Eurographics Conference on Visualization (EuroVis), Posters Track (2017)
A. Puig Puig and T. Isenberg (Editors)

Visual Stratification for Epidemiological Analysis

José Matute† and Lars Linsen ‡

Westfälische Wilhelms-Universität Münster, Münster, Germany

Abstract

Investigating etiology of a disease depends on the combination of tacit medical knowledge and multivariate analysis on a wide array of collected data. Confounding variables may generate a bias when exploring disease determinants, thus, reducing the predictive capabilities of risk factors. Stratified analysis is widely used in epidemiological settings to reduce the effect of confounding factors. We propose a stratified visual analysis approach based on linear projections and interactions in a Star Coordinates Plot (SCP), where the segregation power of dimensions in multiple strata can be explored interactively. We apply our approach to gain insight into three epidemiological results using stratified analysis regarding the prevalence of sleep apnea within age and gender strata and the segregating power of well-defined epidemiological risk factors.

1. Introduction

A wide range of disciplines ranging from biology to social and behavioral sciences contribute to the investigation of determinants of human health and disease prevention [AP05]. Epidemiology is one of such disciplines, which studies the distribution and determinants of disease frequency in man [MP70].

Risk assessment for determinants may be performed through clinical observations or experimental animal exposure. These methods may suffer from strong selection bias or are limited to diseases that have suitable animal models [Gor88]. Epidemiological studies, such as Cohort Studies, do not suffer from these limitations. Cohort Studies allow us to explore the combined effects of multiple factors on disease development in a snapshot of the population. Visual information is intuitive and quickly processed by humans, thus, allowing visual analysis to be used for aiding in the generation of hypotheses and the validation of associations and risks in epidemiological hypotheses.

Hinz et al. [HBS*14] used a combination of time-plots and parallel sets to explore the relationship between hemoglobin levels and time before death for patients diagnosed with Diabetes. Turkay et al. [TLLH13] employed scatterplots to provide a statistical overview of the features and deviation plots, where a sub-group of the population is analyzed and the change in statistical properties is shown in a scatterplot by connecting dimensions with a line before and after the sub-group filtering. Zhang et al. [ZGP15] created Cohort Analysis via Visual Analytics (CAVA) using different types of standard visualization methods coupled with coordinated views whose goal is to find similar patients or at-risk patients within the context of a hypothesis.

From an epidemiological perspective, not only the comparability of the data points should be taken into account. The biological plausibility of the values and their confounding effect should also be considered. Adjustment procedures such as Stratification, where the population is partitioned by selected factors, is often used for correcting bias caused by their confounding effect. Visual stratification allows then for the exploration of subspaces of the data. The previously presented visual approaches allow the exploration of a single stratum.

Yuan et al. [YRWG13] proposed "Dimension Projection Matrix/Tree" a tree-based iterative approach for exploring subspaces. The datasets can be divided into subsets of dimensions or points by interaction in either space. Subsets of dimensions can be selected and a matrix of MDS projections is displayed where the rows and columns are dimension subsets and the projection is calculated from their set union. Given a subset of uniformly weighted dimensions, a variant of a scatterplot matrix based on the projections is also displayed. However, the influence of each individual dimension within each subset cannot be explored. Liu et al. [LWT*14] proposed animated transitions between individual subspaces allowing exploration of its characteristics and its relationship to other subspaces. Given that only the transition between two subspaces is shown, the user may have trouble recalling the relationship between multiple subspaces to each other. Linear projections, such as Star Coordinates Plot (SCP) [Kan01], try to exploit the intrinsic dimensionality of the data by mapping onto a lower-dimensional embedding and allowing interactivity for axes interpretation. We

† e-mail: matutefl@uni-muenster.de
‡ e-mail: linsen@uni-muenster.de

DOI: 10.2312/eurp.20171174

propose a stratified visual analysis approach based on SCP where we interactively explore the segregation power of dimensions in multiple strata.

2. Visual Stratification

SCPs define a linear projection where each dimension is represented by a vector and the dimensional contribution is based on its magnitude and direction. Users can apply rotation and scaling to the dimensional vectors. The combined influence of factors can be studied by rotating the vectors into similar directions while scaling allows for visualizing the effect of a factors upon separability or cohesion. SCP allows for testing the separability of a target and non-target classes, where a target class is defined as having a disease or disease-related outcome.

Stratified analysis is used to examine confounding by partitioning possible confounding factors into different levels or strata. Each stratum is then separatedly analysed. A large difference between stratum-specific and non-stratified results can be seen as evidence of confounding. The confounding effect of any dimension can be explored. In the case of categorical dimensions, each category defines a stratum. In the case of numerical dimensions, the user is prompted to define the ranges of each stratum. The samples are then filtered accordingly and the user is presented with coordinated views of the strata. Given that more than one confounding factor may appear in the data, we allow for multi-level stratification. Each stratum is further partitioned by the newly selected dimension.

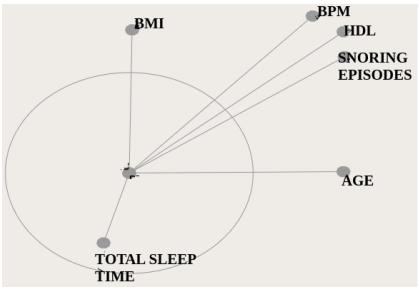

Figure 1: *Configuration for segregating sleep apnea severity. Dimensions are used to validate hypotheses regarding the development of sleep apnea.*

3. Results

As a case study, we use data from the Study of Health in Pomerania (SHIP) [VAS*10] to validate risk factors for central and obstructive sleep apnea. A number of studies have identified risk factors for the development of sleep apnea such as obesity, gender, age, nasal obstruction, and presence of heart conditions [SFP*99, YSP04]. As a target-class we use the Apnea-Hypopnea Index (AHI) which is used to categorize the severity of sleep apnea. AHI values above 15 indicate moderate and above 30 severe apnea. We select previously defined risk factors as our SCP dimensions and attempt to segregate AHI diagnostic levels. Figure 1 shows an SCP configuration for segregating moderate to severe apnea. Figure 2a shows the resulting projection. We can observe a higher concentration of moderate to severe apnea at the upper right region of the projection, yet mild to moderate apnea appears not separated. We apply gender-based stratification as shown in Figure 2b,c and new observations can be made. The configuration is able to separate more clearly mild to severe cases in females than males, which shows that age, BMI, possible heart conditions (defined by higher BPM and high level of cholesterol) and nasal obstructions agree with previously defined hypotheses. We can also observe a higher prevalence of moderate to severe apnea in the male population, which is in agreement with epidemiological results.

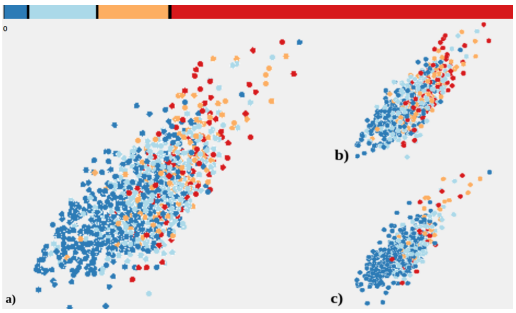

Figure 2: *a) Non-stratified projection of the selected configuration. Gender-based Stratification b) Male c) Female.*

Multiple levels of stratification can be applied to explore the confounding influence of several factors at the same time. Figure 2 displays a two-level stratification analysis given gender and age(from 20 to 80 at 20 years intervals). We are able to observe a higher prevalence of moderate to severe sleep apnea in older population.

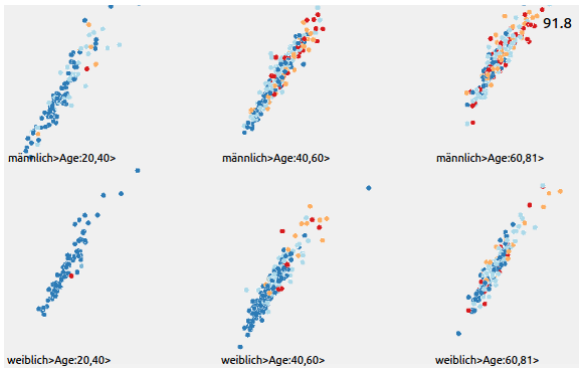

Figure 3: *Multi-level stratification based on gender (rows) and age (columns) from 20 to 80 at 20 years intervals.*

4. Conclusions

Investigating etiology of a disease depends on the combination of tacit medical knowledge and multivariate analysis on large datasets. Stratified analysis is a widely accepted method for studying the confounding effect of factors . We proposed a stratified visual analysis approach based on Star Coordinates Plot where we can explore the segregation power of dimensions in multiple strata can be explored at an interactive rate. We were able to gain insight into three epidemiological results using stratified analysis regarding the prevalence of sleep apnea within age and gender strata and the segregating power of well-defined epidemiological risk factors. Future work involves evaluating the approach through usability tests with epidemiological experts.

Acknowledgments. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under contract LI-23/1.

References

[AP05] AHRENS W., PIGEOT I.: *Handbook of epidemiology*. Springer, 2005. 1

[BKW*15] BAMBERG F., KAUCZOR H.-U., WECKBACH S., SCHLETT C. L., FORSTING M., LADD S. C., GREISER K. H., WEBER M.-A., SCHULZ-MENGER J., NIENDORF T., ET AL.: Whole-body mr imaging in the german national cohort: rationale, design, and technical background. *Radiology 277*, 1 (2015), 206–220.

[Gor88] GORDIS L.: Epidemiology and health risk assessment. In *Epidemiology and health risk assessment*. Oxford University, 1988. 1

[HBS*14] HINZ E., BORLAND D., SHAH H., WEST V. L., HAMMOND W. E.: Temporal visualization of diabetes mellitus via hemoglobin a1c levels. In *Proceedings of the 2014 Workshop on Visual Analytics in Healthcare (VAHC 2014)* (2014). 1

[HBvD*09] HOFMAN A., BRETELER M. M., VAN DUIJN C. M., JANSSEN H. L., KRESTIN G. P., KUIPERS E. J., STRICKER B. H. C., TIEMEIER H., UITTERLINDEN A. G., VINGERLING J. R., ET AL.: The rotterdam study: 2010 objectives and design update. *European journal of epidemiology 24*, 9 (2009), 553–572.

[Kan01] KANDOGAN E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (2001), ACM, pp. 107–116. 1

[LWT*14] LIU S., WANG B., THIAGARAJAN J. J., BREMER P.-T., PASCUCCI V.: Visual exploration of high-dimensional data: Subspace analysis through dynamic projections. *Technical Report UUSCI-2014-003* (2014). 1

[MP70] MACMAHON B., PUGH T. F.: Epidemiology: principles and methods. 1

[SFP*99] SIN D. D., FITZGERALD F., PARKER J. D., NEWTON G., FLORAS J. S., BRADLEY T. D.: Risk factors for central and obstructive sleep apnea in 450 men and women with congestive heart failure. *American journal of respiratory and critical care medicine 160*, 4 (1999), 1101–1106. 2

[TLLH13] TURKAY C., LUNDERVOLD A., LUNDERVOLD A. J., HAUSER H.: Hypothesis generation by interactive visual exploration of heterogeneous medical data. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*. Springer, 2013, pp. 1–12. 1

[VAS*10] VÖLZKE H., ALTE D., SCHMIDT C. O., RADKE D., LORBEER R., FRIEDRICH N., AUMANN N., LAU K., PIONTEK M., BORN G., ET AL.: Cohort profile: the study of health in pomerania. *International journal of epidemiology* (2010), dyp394. 2

[YRWG13] YUAN X., REN D., WANG Z., GUO C.: Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2625–2633. 1

[YSP04] YOUNG T., SKATRUD J., PEPPARD P. E.: Risk factors for obstructive sleep apnea in adults. *Jama 291*, 16 (2004), 2013–2016. 2

[ZGP15] ZHANG Z., GOTZ D., PERER A.: Iterative cohort analysis and exploration. *Information Visualization 14*, 4 (2015), 289–307. 1