes [23] allows users to step through the feature selection process manually. Similar to our work, it uses feature partitions for dimension reduction. It is designed to be a preliminary analysis tool and does not consider cause-and-effect relationships. Fernstad et al. [8] use a set of interestingness measures for dimension filtering and organize correlated variables into clusters for effective subspace visual exploration. These works cannot be directly used to address the problem in this paper since they do not consider the association among explanatory and response variables in the regression relationship.

A few visual analytics approaches have been proposed for regression modeling in recent years. Steed et al. [30] use parallel coordinates to build linear regression models for hurricane activity prediction. Bögl et al. [2] apply line, bar, and scatter plots to a well-known statistical Box-Jenkins methodology for time series data regression modeling. Krause et al. [20] present a tool called "INFUSE" for comparing predictive features across feature selection and classification algorithms. Mühlbacher et al. [24] propose a partition-based framework for predictive linear regression model building. There are several differences from our approach. First, we focus on the visual analytics of multicollinearity, confounding, and weak effect instead of partitions. Second, many measures (i.e. R-squared, RMSE, and OLS) are not as well suited as logit regression because our response variables are not continuous. Third, all of those approaches target predictive modeling while explanatory regression modeling is needed in our project. According to Shumeli [29], explanatory regression modeling is fundamentally different from predictive regression modeling and this distinguishes our work from existing efforts. Predictive regression modeling aims to minimize prediction errors but explanatory regression modeling emphasizes characteristic expressions and the relationships among variables for distinguishing between false-positive variables and true risk factors. The criteria for variable selection differ significantly in these two contexts [29].

3 Requirement Analysis

A team of epidemiologists, computer scientists, and geographic scientists, including most of the co-authors, participated in the environmental health study. A set of essential tasks have been characterized through intensive meetings and discussions.

Task 1: Dimension Reduction. According to a general guideline in logistic regression modeling [14], the complexity of multivariate studies can be reduced by first using univariate statistical indicators to filter out irrelevant or non-significant variables. These variables are ruled out as risk factors and confounders and do not need to be considered in further analysis. Multiple indicators are often used for considering different aspects of measurements in this dimension reduction process.

Task 2: Dimension Relationship Analysis for Model Building. Since univariate analysis may introduce false-positives, a variable should always be analyzed with other correlated variables in a multivariate logit model. To avoid problems such as overfitting, confounding, and multicollinearity, the relationship among the variables needs to be carefully examined. First, variables correlated with the selected variables may need to be included into the model even if they are not in the initial candidate variable set. Second, groups of highly correlated variables need to be identified whose variability can be defined with a smaller set of latent variables. Variables within such groups should not be placed into a single super model. Rather, they should be placed into different models that are smaller and more stable [5]. Third, confounders need to be separated for further epidemiological analyses.

Task 3: Effect Change with Demographic Characteristics. The cases and controls in this project contain maternal and infant characteristics. Some of the characteristics may not be direct causes of birth defects, but instead modify the relationship between environmental exposure and the risk of birth defects in offspring. To assess the effect of a chemical, demographic characteristics need to be considered.

Task 4: Model Evaluation and Result Reporting. Due to challenges such as weak association and multicollinearity, there is no super model that can assess all the variables at the same time. Multiple models need to be built and evaluated interactively and progressively. The results need to be effectively conveyed to users so that they can conduct further model building and report the findings.

To effectively support the above tasks, we argue that our visual analytics system should have the following features: **(1) Integration:** The system should support the general workflow of risk factor analysis and carry out all the aforementioned tasks in a seamless pipeline. **(2) Effective dimension reduction:** Intuitive visualiza-

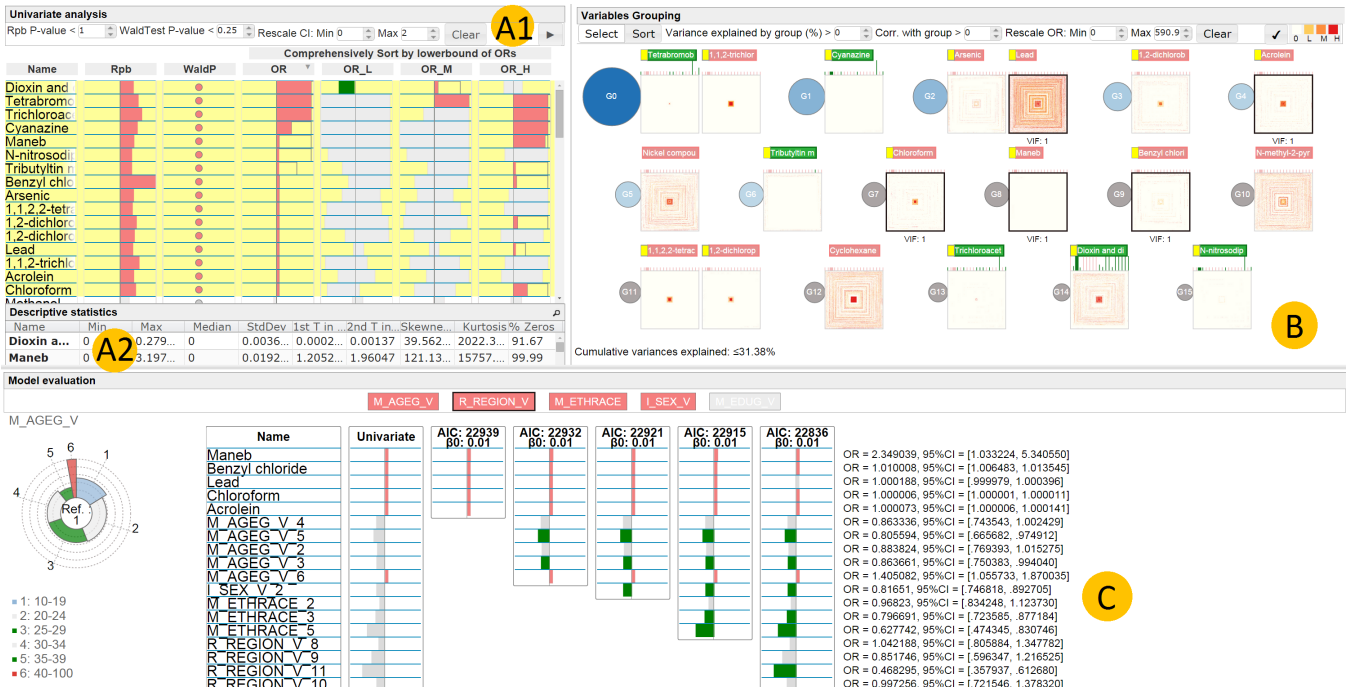

Figure 1: The interface of the prototype where the limb reduction defect is analyzed. (A1) The univariate analysis view. (A2) The descriptive statistics information panel. (B) The variable grouping view. (C) The model evaluation and comparison view.

tion and flexible interactions should be provided so that users can efficiently examine a variety of indicators for a large number of variables to support the dimension reduction task. **(3) Transparent relationship analysis:** The reason a variable is included/excluded from a model should be transparent to users by explicitly providing the confounding and correlation information. **(4) Interactive model building:** The system should allow users to interactively build stable multivariate logit models based on dimension reduction, relation analysis, and domain knowledge. **(5) Complete and accurate results:** The system should effectively facilitate users in finding all potential risk factors from a high-dimensional dataset and contain as few false-positives as possible. **(6) Informative reporting:** The system should provide not only the names of risk factors, but also details and meaningful information on confounders and risk factors. **(7) Intuitive visual interface:** Targeting domain experts such as epidemiologists, the visual encoding should carry clear statistical meaning to users. Easy-to-use interactions should be provided to help the domain experts conduct the tasks.

4 The Visual Analytics Approach

We propose a novel visual analytics approach and fully implemented it in a working prototype. Figure 2 shows the visual analytics pipeline supported in our system. It consists of three steps. The output from a step is the input to the next step. Users can return to a previous step at any time to refine their analysis.

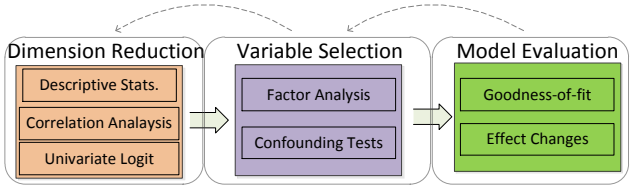

Figure 2: Visual Analytics architecture for risk factors analysis.

4.1 Step 1: Dimension reduction

In this step, users interactively select interesting explanatory variables based on univariate indicators.

4.1.1 Statistical procedures

Our prototype provides the following indicators frequently used in epidemiological analysis. They describe different aspects of the relationship between an explanatory variable and the birth defect variable.

(1) Point Biserial Correlations (rpb): In our system, rpb [32] is used to measure the correlation between a birth defect (measured as a dichotomous variable) and a chemical exposure (measured as a continuous variable). Although correlation does not imply causation, there is a tendency for the two variables to fluctuate in tandem. Rpb is expressed in a range from +1 to -1; values larger/smaller than zero mean positive/negative correlation. The statistical significance of rpb needs to be tested and only correlations whose p-values are smaller than 0.05 are generally considered significant. **(2) Wald test p-value (WaldP):** A WaldP [14] comes from a univariate logit model that contains a chemical exposure variable and a birth defect outcome. It is used for testing the statistical significance of the model. If the p-value is less than or equal to a chosen significance level, the chemical variable is doing much to help explain the birth defect outcome. Generally, the significance level is set as 0.05. **(3) Crude Odds Ratios (OR) and their confidence intervals:** OR is a quantitative measure of the association between a binary explanatory variable X and a binary outcome variable Y. OR and its 95% confidence interval can be calculated from a contingency table for categorical variables [31] or from a logit model for continuous or categorical variables. Higher ORs indicate higher odds of the outcome. The association is significant if the lower/higher bound of the confidence interval is greater/smaller than 1 for a risk/protective factor. OR is crude in this step because it has not been adjusted for other variables in a multivariate model. **(4) Crude ORs based on categorized chemical exposure variables and their confidence intervals:** To explore an association between an exposure and a

birth defect beyond a yes/no (dichotomous) exposure, but not assuming a linear association with a continuous exposure (as in the point-biserial correlation), epidemiologists sometimes categorize the exposure using quartiles or some other quantile based on the reference group. Because a large proportion of the estimated exposures in this dataset were zero, we categorized non-zero exposure values as "low", "medium", or "high" where the number of observations in each group were approximately equal. We mark the odds ratios as OR_L, OR_M, and OR_H for the "low", "medium", and "high" groups, respectively.

4.1.2 Visualization and interactions

Users need to examine the individual indicators as well as the consistency among different indicators (e.g., all positive or negative) for a large number of variables. They also need to select variables of interest according to multiple indicators. Our system supports those tasks using the Univariate Analysis View (UAV) (Figures 1 (A1) and 3). UAV allows users to effectively examine the varying indicators for a large number of variables and select variables of interest flexibly and efficiently.

Since the indicators are measured in different ways, their visual representations are different. To help users judge them intuitively, a consistent color design is used throughout the system: red indicates a statistically significant risk factor, green indicates a statistically significant variable that has been proven to be false-positive or protective, and gray means that the variable is not statistically significant. A variable may turn green from red as the analysis goes further.

As shown in Figure 3, the indicators are displayed in the UAV in a table-like view. Each row represents a variable whose name is displayed on the leftmost cell; each column represents an indicator. With the juxtaposition design [11], it is convenient for users to compare indicators/variables. Inspired by the rank-by-feature framework [28], we allow users to sort the variables according to one or more indicators and then select top ranked variables for further analysis. Following Table Lens [26], we make the inspection intuitive and effective by using visual attributes to represent indicator values.. Horizontal bars encode the rpb values. Zero values, the center of the rpb range, are placed at the center of the cell and marked by a short vertical line. Positive/negative rpb values are represented by a bar on the right/left of the vertical line; the length of the bar represents the absolute rpb value. If the p-value is greater than 0.05, the bar is colored gray. Otherwise it is red/green to indicate statistically significant positive/negative correlation. Significant/non-significant WaldP values are represented by a red/gray dot.

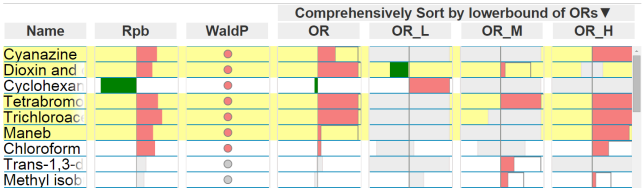

Figure 3: Indicators displayed in the univariate analysis view. The variables selected are highlighted in yellow.

For OR, OR_L, OR_M, and OR_H, a horizontal axis is used in the cell and 1 is marked by a vertical line. If 1 is between the higher and lower bounds, the association is non-significant and we color the rectangle between the lower and higher bounds gray. It is desired to display a larger red/green portion in a cell of a more risky/protective variable. Therefore, we color the rectangle between the lower bound and the vertical line marking 1 red if the lower bound is higher than 1.We color the rectangle between the higher bound and the vertical line green if the higher bound is smaller than 1 (see Figure 3). This encoding makes it possible to rescale the cells because even if the confidence intervals extend beyond the cells, the risk/protective factors are still indicated by cell color (see Figure 3).

This view provides a set of interactions for dimension reduction. Users can filter out variables whose WaldP or rpb p-values are higher than a threshold. They can also sort the variables based on any of the indicators. An interesting interaction called comprehensive sorting is provided. It sorts the variables by the maximum of the lower bounds of OR, OR_L, OR_M, and OR_H. This is a useful interaction since a variable may have a strong association as long as any of the indicators are significant. Users can interactively click a variable to select/unselect it or use shift + click to bulk select. Basic descriptive statistics of the selected variables are displayed in the bottom of the UVA (see Figure 1 (A2)).

Users can send the selected variables to the next step by clicking a button. Since they can be unaware of risky variables correlated to a selected variable, the system will automatically examine if such variables exist and add them to the next step. In particular, confounding tests (see Section 4.2) are conducted for each selected variable. All unselected variables which either confound or are confounded by the selected variable are included into the selection. Real-time interactions are possible because confounding tests for all chemical variables are conducted during pre-processing and the results are stored in a matrix. The system only needs to find the correct matrix location during the interactive selection. Variables selected in the UAV view by a user are marked in the next view by a small yellow block in front of the labels.

4.2 Step 2: Relation analysis for model building

In this step, users inspect the relationships among the variables from step 1 and interactively select variables to build logit models. Information such as correlations among the variables, variable stability, and confounding are automatically analyzed and visually presented to users.

4.2.1 Statistical procedures

Factor Analysis is often conducted to find intercorrelated variable groups, variances explained by each group, and the contribution of each variable in the group. This information can guide users to select variables that contribute more to the variability of the dataset thereby avoiding superfluous variables whose response patterns are caused by their association with an underlying latent variable. In addition, group information is helpful for identifying variables with weak associations whose significant associations can only be observed when other variables in the same group are excluded from a model (see Figure 5). Our system uses varimax rotation to impose a partition where each variable has a large correlation coefficient with the group it belongs to and small correlation coefficients with other groups [18]. Each group defines a "factor". The eigenvalue of a factor reflects its contribution to the total variance in the correlation coefficient matrix.

Confounding tests are conducted as follows: a bi-variate logit model is constructed for the response variable and each pair of explanatory variables. A explanatory variable having a change in OR greater than 10% compared to its univariate logit model OR in any bi-variate logit model is considered a confounder and the pairing variable caused the confounding [13].

4.2.2 Visualization and interactions

In order to select variables for model building, users need to examine variable correlation, inspect Factor Analysis and confounding test results, and study data distributions of the variables. Data distribution is important for domain experts to understand the quality of data collection, inspect levels of exposures, and validate and hypothesize the correlation between variables. Our system provides the Variable Groups View (VGV) for interactive variable selection

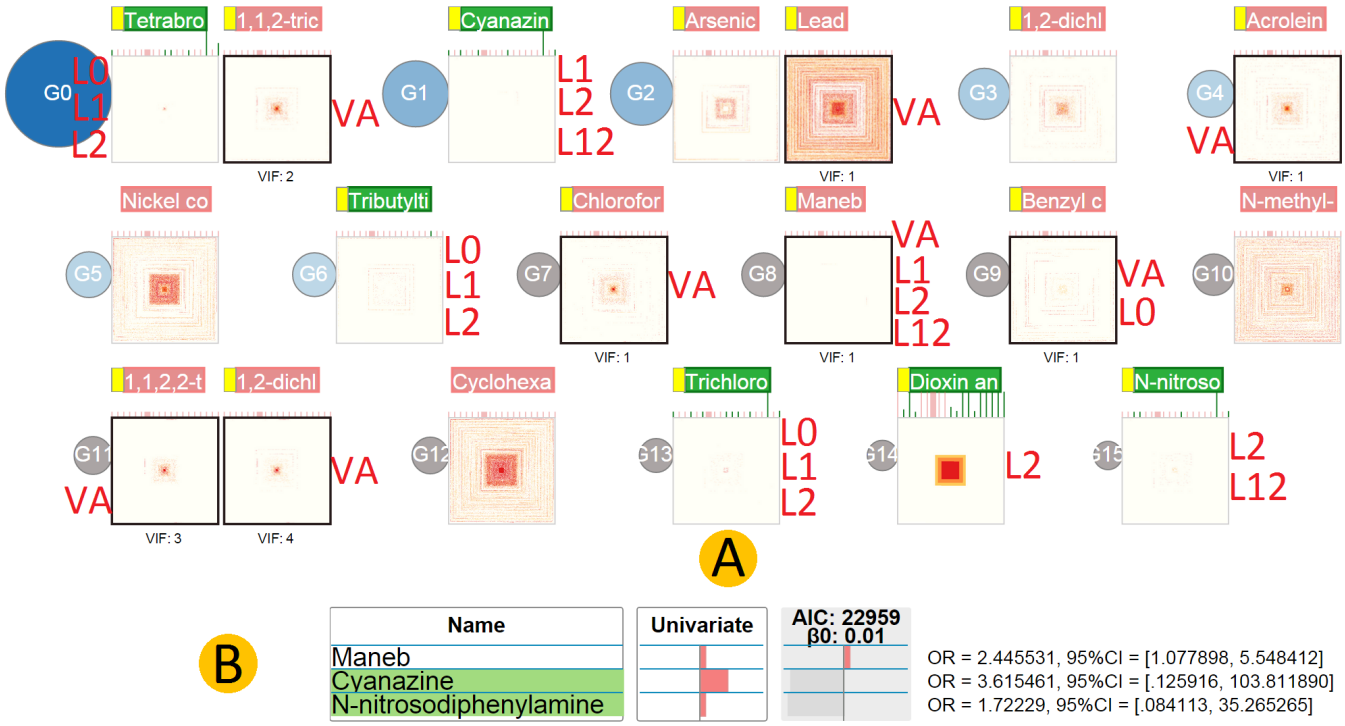

Figure 4: A. The variable groups view. L0, L1, L12, and VA indicate risk factors identified by Forward Stepwise selection, LASSO, Elastic Net, and our visual analytics approach, respectively. B. The result of a model with variables selected by L12. The gray in the rightmost column indicates that the model is not statistically significant. The green background behind the chemical names in the first column indicates the presence of confounding.

which supports the above tasks (see Figure 4). Compared with the UAV, the VGV needs to provide many more details for a smaller number of variables.

Presenting our dataset is challenging since its large proportion of zeroes causes clutter in many visualization techniques. We selected the pixel-oriented technique [19] since it can present datasets with a large proportion of a single value without clutter. In addition, it scales to large, high-dimensional datasets and allows users to inspect variable distributions and dimension correlations at the same time. Other popular techniques such as scatterplot matrices [12] and Parallel Coordinates [17] do not have all of these characteristics.

In Figure 4, each square represents a variable. In a square, each data item is represented by a pixel whose color represents its value for the variable: white means zero; dark, medium, and light red represents high, medium, and low values of the exposure to that chemical (calculated with the same partitioning approach used in the univariate analysis), respectively. The squares can help users examine the distributions of the variables. For example, Figure 4 shows that *lead* is widely distributed while *Cyanazine* is sparsely distributed. A data item has the same position in all the squares. Therefore, correlated variables have similar textures in their squares. We can observe that *1,1,2,2-tetrachloroethane* and *1,2-dichloropropane* are highly related from Figure 4. Users can sort the items by clicking a variable. The sorting will place items with high values for that variable in the center of the display. This allows the correlation between that variable and other variables to be clearly observed. In Figure 4, the records are sorted by *Dioxin and Dioxin-like Compounds*.

Variables are grouped and placed in the VGV based on the Factor Analysis results. To use the space efficiently without clutter, a grid layout is used and the variables are placed on the grid line by line. Variables of the same group are placed adjacent to each other. Different groups are bounded by circles. Groups with larger eigenvalues are considered more important and their circles are bigger and darker. Groups are sorted in descending order by eigenvalues. Within a group, a variable that is more correlated with the group is placed in front of variables that are less correlated.

Selection mode in the VGV allows users to add a variable into a logit model by clicking it. Dynamic tips for Variance Inflation Factor (VIF) [18] are provided during variable selection. It provides an estimation of the severity of multicollinearity in selected variables according to the correlation matrix. In addition, the cumulative variance of selected variables is also displayed at the bottom of the VGV during the variable selection process. In this way users can learn how much variability has been explained by the selected variables (see Figure 1 B).

For the confounding tests, analysts are interested in OR changes of 10% or more in the N-1 (N is the number of variables in the VGV) bi-variate logit models in which a variable participates. Very large OR values are also interesting since they are an extreme case of confounding. We propose a compact grass view attached to each pixel-oriented display for visualizing the variable's OR values and changes in the N-1 bi-variate logit models (see Figure 4). N-1 grass blades (tiny vertical lines) grow on top of each square. Each grass blade represents the result of one logit model. If the OR change is 10% or more, the grass blade and the label are green to indicate that the variable is false-positive (a confounder). Otherwise the grass blade is red. The label of a variable is red if all its OR changes are less than 10% (it is not a confounder). The length of a grass blade represents the OR value. It is normalized among all the variables to enable comparison. Tall grass blades, which mean inflated ORs, can be easily spotted. Since extremely high OR values of the confounders may skew the OR distribution, we set the upper bound of the normalization to be the maximum of 20 and the largest OR.

When users hover a mouse over a grass blade, the blade of the paired variable is highlighted by an increased width. Users can examine the OR value or click the label of that variable to find out whether it confounds other variables (they will be highlighted). By

comparing the textures of the squares, users can learn the correlation among the variables and thus get a deeper understanding of the confounding. Users can also examine the pairing variables one by one through a navigation widget triggered by clicking the grass.

Interactive visualization provides the flexibility of building models using different strategies. For example, users can start by adding non-confounders with distinct pixel textures into the model. If the model is good (the background color of the result column in the Model Evaluation View is not gray), more variables whose group mates are not in the model can be included. In addition, the weak association between the outcome and a variable can be suppressed if its group mates are in the same model. Removing the group mates will allow the model to reveal the weak association (see Figure 5 for an example). Users can also use a full model with all non-confounders, where they only pick one variable from each group to build the model. Later, they can replace any variables with their group mates to check the effect of the group mates. At any time during interactive model building, if the model is not good (indicated by a gray column in the Model Evaluation View), users can remove the variables with unstable estimation from the model to improve its stability.

4.3 Step 3: Model Evaluation and Effect Change Assessment

After users click a button in the VGV, a multivariate logit model will be built with the variables selected in it. The users can interactively examine the results of the model for refinement or for testing other variables. The categorical demographic characteristics can also be added into the model for effect change analysis.

4.3.1 Statistical procedures

The Newton-Raphson technique [36] is used to optimize logistic regression coefficient computations. Their results include ORs and confidence intervals for each variable. If a variable has a change in OR greater than 10% compared to its OR in the univariate logit model, it is considered a confounder. A non-confounder whose lower confidence interval bound is larger than 1 is considered a risk factor. If there are one or more variables with large OR values (such as an OR larger than 20), the model is considered unstable and needs to be improved.

The Likelihood Ratio test [14] is used to measure the statistical significance of the model. The model is significant if the p-value is smaller than 0.05. Akaike's Information Criterion (AIC) is another measure to assess the goodness of fit [14]. The smaller the AIC, the better the model is.

To help users analyze categorical variables, a Chi-square independence test [14] is conducted. It is not interesting to study a categorical variable which has a non-significant association with the birth defect being studied. For each category in a variable, a contingency table is constructed using a reference category assigned by domain experts. ORs and confidence intervals of these groups are calculated to discover groups vulnerable to the birth defect.

4.3.2 Visualization and interactions

The Model Evaluation View (MEV) allows users to examine the results of a set of multivariable logit models. Besides the model consisting of all the variables selected from the VGV, users can interactively build more models with those variables and one or more categorical variables so that the effect change of the categorical variables can be evaluated. The results are presented in a table (see Figure 1 C). Following SAS [27], the first column of the table shows variable names and the other columns record the results of the models. The names of the confounders are highlighted in green (see Figure 4 B). The second column shows the ORs and the confidence intervals resulting from univariate logit models. The other columns show the ORs and the confidence intervals resulting from the multivariate logit models. The ORs and the confidence intervals are displayed using the same visual encoding of the UAV. Their numeric values in the last model are displayed after the last column so users can read the results accurately. Once users have a stable model containing chemicals only, they can add categorical variables into the model to examine their effect change. As shown in Figure 1 C, the color of the categorical variable name indicates their significance in the Chi-square independence test (gray for non-significant variables and red for significant variables). Clicking the name of a categorical variable will trigger its Rose Plot (see Figure 1 C on the left). A Rose Plot visually presents the population sizes, the proportion of cases, and the significance of the ORs associated with each variable category. In particular, it consists of multiple sectors, each of which represents a population group defined by a category. The angle of a sector is the proportional to the number of mothers in this group. The radius is the ratio of cases to the number of mothers in this group. The red/green color indicates this population group has a significant OR with a confidence interval lower bound larger than 1/higher bound smaller than 1 (vulnerable/resistant to the birth defect). The reference group is colored blue and other groups are colored gray. The Rose Plot is compact and allows users to effectively compare populations and risks. It works well for variables with a small number of categories, which is the case in this application.

Users can gradually add the categorical variables into the model by double clicking them. The results are shown column by column. The effect change can be examined by comparing these columns. For example, as shown in the Figure 1 C, column 3, 4, 5, and 6 do not show any big differences in ORs or in the confidence intervals for the chemical *lead*. However, the effect for *lead* becomes non-significant in the last column. The column has an additional categorical variable *region* in the model. This tends to explain that the effect of *lead* varies by *region*.

Weak effects can be uncovered by interactive model building. For example, by looking at the data distribution in the VGV (Figure 4 A), we find three chemicals tightly correlated. They are *1,2-dichloropropane*, *1,1,2,2-tetrachloroethane*, and *1,1,2-trichloroethane*. A weak association between each of them and the birth defect can be identified by the model shown in Figure 5.

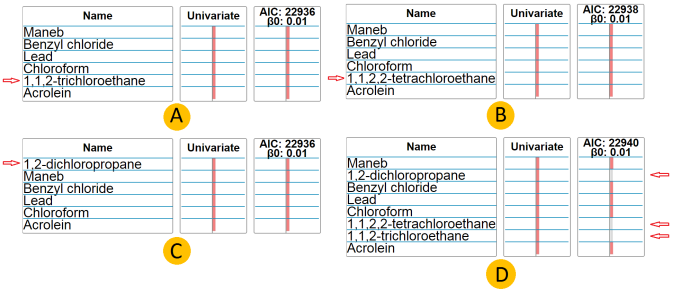

Figure 5: Weak associations. A. *1,1,2-trichloroethane* is included in a model with stable variables. It has a red bar indicating its statistical significance. B and C show the same pattern for *1,1,2,2-tetrachloroethane* and *1,2-dichloropropane*, respectively. D. The three of them are added into the same model. They all have gray bars which mean that they become non-significant.

5 Case Studies

In this section, we first report a case study of identifying risk factors for Limb Reduction Defects (LRD) conducted by the authors. The results are compared with five automatic dimension reduction methods. Then we present a case study on the COIL Challenge 2000 benchmark dataset [34] to characterize caravan policy holders. The study was pair analytics [1] by a senior Ph.D. student in Statistics and a visualization expert.

5.1 Identifying risk factors for Limb Reduction Defects

We conducted a case study for LRD using the environmental birth defect dataset introduced in Section 1. Sixteen variables were selected from the UAV (Figure 1 A1) with a 0.25 WaldP threshold ([5] and [14] suggested this threshold for dimensional reduction). The variables are sent to the VGV (Figure 1 B) together with three other variables correlated to them (the three variables were later proven to be non-risk factors). A robust model was quickly built. It identified five risk factors (the top five rows in Figure 1 C). Then, we further identified three risk factors having weak associations with LRD through interactive model building (Figure 5 illustrates this process).

To compare our approach with existing approaches, we fed the 16 variables selected from the UAV into 4 automatic approaches and compared their results with ours. They were Forward Stepwise selection [7], LASSO [33], Ridge Regression [6], and the Elastic Net [39]. The stats package [25] and the glmnet package [9] in R were used. The penalty parameter lambda was chosen based on cross-validation provided in the package.

The results favored our approach. First, only a small number of identified risk factors were consistent among the automatic approaches. Second, several risk factors identified by one or more automatic approaches were not identified in our case study. We have proved that they are all confounders (see Figure 4 A for an example). Third, our approach identified several risk factors that were not identified by any automatic approaches. Among them, *lead* is a well-studied metal, so we conducted a literature search to find the ground truth. We found several articles [10, 21] suggesting that *lead* levels in hair and blood of mothers are related to LRD, which is consistent with our result that *lead* may be a risk factor for LRD.

5.2 Finding characteristics of caravan policy holders

The COIL 2000 Challenge benchmark dataset contains customer information from an insurance company [34]. It consists of 9,000 data items and 86 variables, including product usage and socio-demographic attributes. This dataset was selected because it represents many domain-specific problems: noisy data, correlated items, redundancy, high-dimensional variables, and weak associations between the explanatory and response variables. The task is to characterize caravan insurance policy holders and provide insights into why customers have the insurance.

The participants were Jean, a senior Ph.D. student in Statistics whose research area is variable selection and regression model building, and Joe, a senior Ph.D. student in visualization who is familiar with our visual analytics prototype. Both of them were not familiar with the dataset before the study. They did not know the semantic meaning of the attributes (the text inside [ ] following a variable name below) until the end of the study.

Jean and Joe explored the dataset side by side in front of a desktop where the dataset was loaded into the prototype. Joe briefly introduced the overall workflow to Jean at the beginning of the study. Jean then took charge of the reasoning process and Joe helped her in manipulating the visual interface and explaining the visual encoding of the data displayed.

First, Jean filtered, sorted, and selected attributes through the UAV. She used 0.25 as the threshold for the rpb P-value and Wald test P-value. Thirty-two attributes were selected and sent to the VGV. Extra attributes were automatically sent to the VGV since they correlate to one or more variables she selected. From the textures of the squares in the VGV, Jean commented that she could tell that the data was redundant and correlated. She found several variable groups, such as a group consisting of MOSTYPE [customer subtype] and MOSHOOFD [customer main type], as well as a group consisting of MGEMOMV [avg. size household 1-6], and MFWEKIND [household with children]. She noticed that MKOOPKLA [purchasing power class] confounded the relationship between the caravan policy and MGEMOMV [avg. size household 1-6]. She also found correlations between attributes after sorting the textures. For example, MKOOPKLA [purchasing power class] had a negative correlation with the group of MOSTYPE [customer subtype] and MOSHOOFD [customer main type]. APERSAUT [number of car policies] and PPERSAUT [contribution car policies] had a positive correlation. After three iterations in the steps of model building, Jean obtained a good model with 21 variables. There were 13 non-significant and 8 significant attributes as the color coding indicated. Excluding 2 confounders indicated by a green background, Jean concluded that 6 significant attributes were likely to describe the characteristics of caravan policy holders. They were *PPERSAUT* [contribution car policies], *MAUT1* [1 car], *MBERMIDD* [middle management], *MOPLHOOG* [high level education], *PBRAND* [contribution fire policies], and *ALEVEN* [number of life insurances]. These attributes were commonly acknowledged in [34]. They were likely to describe a group of rich people with a more expensive car, a high level of education, and fire insurance due to the need to carry gas for cooking in the caravan.

5.3 Expert Feedback

The LRD case study and experiment results were shared with the epidemiologists in a written document. One epidemiologist commented that the models from the automatic approaches were behaving oddly. For example, one model had an OR of around 262 for *Tetrabromobisphenol A*. The epidemiologist said it was extremely unlikely to be valid. He suspected that the inconsistency among the results of the automatic approaches was more likely caused by the high correlation among the chemicals than caused by missing observations, since our data contains 60,613 cases.

The univariate analysis view was demonstrated to the epidemiologists. They were excited about the visualization and commented that it was intuitive and makes their tasks of comparing the indicators much easier. Since they are not experts in high-dimensional logit modeling, they suggested that we consult a statistician for feedback on the multivariate analysis part of the prototype.

We followed their advice and interviewed a statistician through Skype. He has a PhD degree in statistics and currently is a professor and active researcher in the field. He has conducted intensive research on high-dimensional logit modeling. The interview lasted one hour, before which he had read a written document illustrating our approach. We showed him a live demo of the system in the interview. During the demo, we explained the statistical procedures and visualization techniques to him. He validated the statistics procedures we used and made the following comments:

"This is a cool and useful system."

"It allows statisticians to communicate with users much easier. It conveys the modeling process to users in a visible, intuitive, user-friendly way rather than using tedious word descriptions."

"It follows the high-dimensional logit modeling pipeline we use. It nicely integrates a variety of statistical procedures together for effective logit modeling."

6 Conclusion and Future Work

In this paper, we present a novel visual analytics approach to high-dimensional logit modeling for risk factor identification. It integrates a set of useful statistical techniques for dimension reduction and model building into a smooth analysis pipeline. It enhances the analysis process with intuitive visualizations and interactions that allow users to easily compare the results from varying statistical analyses, and allows users to effectively and efficiently conduct dimension reduction and model building iteratively. Case studies have been conducted where the prototype has been used to find potential risk factors for limb reduction defects and characterize the caravan insurance policy holders. The results present the effectiveness and efficiency of our approach. Positive feedback has also

been received from two epidemiologists and a statistician. Our approach is limited to datasets where most variables to be examined are numerical dimensions. In the future, we would like to extend our approach to support analyzing datasets with a large number of categorical variables.

ACKNOWLEDGEMENTS

We thank Tamara Munzner and Daniel A. Keim for feedback on the design study. We thank Jun Li, Jiancheng Jiang, and Weihua Zhou for comments and suggestions on the statistical methodologies. The research reported in this publication was made possible in part by a USEPA grant (#R834790). It was also partly supported through a cooperative agreement (U01dd000494) between the Centers for Disease Control and Prevention (CDC) and the Texas Department of State Health Services (TX DSHS). The contents are solely the responsibility of the authors and do not necessarily represent the official views of the USEPA, the CDC, or the TX DSHS. Further, USEPA and the authors do not endorse the purchase of any commercial products or services mentioned in the publication.

REFERENCES

[1] R. Arias-Hernandez, L. Kaastra, T. Green, and B. Fisher. Pair analytics: Capturing reasoning processes in collaborative visual analytics. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10, Jan 2011.

[2] M. Bögl, W. Aigner, P. Filzmoser, T. Lammarsch, S. Miksch, and A. Rind. Visual analytics for model selection in time series analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2237–2246, Dec 2013.

[3] J. D. Brender, M. U. Shinde, F. B. Zhan, X. Gong, and P. H. Langlois. Maternal residential proximity to chlorinated solvent emissions and birth defects in offspring: a case-control study. *Environmental Health*, 13(1):96, 2014.

[4] R. L. Brent. Environmental causes of human congenital malformations: the pediatricians role in dealing with these complex clinical problems caused by a multiplicity of environmental and genetic factors. *Pediatrics*, 113(Supplement 3):957–968, 2004.

[5] Z. Bursac, C. H. Gauss, D. K. Williams, and D. W. Hosmer. Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, 3(1):17, 2008.

[6] S. L. Cessie and J. C. V. Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1):191–201, 1992.

[7] M. Efroymson. *Multiple regression analysis*. Mathematical Methods for Digital Computers, Wiley, New York, 1960.

[8] S. J. Fernstad, J. Shaw, and J. Johansson. Quality-based guidance for exploratory dimensionality reduction. *Information Visualization*, 12(1):44–64, 2012.

[9] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[10] E. Gilbert-Barness. Teratogenic causes of malformations. *Annals of Clinical & Laboratory Science*, 40(2):99–114, 2010.

[11] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[12] J. A. Hartigan. Printer graphics for clustering. *Journal of Statistical Computation and Simulation*, 4(3):187–213, 1975.

[13] M. A. Hernán, S. Hernández-Díaz, M. M. Werler, and A. A. Mitchell. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American Journal of Epidemiology*, 155(2):176–184, 2002.

[14] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[15] IBM Corp. . *Released 2013. IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY, 2013.

[16] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. DimStiller: Workflows for dimensional analysis and reduction. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 3–10. IEEE, 2010.

[17] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[18] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*. Prenticee Hall, London, 2002.

[19] D. A. Keim. Designing Pixel-Oriented Visualization Techniques: Theory and Applications. *IEEE Transactions on Visualization and Computer Graphics*, 6:59–78, 2000.

[20] J. Krause, A. Perer, and E. Bertini. Infuse: interactive feature selection for predictive modeling of high dimensional data. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1614–1623, 2014.

[21] F. Levine and M. Muenke. VACTERL association with high prenatal lead exposure: similarities to animal models of lead teratogenicity. *Pediatrics*, 87(3):390–392, 1991.

[22] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Ann. Statist.*, 42(2):413–468, 04 2014.

[23] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer. Guiding feature subset selection with an interactive visualization. In *Visual Analytics Science and Technology, 2011 IEEE Conference on*, pages 111–120. IEEE, 2011.

[24] T. Mühlbacher and H. Piringer. A partition-based framework for building and validating regression models. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):1962–1971, 2013.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[26] R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '94, pages 318–322, New York, NY, USA, 1994. ACM.

[27] SAS Institute Inc. *SAS/STAT Software, Version 9.1*. Cary, NC, 2003.

[28] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 65–72, 2004.

[29] G. Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[30] C. Steed, J. Swan, T. Jankun-Kelly, and P. Fitzpatrick. Guided analysis of hurricane trends using statistical processes integrated with interactive parallel coordinates. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 19–26, Oct 2009.

[31] L. M. Sullivan. *Essentials of Biostatistics in Public Health*. Jones & Bartlett Learning, second edition, 2011.

[32] R. F. Tate. The theory of correlation between two continuous variables when one is dichotomized. *Biometrika*, 42(1/2):205–216, 1955.

[33] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[34] P. van der Putten and M. van Someren. CoIL challenge 2000: The insurance company case. Technical report, Leiden Institute of Advanced Computer Science, 2000.

[35] J. Yang, D. Hubball, M. O. Ward, E. A. Rundensteiner, and W. Ribarsky. Value and Relation Display: Interactive Visual Exploration of Large Data Sets with Hundreds of Dimensions. *Visualization and Computer Graphics, IEEE Transactions on*, 13:494–507, 2007.

[36] T. J. Ypma. Historical development of the newton-raphson method. *SIAM Review*, 37(4):531–551, 1995.

[37] F. B. Zhan, D. J. Brender, H. P. Langlois, and J. Yang. Air pollution-exposure-health effect indicators: Mining massive geographically-referenced environmental health data to identify risk factors for birth defects. US Environmental Protection Agency, 2015. Final report (2011-2015, 325 pages).

[38] B. Zou, J. G. Wilson, F. B. Zhan, and Y. Zeng. An emission-weighted proximity model for air pollution exposure assessment. *Science of The Total Environment*, 407(17):4939–4945, 2009.

[39] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.