ABSTRACT
The KDD community has described a multitude of methods for knowledge discovery on large datasets. We consider some of these methods and integrate them into an analyst’s workflow that proceeds from the data-centric descriptive level to the model-centric causal level. Examples of the workflow are shown for the Health Indicators Warehouse, which is a public database for community health information that is a potent resource for conducting data science on a medium scale. We demonstrate the potential of the HIW as a source of serious visual analytics efforts by showing correlation matrix visualizations, multivariate outlier analysis, multiple linear regression of Medicare costs, and scatterplot matrices for a broad set of health indicators. We conclude by sketching the first steps toward a causal dependence hypothesis.

Categories and Subject Descriptors
K.4.1 [Computers and Society]: Computer-related Health Issues

General Terms
Algorithms, Performance.

Keywords
Community health indicators, multivariate statistics, machine learning, visual analytics.

1. INTRODUCTION
Data mining and visual analytics are maturing disciplines that have outlined an impressive array of methods for data analysis of large datasets. Despite the widespread availability of methods and toolkits for analytics, there is little guidance provided for a systematic knowledge discovery workflow that would allow the analyst to progressively gain a deeper understanding of the dataset they are working with. In health informatics this is particularly important, because the goal is discovery of actionable knowledge with the potential to improve health outcomes in the general public.

We wish to create an analyst’s workflow that winnows down the dataset from a large number of dimensions to a manageable number for a more detailed discovery process. The analyst workflow proceeds from descriptive evaluations to causal hypotheses, working in a multivariate data space. At one extreme, the analyst drills down to possibly univariate levels using descriptive statistics or database queries. At the opposite extreme, descriptive roll-ups are performed to provide aggregate views of an entire dataset. High level visualizations often provide critical opportunities for analysts to use their perceptual skills to see globally-distributed relationships among multiple variables – a “zero’th order” discovery process. However, the ultimate goal is usually to understand the underlying abstract structures and processes that generated the “surface structure” of the dataset itself.

Several important online databases exist that provide important sources of data for health informatics. The Health Indicators Warehouse (HIW) is an online public database for community health information with over a thousand indicators that list health indicators at national levels partitioned by demographic axes including ethnicity and sex, and a variety of health condition and activity categorizations. Some indicators are also expressed to show geographic variation at granular levels such as state, county and hospital referral region (HRR). The HIW may be found here: http://healthindicators.gov/. We have employed the HIW in our exploration of a knowledge discovery process and produced several analytic artifacts that illustrate a proposed analyst workflow for such moderately sized datasets. We also do not consider special characteristics of datasets (e.g., sparsity, or significant “type” variations) in this proposal for an analyst workflow.

The remainder of the paper is organized as follows. Section 2 presents the analyst workflow diagrammatically. Sections 3, 4 and 5 respectively present analytical results from the HIW for correlations, multivariate outlier analysis, and predictive analytics for cost drivers of community healthcare. Section 6 presents HIW data in an expanded format called the partial correlation matrix. The first intimation of a causal analysis of HIW data is presented in Section 7. Finally, Section 8 concludes with suggestions for future work.

2. ANALYST WORKFLOW
The workflow resulting from working with a multivariate dataset is depicted in Figure 1. The steps from top to bottom proceed...
After selecting the right indicators for further analysis, the workflow proceeds in two main directions: descriptive statistics or model-based analytics. There are two main descriptive routes: mean structure and correlation (covariance) structure. The scatterplot matrix and correlation matrix express the same data at different levels, and they are combined in the software tool described in Section 3. Model-based analytics are almost always required for high-dimensional multivariate datasets. Predictive methods are also a common tool in the analyst toolkit.

3. CORRELATION STRUCTURES

All pairwise correlations among a large set of indicators may be expressed in a square data matrix, where the diagonals are all set to the value of one. (The Pearson correlation coefficient is an example conventional measure of association between two numerical variables.) An initial but powerful visual analytics step emerges when such a correlation matrix is displayed as a heatmap [1]. We created a correlation matrix heatmap for a set of 150 indicators from the HIW expressed at the hospital referral region level (available online at http://cda.ornl.gov/heat/heatmap.html). The heatmap portrays a global view of the associations among indicators, but also allows the user to drill down to the level of individual scatterplots by panning across and zooming into a local region of the heatmap.

The heatmap, shown in Figure 2, is color-coded on a graded red-blue scale. Extreme negative correlations are colored as saturated blue, whereas high positive correlations show up as saturated red hues. Neutral correlations are displayed along a gray scale with a whitish hue.

Figure 2. Heatmap showing correlation matrix from a set of HIW indicators.

The user may drill down to the scatterplot level and view scatterplots of individual HRR data points as in Figure 3.

Using tools readily available in analytic systems such as R [2], we transform the correlation matrix heatmap such that adjacent row and columns are similar. The reordered heatmap can be used to identify clusters of indicators that possess similar correlation profiles across the entire set of indicators. We note that color pattern perception is a superior human ability worth exploiting for visual analytics. Reordering indicators often leads us to identify the emergence of color patterns and the subsequent identification of candidate latent factors in the clusters. Friendly [3] illustrated the use of reordering techniques for baseball and automobile data; our approach is to use hierarchical clustering as a basis for reordering rows and columns. If the initial order of the indicators is mostly random, a clustering transformation of a random-
appearing color matrix can generate a highly patterned image that delineates regions of high negative and positive correlation.

Figure 3. Zoomed-in view of HRR-level health indicators.

Clustered heatmaps have been used previously to profile gene expression data from micro-arrays (e.g., [4]), but we are not aware of applications to healthcare indicators. As an example of a healthcare application, Figure 4 illustrates a rectangular section of a correlation heatmap containing a set of disease indicators (DI) and Community Health Status Indicators (CHSI) variables in the HIW that has been reordered using a hierarchical clustering method – here we make the point that dendograms along the top and left sides show how we may deduce co-organized structures from such indicator interactions.

Figure 4. Correlation submatrix heatmap shows DI/CHSI indicator interactions.

A few color patterns emerge directly from the induced reordering in Figure 4. One striking region near the bottom center of the matrix contains a group of moderate to intense positive correlation values. The corresponding CHSI cluster includes the last 8 indicators (by row), whereas the DI cluster is composed of 6 indicators near the middle columns, including several admission rates and the percent of beneficiary population using in-patient (IP) services. We interpret this pattern as demonstrating a strong positive association between general population risk factors and beneficiary IP admission rates at the state level. Another mostly blue region at top center of the submatrix reveals generally negative associations between utilization of general population preventive health services and beneficiary post-acute care (PAC) costs and utilizations.

4. MULTIVARIATE OUTLIER ANALYSIS

Univariate outlier analysis is often performed in one or two dimensions using either a box plot (unidimensional) or scatterplot with elliptical boundaries (bidimensional). Bag plots are another distribution-free visual analytic tool to isolate outliers [5]. Summarizing high-dimensional outlier information into a single chi-square distributed score assumes the data are multivariate Normal. We compute the Mahalanobis distance [6] for each data point with respect to the sample mean and covariance matrix thus standardizing the measures. Obviously we would use this measure to group indicators with similar metrics. For example, the Mahalanobis distance is not likely to have much meaning when combining count data and proportions. Sometimes it also makes sense to group variables from the same category, e.g., the 12 proportion indicators from “disease prevalence” in the DI from the HIW. Figure 5 shows the top 13 outlying HRRs with respect to disease prevalence in a chart that we call a sorted bar-heatmap. The heatmap to the right of the horizontal bar chart presents a disease prevalence profile for each HRR. The color scale of the heatmap varies along a continuum from white (0 percent) to reddish-brown (100 percent). Whereas the bar chart sorts HRRs by their distances (from largest to smallest) from the center of the distribution, the color profiles impart information relevant to the reason for the uniqueness of an HRR. Most HRR profiles score high both on probability of diabetes and probability of ischemic heart disease. Honolulu appears to have a different profile than the other top outliers because prevalence appears to be low for nearly all diseases.

Figure 5. Sorted bar-heatmap displays most extreme multivariate outliers.

We also encode a cost attribute in Figure 5 as the degree of saturation of green fill in the horizontal bars. The amount of “greenness” increases with the standardized per capita cost (SPCC) of services in the HRR. The top two outliers, Miami and McAllen Texas, also have the greatest SPCC in the top 13. Honolulu is associated with the lowest costs. A cautionary note on this type of outlier analysis: the Mahalanobis distance method implicitly assumes the normal distribution of data. Since these data are proportions, it may be appropriate to use an arcsine-root transformation before proceeding with the analysis.
This section described the top half of Figure 1 in helping an analyst reduce the large parameter space into a limited set for further sophisticated analysis which we discuss below.

5. PREDICTIVE ANALYTICS
An important aspect of mitigating escalating health care costs, especially in the public sector, is the identification of major cost drivers in each HRR. Comparative analysis across HRRs could generate important recommendations for policy adaptations to control healthcare costs. We address this by aiming to regress cost metrics onto health indicators to enable the determination of those major drivers. Ultimately a distribution-free supervised learning technique should be employed for this purpose to obtain the most accurate predictions, but as a first step we used multiple linear regression to explore the ability of a simple linear combination of indicators to account for SPCC.

Five predictors from the HIW were selected according to the following criteria: (1) high correlation with SPCC; and (2) collective ability to account for a large degree of variance in SPCC. The selected predictors were: (1) probability beneficiary is female; (2) probability beneficiary is African-American; (3) probability beneficiary received post-acute care; (4) probability beneficiary received IP services; and (5) average normalized Hierarchical Condition Category (HCC) score. A plot of predicted/observed SPCC is shown in Figure 6. These five predictors attained an $R^2=0.87$ – a strong fit ($p<.0001$).

There are a few (labeled) regression outliers [7] in Figure 6 that might not be detected from conventional outlier analysis.

6. Comparison across Sources in the HIW
Rarely have datasets from public health sources been studied together with medical history data to allow the knowledge contained in each dataset to co-inform each other. We followed our procedure outlined in Figure 1 and used a scatterplot matrix to display relationships among multiple variables. A few enhancements allow us to improve the utilization of the available real estate in the conventional scatter plot view using the partial scatterplot matrix [8] approach. The authors introduced a mixed partial scatterplot matrix that displays univariate histograms along the diagonal (which typically contains no quantitative information in a conventional scatterplot matrix), and partial regression residuals after adjusting for all other variables along the lower panel. The information in the lower panel represents a step closer to the causal level of description. Figure 7 shows a similar scatterplot matrix for one DI (congestive heart failure admission rate among beneficiaries aged 65-74) and two CHSI indicators (percent of population with high blood pressure and percent of population that does not exercise) from the HIW.
effect of percent of high blood pressure in the population, the association between CHF admission rate among beneficiaries and percent of population that doesn’t exercise is reduced from .622 to .202. This suggests the lack of a direct linkage between these indicators, and that most of the relationship between them can be explained by an intervening indicator: percent of the general population with high blood pressure.

7. Causal-Level Abstraction of CHSI Data

Partial correlations represent an intermediate step between the descriptive and causal levels of analysis. A more complete causal explanation is possible with graphical modeling where nodes represent indicators and factors, and links show the strength and direction of relationships between nodes. For example, structural equation modeling (SEM) [10] and Bayesian networks [11] are two graphical models that explain data at the causal level.

When working with a large set of indicators it is worth asking whether some indicators are simply proxies for more general latent (unobserved) factors. A causal model that incorporates a large set of indicators is unwieldy and difficult to comprehend. It is often easier (and more intuitive) to postulate a smaller set of latent factors that underlie the entire indicator set and then propose causal links between the reduced set of factors. For example, many health care studies have capitalized on the explanatory power of socioeconomic status, which is itself a latent factor. This is a primary purpose of confirmatory factor analysis, itself a part of SEM.

We performed an exploratory factor analysis (EFA) in R using a set of 17 CHSI indicators expressed at the state level in the HIW. We then used the output of EFA to hypothesize a factor model and ran it through a confirmatory factor analysis using the ‘sem’ package in R. Figure 8 reveals the structure of the confirmatory model incorporating the original set of CHSI indicators. Latent factors are represented by ovals, and are named: (1) overall risk, (2) screening, (3) cancer/heart mortality and (4) prevention. (The four current causal links shown with the dashed lines in Figure 8 are hypothesized as negative relationships.)

8. FUTURE WORK

8.1 Expanding the Scope of Analytics

The types of analytics presented in this paper represent only a beginning to the exploration of the full potential of the HIW and similar datasets. We have limited our analysis to a subset of the indicators.

One challenge to full integration of data sources is the level of data expression, which varies among sources. “Crosswalks” are needed to move seamlessly between state, county, zip code, and HRR levels. Some crosswalk tables already exist, but others will need to be developed in order for analysts to pre-process data from different sources into compatible formats and projections. Among the most difficult crosswalks are between levels that have many-many mappings, e.g., county to HRR and HRR to county.

8.2 Visualizing Pairwise Relationships

The correlation coefficient is a basic metric of association, but somewhat limited because it reveals only the strength of a possible linear relationship. Numerical healthcare data, generally speaking, exhibits a much greater degree of diversity than what can be explained by a linear system. Interesting relationships among pairs of indicators can assume many other functional forms such as piecewise linear, parabolic, cubic, exponential, elliptical, sinusoidal and non-coexistent, to name a few. Recently, a maximal information coefficient was defined that identifies a potentially large number of pairwise relationships based on a grid algorithm [12]. We are currently working on a new scheme that generalizes the correlation matrix visualization while incorporating these new relationship types. This should facilitate movement from the top to bottom of the analyst workflow.

9. ACKNOWLEDGMENTS

This manuscript is authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy.

10. REFERENCES


Figure 8. Hypothesized structural equation model for CHSI indicators.

Although significant further analysis is needed to refine the factors and their labels, and to assist in verifying causal relationships among the four factors, we offer this example as an illustration of how we move to a focused causal hypothesis from a seemingly disparate dataset.
http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.ss/1177013622


