Journal Title
XX(X):1–17
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Wiggum: Interactive Visual Analytics for Examining Mix Effects

Chenguang Xu[1], Sarah M Brown[2], Christan Grant[3], and Chris Weaver[4]

Abstract

The importance of data-driven decision-making is rapidly increasing thanks in part to the growing availability and accessibility of data sets and analysis tools. Yet, applicable insight can be difficult due to biases and anomalies in data. An often overlooked phenomenon is *mix effects*, in which subgroups of data exhibit patterns opposite to the data as a whole. This phenomenon is widespread and often leads inexperienced analysts to draw incorrect statistical conclusions. In this paper we present Wiggum, an interactive visual analysis system for uncovering both mix effects and special cases known as *Simpson's paradox*. A Python-based web implementation of Wiggum lets users interactively analyze multidimensional data sets to reveal various forms of mix effects. Through use cases, we describe how Wiggum supports the examination of mix effects in three real data sets and demonstrate how a combination of visualization techniques—heatmaps, trend plots, small multiples, coordinated multiple views, and dynamic queries for multi-attribute drill-down—are effective for analyzing mix effects. We conducted a user study to evaluate Wiggum, focusing on users' ability to comprehend the statistical concepts, identify the corresponding visual patterns, and perform common analysis tasks correctly and efficiently. We discuss usability issues, utility limitations, and outline future directions to improve Wiggum.

Keywords

Visual analytics, human-in-the-loop, mix effects, Simpson's paradox

Introduction

Developing a deep understanding of data is an essential part of decision-making processes. It often takes substantial time and effort to develop enough understanding to make well-informed decisions. Data analysts often perform statistical analyses to develop such understanding. A variety of challenging statistical phenomena can manifest in data sets. This is especially common for large, multi-attribute data sets. An inability to examine them, or even mere ignorance of the phenomena themselves, can restrict analysis and lead to incorrect conclusions[1,2].

Data analysts often use visualizations to explore data sets and examine trends in them. For purposes of discussion in this paper, we define a *trend* as a tendency of a variable in relation to another variable that is changing in the data set. We focus on two trend types. A linear relationship between two variables defines a *regression* (short for *linear regression*) *trend*. A ranking of groups defined by one variable according to a statistic calculated on a second variable defines a *rank trend*. Both types of trends can be identified in a whole data set or in a *subgroup* created by grouping data items under a certain condition. A *mix effect*[3] occurs when a trend reverses in subgroups relative to the aggregate trend.

A prominent case of mix effects arose in an audit of gender bias in graduate school admissions at the University of California, Berkeley[1]. University officials observed that men were admitted at a higher rate than women university wide; however, upon further inspection, this was not true for admissions to individual departments. If *all* departments had shown the reverse trend—they did not—it would have been a special case of mix effects known as *Simpson's paradox* in which all subgroups partitioned by a certain condition present the opposite pattern as the aggregate data set.

The main motivation stems from the substantial challenges inherent in conducting statistical analysis, given the laborious nature of tasks like scrutinizing combinatorial dimensions within the data. An integrated automated system with visualization capabilities could provide a more efficient way to navigate and analyze multidimensional data, from obtaining an overview to scanning the characteristics across multiple dimensions. Data visualization, in particular, makes it possible for a diverse spectrum of users, ranging from researchers to lay audiences, to efficiently and effectively identify interesting patterns[4]. Wiggum represents a punctuation point in our work to develop visual analysis tools for a broad community of users concerned with paradoxes and other complex anomalies in multidimensional data sets. This work aims to examine mix effects and address critical challenges through an overall analytic process. We pursue four main visual analysis goals (**AG1–AG4**)

[1]Oklahoma City University, Oklahoma City, OK, USA
[2]The University of Rhode Island, Kingston, RI, USA
[3]The University of Florida, Gainesville, FL, USA
[4]The University of Oklahoma, Norman, OK, USA

Corresponding author:
Chenguang Xu, Oklahoma City University, Oklahoma City, OK, USA.
Email: shine.xu@okcu.edu

for Wiggum to confront these challenges and tackle the unresolved questions: How can we identify and examine mix effects across various trend types in multidimensional data using an integrated visualization system, and how can users understand such complex statistical concepts through this visualization approach? Although Wiggum primarily focuses on foraging activities, these goals represent the overarching objectives of the analytic process, which is organized into the foraging loop and the sensemaking loop[5], as well as other conceptual models of the analytical reasoning process[6–9].

AG1 Flexibility of Dimensions: Checking a multidimensional data set for mix effects can be a challenging task because it requires examination of relationships between subsets of dimensions in terms of the partitions of data in those dimensions on subgrouping criteria. Wiggum provides features to flexibly drill-down into dimensions and specify trend type and criteria.

AG2 Flexibility of Perspectives: Visual identification and characterization of mix effects can be challenging because the interpretation of trend strengths against the underlying data distribution varies with the trend type and the data types of dimensions considered. Wiggum generalizes the exploration process across different combinations of mix effects and dimension types while also providing visualizations suitable for examining each one.

AG3 Integrability: Exploring mix effects in multidimensional data can be difficult because of the need to bring together independent tools for tasks including selection of dimensions, definition of subgroups, and examination of trends. Wiggum integrates visualizations for examining mix effects with a user interface for loading data, browsing dimensions, specifying groupings, and selecting groups to be examined.

AG4 Understandability: Designing usable visualization tools can be challenging because users differ in their understanding of statistics concepts and how those concepts manifest in data, especially when high dimensional data is transformed along a pipeline for visualization. A participant-based evaluation of Wiggum sheds light on how well it operates with respect to the regression trend of mixed effects in high-dimensional data.

In addition to the general development goals, we set a preliminary goal to evaluate the participants' understanding of mixed effects and associated concepts through a knowledge assessment. Assessing users' basic understanding of statistical concepts, in turn, helps us evaluate AG4.

Taken together, Wiggum's features support interactive visual identification and examination of mix effects. While automated approaches to find mix effect candidates exist, one must select and examine them to interpret their character, strength, and meaning. Automated methods can thus complement but not supplant manual identification. For the current version of Wiggum, we focus on supporting examination of candidates detected by *some* means, and leave addition of automated methods as future work.

To help data analysts better identify and interpret mix effects, Wiggum includes data annotation features to specify variable types and roles. It also offers data augmentation features to generate clusters, quantiles, or intersectional subgroups[10,11].

We introduce *trend strength* and *trend distance* metrics to support comparison of aggregate and subgroup trends. A *distance heatmap view* provides an overview for efficient exploration of trend measures and lets users inspect patterns of relationships across dimensions for defined subgroups. A second view (specific to trend type) shows the details of the selected trend for examination and interpretation. Interactive filtering and ranking features allow users to efficiently organize trends for exploratory analyses.

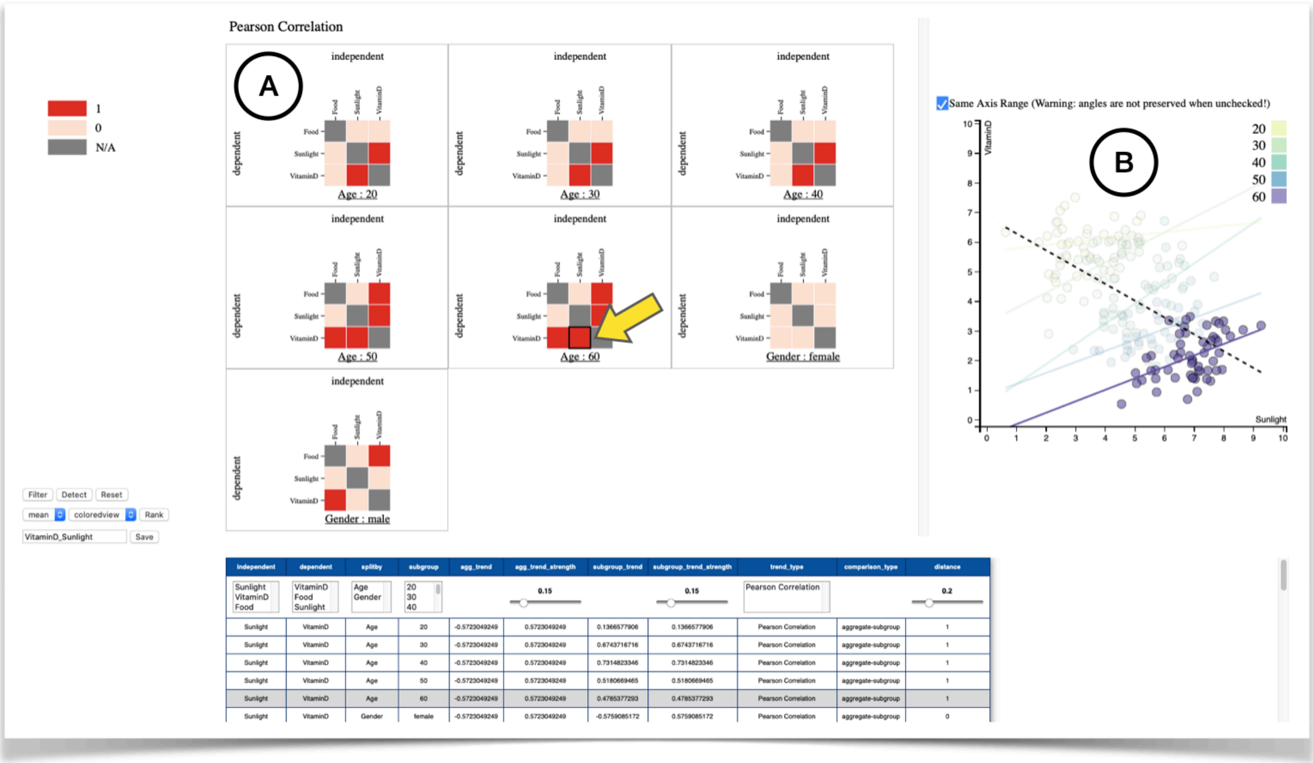
We present use cases of exploring mix effects in three well-known, real-world data sets. The third use case considers a data set of 32,561 records with eight data attributes to demonstrate the practical effectiveness of our system. We conducted a user study to assess the utility of Wiggum's features and usability of its user interface design. Our contributions are as follows:

- a **visual analysis system** that supports interactive exploration to examine patterns that reveal mix effects;
- a **processing pipeline** to map data, annotated and augmented by the user, into statistical results for display;
- **mathematical equations** to formalize metrics of mix effects;
- two new **view designs**, adapted from heatmaps and trend plots, to understand and efficiently examine mix effects;
- a comprehensive **approach** to help users examine *multiple* trend types and explore *different* trend types simultaneously;
- an **evaluation** to assess Wiggum's usability and utility.

We start with an illustrative example of mix effects examination and a review of related work. We then describe how Wiggum prepares and processes data for visual analysis of mix effects. Next, we outline key goals to support visual analysis and the design of interaction visualization features in Wiggum to achieve them. After that, we present use cases of Wiggum followed by the user study. Finally, we discuss limitations and future directions before concluding. To aid the reader, we added an appendix to clearly define key terms used throughout the article.

Illustrative Example

We illustrate how Wiggum supports visual data exploration of a synthetic data set relating vitamin D level, sunlight exposure level (sunlight), food level in vitamin D (food), age, and gender. (Key concepts are italicized throughout, and will be described more formally in later sections.) The target user for Wiggum is diverse and not confined to any specific domain. It is designed to support users from a wide variety of fields who wish to incorporate more sophisticated statistical analysis into their work. Wiggum's target users possess basic statistical knowledge such as Pearson's correlation. These users have a specific interest in exploring and analyzing the differences in statistical results between aggregate data and subgroups. They see value in conducting exploratory



independent	dependent	splitby	subgroup	agg_trend	agg_trend_strength	subgroup_trend	subgroup_trend_strength	trend_type	comparison_type	distance
Sunlight VitaminD Food	VitaminD Food Sunlight	Age Gender	20 30 40		0.15		0.15	Pearson Correlation		0.2
Sunlight	VitaminD	Age	20	-0.5723049249	0.5723049249	0.1366577906	0.1366577906	Pearson Correlation	aggregate-subgroup	1
Sunlight	VitaminD	Age	30	-0.5723049249	0.5723049249	0.6743716716	0.6743716716	Pearson Correlation	aggregate-subgroup	1
Sunlight	VitaminD	Age	40	-0.5723049249	0.5723049249	0.7314823346	0.7314823346	Pearson Correlation	aggregate-subgroup	1
Sunlight	VitaminD	Age	50	-0.5723049249	0.5723049249	0.5180669465	0.5180669465	Pearson Correlation	aggregate-subgroup	1
Sunlight	VitaminD	Age	60	-0.5723049249	0.5723049249	0.4785377293	0.4785377293	Pearson Correlation	aggregate-subgroup	1
Sunlight	VitaminD	Gender	female	-0.5723049249	0.5723049249	-0.5759085172	0.5759085172	Pearson Correlation	aggregate-subgroup	0

Figure 1. Selecting a subgroup and examining trends on the Wiggum visualization page. After selecting the age 60 group and clicking a dark red cell in its distance heatmap (A), the detail view (B) displays the trend reversal between the aggregate data and the age subgroups in a scatter plot.

data analysis to uncover patterns, correlations, and variations within their data.

After loading the data set, we annotate the data type and role of each variable on the data configuration page. To study correlations with vitamin D level, we choose the *Pearson correlation* trend type then click the *Visualize Trends* button. On the visualization page, we observe seven 3×3 distance heatmaps (Fig. 1A). Each distance heatmap represents one subgroup of either the age group or the gender group. Each cell color encodes the pairwise attribute *trend distance* that measures the discrepancy between the aggregate and subgroup trends. A dark red cell encodes a distance of 1, indicating a trend reversal between the aggregate and subgroup.

Clicking the *VitaminD x Sunlight* cell in the heatmap for the age 60 subgroup, the detail view (Fig. 1B) displays a scatter plot of the age subgroups. Each subgroup and its Pearson correlation trend are color encoded with the age 60 subgroup highlighted in solid purple. Its regression line in the plot indicates a positive relationship between vitamin D level and sunlight exposure level. In contrast, the dashed black regression line indicates a negative relationship between them in the aggregate data set. This reversal pattern matches the meaning of the selected dark red cell in the heatmap. The regression lines for the other age groups are similarly reversed from the aggregate data set, although to different degrees.

Mix effects and Simpson's paradox can be expressed mathematically. A trend distance d is defined in Equation (1), using a distance function, dist, and a binary relationship, trend_b, for some summary statistic, stat, of two variables x_1 and x_2. For a single value of $x_3 = y$ (y is a value of the variable x_3, e.g., $age = 60$), the value of d is the *trend distance* between a *subgroup trend* y (i.e., age 60) and the aggregate data. We define a *subgroup trend* as a trend between two variables x_1 and x_2 given $x_3 = y$. In the synthetic data, the stat is the Pearson correlation between vitamin D level (x_1) and sunlight exposure level (x_2) conditioned on age (x_3). The trend_b is positive or negative according to the direction of the correlation. A distance of 0 indicates the trends have the same direction, and 1 indicates a reversal. In this case, the trend between vitamin D level and sunlight exposure level is negative for the aggregate data but positive for the age 60 subgroup. Since the two trends are reversed, the distance between them is 1.

$$d_{x_1,x_2,y} = \text{dist}(\text{trend}_\text{b}(\text{stat}(x_1,x_2)), \text{trend}_\text{b}(\text{stat}(x_1,x_2|x_3 = y))) \tag{1}$$

Mix effects occur when reversal ($d = 1$) happens for *some* values of x_3. If reversal occurs for *all* values of x_3, it is a special case known as *Simpson's paradox*. The synthetic data exhibits both Simpson's paradox and mix effects. The correlation between vitamin D level and food illustrates mix effects, since only two out of five age subgroups (50 and 60) have the reverse trend. The correlation between vitamin D level and sunlight exemplifies Simpson's paradox, since all five age subgroups have the reverse trend. Ignoring these phenomena could lead to incorrect conclusions and decision making.

Background and Related Work

This section reviews research most relevant to the development of Wiggum, namely mix effects, Simpson's paradox, and existing visualization and visual analysis techniques for exploring them.

Mix Effects and Simpson's Paradox

A central objective of mix effects examination is to search for diversity of trends between aggregate data and its partitions, and account for that diversity in one's analysis. Undetected cases of mix effects can lead an analyst to draw incorrect conclusions. The phenomenon has many names, but this most popular one is attributed to Blyth, crediting one of the earliest formal accounts of the paradox [12]. Two well-known examples are gender disparities in university admission rates that reverse at the department level (the Berkeley Admissions Example [1]), and the average income increasing over time but decreasing at every level of education [13].

In the Berkeley Admission Example, the relative rank of genders by admission rate is flipped. We refer to this form of mix effects as a *rank trend* mix effects. Other examples of this type include tuberculosis mortality rates by race [14] and the on-time departure rate of flights by airline and airport [15]. Mix effects can also be found in sports data sets [16,17] used to rank players.

In the average income example, the summary statistic is correlation between time (x_1 in Equation (1)) and income (x_2) conditioned on educational attainment (x_3), the trend$_b$ is positive or negative, and the distance is a 0/1 loss. By considering the direction of the correlation, this form relies on assuming a linear relationship between the variables, and is referred to as a *regression trend* mix effects. Other examples of this form include the positive relationship between coffee and neuroticism by gender [18], and between petal width and sepal by species of iris [19]. Wiggum is designed to support regression and rank analysis in data sets of similar size and complexity, but works for larger and more complex data sets too.

As a special example of mix effects, Simpson's paradox is of significant interest and often concern in many disciplines, including epidemiology [20,21], social science [22–24], psychology [18], and sports analytics [16,25]. In one famous example, researchers quantified the success rate of a medical treatment [26]. An error arising from the paradox was not discovered until years later [27]. The paradox has also been studied in its impact on association rules [28]. In the statistics community, it is a well-known phenomenon widely understood through causal explanations [29,30]. The *HypDB* [31] system was developed to detect, explain, and resolve bias leading to statistical anomalies like Simpson's paradox. Detecting bias in *HypDB* happens via a causal DAG (Directed Acyclic Graph) that can be difficult to create and interpret. Wiggum's interactive visualization approach aims to help users comprehend and apply statistical concepts for analysis without requiring considerations of causality.

Visual Exploration of Mix Effects

In visualization research, mix effects' impact on causality has been studied with respect to reliability of correlation analysis [32,33]. In causal inference, a confounding variable that predicts both treatment and outcome can be closely linked to mix effects [34]. Existing visualization techniques generally fall into two categories based on how trends, including rank trend and regression trend, manifest in data. Tabular techniques usually focus on rank trend mix effects [1,25,35–37]. Scatterplots are more appropriate when the purpose of the visualization is to present regression trends of two variables [18,22,23,38,39]. A line graph that consists of only two points on the x-axis has also been proposed to illustrate mix effects visually [38]. Causal network visualizations [40–44] offer additional means to explore and interpret mix effects.

Efforts to visualize mix effects demonstrate how detection can be included in visual exploration environments [45]. *Vizdom* allows users to observe instances of mix effects by using multiple bar charts to compare aggregate and partition trends [46]. A grouped bar chart can be used to explain mix effects by observing trend reversal in each subgroup [15]. The *comet chart* [3] supports detection of mix effects, approaching the problem with an explicit goal to explain the phenomenon mathematically. However, it does not address effectiveness for exploring multidimensional data sets, is inherently unsuited to exploring more than two groups—such as race in the adult income example—and doesn't support regression trend mix effects. Wiggum supports flexible multidimensional anomaly detection in terms of trend types and offers more general trend comparison, including degree of difference in trends, with detail visualizations to examine both data and trends.

Building on static visualization approaches [3,18], interactive approaches hold clear promise to facilitate the mix effects detection process [45]. While automated methods of detection can facilitate discovery [19], their black-box nature may not be suitable for sense-making. The relatively new area of auditing fairness in machine learning [11,47–49] brings visual analytics techniques into the domain of interpretable machine learning [50,51], concerned with understanding the results of ML-based decisions. In Wiggum, we apply interactive visualization techniques to create a human-in-the-loop design that improves interpretability as a part of sense-making by giving analysts access to and control over the examination process.

Data Preparation and Processing

In this section, we describe the processing pipeline that Wiggum applies to map a data set into the statistical information displayed in its visualizations. The process begins with a data preparation phase, in which the user provides metadata and applies any preprocessing needed for their investigation. Next, Wiggum extracts features and partitions the data as inputs to trend generation. Finally, it calculates trend measures for each subgroup.

Data Preparation

Data preparation happens in two stages, both under interactive user control. In *data annotation*, the user labels the data attributes of the loaded data set with the role they should play in trend calculations. In *data augmentation*, the user can define additional categorical variables for use in partitioning data.

Data Annotation Wiggum needs type and role information for each data dimension (variable) to determine how to apply rank and regression trend calculations. Upon loading a data

Table 1. User-defined types and roles of features in instances of mix effects, for each of the two types of trends.

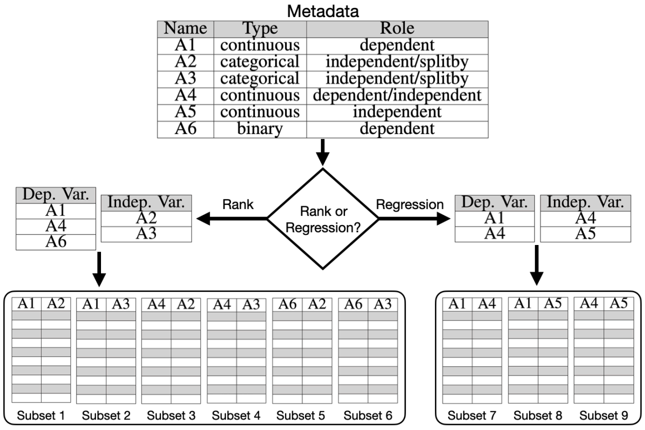
Feature in Eq. 1	Rank		Regression	
	Type	Role	Type	Role
x_1	binary/continuous	dependent	continuous/ordinal	dependent
x_2	categorical	independent	continuous/ordinal	independent
x_3	any	splitby	any	splitby

set, Wiggum applies an automatic type-mapping function to infer a type for each variable. The user can edit the type to be binary, categorical, continuous, or ordinal. They can also specify each variable's role to be *dependent*, *independent*, or *splitby*.

A *feature* is a variable that has been cast into a type and put into a role for the purpose of calculating trends, as shown in Table 1. An instance of mix effects involves three features. The first two features, one dependent and one independent, are used to compute a trend. For a rank trend, an aggregation (e.g., average) is applied to values of the *dependent* feature. The *independent* feature partitions a population into ranked groups (e.g., {Women, Men} for *Gender*) based on the aggregation result. For example, in the admissions data set, admission rate is calculated as the mean of a *dependent* feature (*Accept Status* as a binary type). The rank by admission rate of the *independent* feature (*Gender* as a categorical type) determines the rank trend. For a regression trend, Wiggum models the relationship between the features (details in Section 4.2). For both types of trend, a third *splitby* feature groups rows by the other features' values.

A drop-down list lets users select multiple roles for a variable in a data set, making it possible to have multiple features in which the same variable plays different roles. Users may want to consider only a subset of variables, especially for high-dimensional data sets. Wiggum provides an *ignore* role option for users to exclude variables that are not of interest. (Note: For the rest of the paper we use the terms *variable* and *feature* interchangeably.)

Data Augmentation For many data sets and problem definitions, existing categorical variables can be used as *splitby* attributes for partitioning. Wiggum also lets users define calculated categorical variables. The values of each calculated variable populate an additional column in the data. Wiggum currently supports calculations to discretize the values of a quantitative attribute into quantiles. Users can also create intersectional combinations of other categorical variables. For example, a user can select features "race" and "gender" to generate a new categorical column called "race_gender" containing values such as "Asian male" from "Asian" and "male". Creation of intersectional variables allows users to examine reverse trends in more complex groupings. Wiggum also lets users define categories via clustering using the Dirichlet Process Gaussian Mixture Model implemented in the Scikit-learn library[52]. The user is not required to specify a number of clusters. We set the maximum number of mixture components to 20, which limits the number of clusters to 20. The user can interact with a range slider to adjust cluster quality and tune the resulting clusters.



Metadata

Name	Type	Role
A1	continuous	dependent
A2	categorical	independent/splitby
A3	categorical	independent/splitby
A4	continuous	dependent/independent
A5	continuous	independent
A6	binary	dependent

Figure 2. Subset generation. For each user-annotated feature (A1–A6), Wiggum generates dependent and independent feature lists appropriate to the user-specified trend type. Generated subsets consist of all possible non-same pairings between dependent and independent lists.

Trend Extraction

Wiggum selects subsets of the available features and partitions them into subgroups. It then generates rank and regression trends based on variable types and roles.

Feature Subset and Subgroup Creation A *subset* is a pair of features along with their respective columns populated with the data values of each feature's variable cast into the feature's type. Wiggum uses the user's entered data annotations to calculate all subsets of the entire data set that are relevant to the selected trend type (see Figure 2).

In a typical rank trend analysis, users are interested in ranking different groups on a particular statistic. We define a dependent variable as a data column whose values are used to calculate such a statistic. In Equation (1), the dependent and independent variables correspond to features x_1 and x_2, respectively. Wiggum processes the data to find all column pairings that contain one dependent variable and one independent variable. The total number of subsets detected for rank trend mix effects is the number of dependent variables times the number of independent variables. To create *subgroups* for rank trend, Wiggum iterates over the list of splitby variables and partitions each subset on the unique values of each splitby variable. If a splitby variable is also in a subset as an independent variable, it will be skipped and will not be used to partition the subset.

In a regression trend analysis, the idea is to model the relationship between a pair of continuous or ordinal variables. Variables specified as *continuous* or *ordinal* type in a *dependent* or *independent* role are used for subset generation. The subsets are all possible pairs of dependent and independent variables, excluding pairings of a variable with itself. Variable pairs are treated as x_1 and x_2 in Equation (1) to examine instances of regression trend mix effects. For M dependent variables, N independent variables, and P same variable pairs, the total number of subsets for regression trend examination is $M \times N - P$. Unlike rank trend, there is no conflict between *splitby* variables and *independent* variables, so all of the *splitby* variables can be used to partition subsets.

In practice, a data variable may play multiple roles in a data set. For instance, feature *A4* in Figure 2 is annotated as playing both *dependent* and *independent* roles. In subset generation, *A4* is paired both with continuous/dependent feature *A1* in Subset 7 and with continuous/independent variable *A5* in Subset 9. Both subsets are used for regression trend examination.

Generating Trends Given subgroups of different subsets, the next (vital) step is to compute trends for each subgroup in each subset. Trend generation applies the same method to both the aggregate data and subgroups. For rank trends, Wiggum computes a summary statistic, such as mean or median, on X_{Dep} (e.g., for the admitted variable: 1 admit, 0 do not admit) for every group value $x_i \in X_{Indep} = \{x_1,...,x_n\}$ (e.g., gender = {Men, Women}). The next step is to rank the n groups by the summary statistic. The trend is an ordered list of ranked groups. In the Berkeley Admissions Example, the trend for aggregate data can be denoted as $t_a = [Women, Men]$ indicating that the admission rate is higher for men than for women.

For a regression trend, we first consider correlation for a subset. Wiggum applies Pearson correlation on the pair of columns in the subset. We denote the *trend* as $t = corr(X_{Dep}, X_{Indep})$, where X_{Dep} is the variable of the *dependent* role and X_{Indep} is the variable of the *independent* role. A difference in the signs of the correlations for the subgroup and aggregate data indicates a reverse trend.

Despite being a powerful method for examining mix effects, correlation alone does not suffice to determine the difference between trends without trend reversal. To allow users to track changes in trends even without reversal, the slope of a linear regression line is used as the trend value. Then $t = b$ where b is derived in the equation of a linear regression line $X_{Dep} = b \cdot X_{Indep} + a$.

Trend Measurement

Wiggum introduces a practical set of measures to support discovering mix effects. We define two basic measures: *trend strength* and *trend distance*. These metrics allow users to identify their specific area of interest within trends, detect trend reversals, and assess disparities between trends. The metrics we choose for this study, Pearson's correlation and Kendall's tau, are commonly used in practice[53]. They are suitable for measuring the association in the two prevalent trend types of mix effects; i.e., linear regression trend and rank trend[54,55]. Because Wiggum is implemented as a modular framework that allows extension and modification of its data processing and visualization components, it would be straightforward to add other types of trends and metrics. For example, if users want to study non-linear trends, other metrics such as Spearman's correlation could be readily integrated into Wiggum.

Trend Strength Trend strength is a metric for assessing how well a trend represents data being examined. For a rank trend, given a subgroup trend t, the strength $s(t)$ indicates how the subgroup trend list and the element-wise sorted list are dissimilar to each other after repeating each element of the subgroup trend in accordance with its representation in the data. We denote the extended subgroup trend list as L_s. Wiggum generates an element-wise sorted list denoted as L_a by sorting on the X_{Dep} column of the subset, then creates a list of the elements in X_{Indep}. The strength is computed using the absolute value of Kendall's tau similarity between the two lists L_a and L_s. We formulate the strength as:

$$s(t) = |\tau(L_a, L_s)| = \left| \frac{P-Q}{\sqrt{(P+Q+T)(P+Q+U)}} \right| \tag{2}$$

in which P is the number of concordant pairs, Q is the number of discordant pairs, T is the number of ties only in L_a, and U is the number of ties only in L_s. For example, if $L_a = [M,F]$ and $L_s = [F,M]$, then $P = 0, Q = 2, T = 0, U = 0$ and $s(t) = 1$. For a regression trend, the absolute correlation indicates the strength of the association of the two attributes in the observed subgroup. We denote the trend strength for a regression trend as:

$$s(t) = |\operatorname{corr}(X_{Dep}, X_{Indep})| \tag{3}$$

which quantifies the strength of the relationship between dependent and independent variables to measure how well the regression trend fits the data.

Trend Distance Distance is a fundamental concept in Wiggum. Given an aggregate trend t_a and a subgroup trend t_s, a pairwise distance $d(t_a, t_s)$ indicates the discrepancy between two trends. To make multiple instances in different types of trends consistent and comparable, the distances in Wiggum are normalized to $[0,1]$; 0 is considered "the exact same trend" and 1 "the largest possible difference". For a rank trend we use Kendall's Tau similarity between the aggregate and subgroup trends, since they are represented as lists. We formulate the normalized distance for a rank trend as:

$$d_\tau(t_a, t_s) = 1 - \frac{(\tau(t_a, t_s) + 1)}{2}. \tag{4}$$

For regression trends, we use two different strategies of examination to increase the flexibility of the trend comparison. When users only care if the subgroup exhibits the opposite pattern of the aggregate data, Wiggum checks the sign of the correlation. The distance for a subgroup trend is defined as:

$$d_\oplus(t_a, t_s) = \operatorname{sign}(t_a) \oplus \operatorname{sign}(t_s) \tag{5}$$

in which *sign* refers to mapping the correlation by its sign to 0 or 1 (1 for positive, 0 for negative) and $\oplus$ is the logic operation for exclusive-OR. Sign on its own is useful for focusing qualitatively on *whether* trends are reversed.

Wiggum users can also focus quantitatively on *how much* trends differ by exploring a measured difference between trends, such as to identify when a trend is a recurring phenomenon. We use the slope as the trend value, with distance defined as the normalized angle between two linear regression lines, as:

$$d_\measuredangle(t_a, t_s) = \frac{2}{\pi}\left(\left|\tan^{-1}(t_a) - \tan^{-1}(t_s)\right| \% \frac{\pi}{2}\right) \tag{6}$$

in which % is the modulo operator, t_a is the slope for the aggregate data, and t_s is the slope for the subgroup data. The normalization step generates a distance $d_\measuredangle \in [0,1]$ where $d_\measuredangle = 1$ indicates a right angle and $d_\measuredangle = 0$ indicates parallel lines.

Wiggum: Examining Mix Effects

Wiggum is a web application consisting of a data preparation page and a visualization page, as shown in Figure 3. In the data preparation page, the user can pick trend types, label metadata, and perform data augmentation. Wiggum is motivated by the Visual Information Seeking Mantra: "Overview first, zoom and filter, then details-on-demand"[56]. The visualization page is made up of a distance heatmap view collection for overview, a trend plot and result table for detail, and a control panel area to support zoom and filter operations.

Foraging Goals

Wiggum was designed to be an interactive visual system for exploring mix effects in multidimensional data. We established a set of foraging goals to guide design and evaluation. These goals focus on seeking information, identifying relationships, and detecting patterns. They aim to interpret the patterns through the data and understand that data through its patterns as a foraging activity in an overall visual analysis process.

FG1 **Provide easy generation of variables for partitioning data.** Although some data sets have categorical or ordinal variables which can be used for partitioning, it is often useful to provide users with a set of approaches to generate subgroups without direct application to existing categorical or ordinal variables. Users should be able to specify the inputs for such data augmentation and quickly generate variables for partitioning (**AG3**).

FG2 **Support multiple trend types.** The system should support examination of multiple, statistically important trend types (e.g., rank trend and regression trend). The visualization design should be general enough to reveal the statistical characteristics for different trend types (**AG2**).

FG3 **Present an overview of the trend distance.** The visual design should clearly show the trend distances generated by the algorithm in a way that lets users explore them quickly to identify and analyze interesting patterns (**AG1**).

FG4 **Facilitate the interpretation of exploration results.** The system should help users understand trend measurements to discern how a subgroup trend reverses compared to the aggregate trend. Users should be able to visualize all subgroups' trends to examine whether mix effects or Simpson's paradox exists in the data set (**AG4**).

FG5 **Allow flexible selection of subgroup trends.** Since users may have domain knowledge about important feature attributes or subgroups to check, users should be able to select records in the result table using quick, simple interactions (**AG3**).

Preparation Page

On the preparation page, Wiggum lets users load data from a new or previously saved CSV file. Data annotation and augmentation are integrated into the same page. Wiggum allows users to select variables to generate quantiles or make intersectional subgroups. It also supports Dirichlet Process Gaussian Mixture Model clustering over pairs of continuous variables by setting a cluster quality threshold (**FG1**). Wiggum equips a list of available trend types such as Pearson correlation, linear regression, and rank trend in a selection box to allow users to select one or multiple options at a time (**FG2**). Upon clicking the Visualize Trends button, Wiggum redirects users to the visualization web page that includes the views discussed below.

Result Table View

The output of the data processing pipeline is a result table (Figure 3B) for each trend-level comparison. After executing the statistical computation in the Wiggum back-end, the user is provided with a result table view containing information on *dependent*, *independent*, and *splitby* variables, subgroups, and trend types. Trend, trend strength, and trend distance for subgroup and aggregate, respectively, are also included in the result table for each detected subgroup trend. Each record in the result table represents a subgroup trend from a feature subset. If there are m subsets (see Section 4.2.1), n splitby variables, and the ith *splitby* variable has k_i values, then the total number of rows in the result table will be $m \times \sum_{i=1}^{n} k_i$. In this view, users can find detailed statistics for a trend in an observational subgroup (**FG4**). The result table is a basic component containing all information needed to generate the distance heatmap view. To further investigate subgroup trends, the area below the table header provides interactive selection boxes and sliders for use in conjunction with the Filter and Detect buttons.

Distance Heatmap View

The distance heatmap view provides an overview of the subgroup trends from the result table (**FG3**). The view consists of multiple distance matrix heatmaps. We designed the distance heatmap view to show the measured distances between different paired features for each subgroup for each trend type. The different trend types generate heatmaps separately. For example, if there are n *splitby* variables used for partitioning in a data set and the ith *splitby* variable has k_i values, then there are $\sum_{i=1}^{n} k_i$ separate heatmaps for each trend type. If there are T trend types that are selected by users, the total number of heatmaps will be $T \times \sum_{i=1}^{n} k_i$. To use the central space more efficiently, Wiggum lays out three heatmaps per row with vertical scrolling. Although automated approaches can provides rankings for instances of mix effects, analysts may want to explore relationships between instances or delve into the attribute pairings with more subtle cases. Consequently, the distance heatmap view emerges as a compelling solution to address these requirements.

The first step in generating a distance heatmap is to select rows from a subset of the result table which contains *dependent*, *independent*, *splitby*, subgroup, trend type, and distance information. For example, in the table in Figure 4, each row represents a subgroup trend in the Auto MPG data set. The selection iterates over trend type, *splitby*, and subgroup; each subset has the same values of those three things. The second step is to reshape the subset data into a two-dimensional matrix, as illustrated in Figure 4B. The

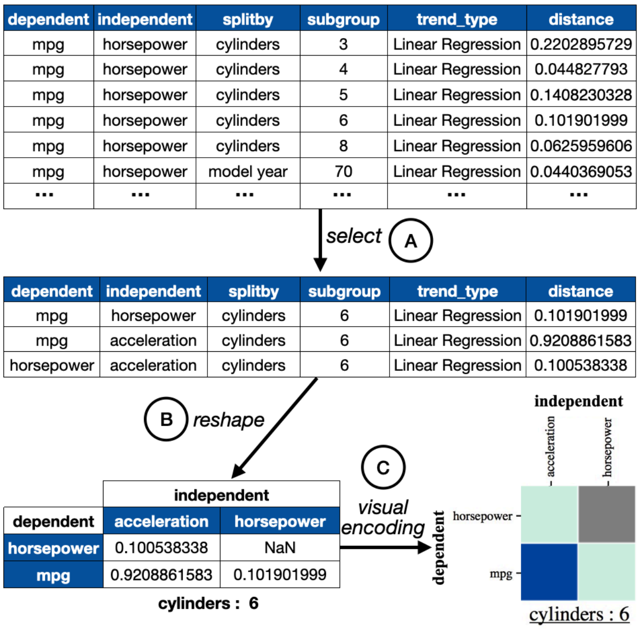
Figure 3. The Auto MPG data set[57] in Wiggum. (A) Control panel for filtering, detecting, and ranking potential instances of mix effects. (B) Result table, for examining statistical details of each subgroup trend. (C) Scrollable array with a distance heatmap for each result subgroup trend. (D) Legend of heatmap cell colors. (E) Trend plot for exploring details of regression trends.

matrix's values are the distance values, and missing values will be set to not-a-number (*NaN*). Rows show *dependent* values and columns show *independent* values. Each distance matrix heatmap's size is the product of the number of unique values in the *dependent* variable and the number of unique values in the *independent* variable for the current choice of subgroup and trend type. For example, as the subset table after selection in Figure 4 shows, there are two unique values of *dependent* variable (mpg and horsepower) and two unique values of *independent* variable (horsepower and acceleration), giving the distance matrix a size of 2 × 2. The corresponding distance matrix heatmap (Figure 4C) is a visual representation of the cells of that distance matrix with each cell representing a subgroup trend. The cells of the matrix are color-encoded to show distance values. Subgroup trends with higher distance are more saturated. In addition, a dark/light red color encodes a binary distance (i.e., 1 for a reverse trend and 0 for a same trend) for a Pearson correlation trend type. Grey indicates a cell with an absent (*NaN*) distance value.

Detail Views

Wiggum includes detail views to help users investigate and understand relationships between aggregate and subgroup trends (**FG4**). For each trend type, we choose different visual techniques to help explain various situations in the data. We describe the detail views corresponding to the two basic trend types, **rank trend** and **regression trend**, as follows.

pictures/2021_overview.pdf



dependent	independent	splitby	subgroup	trend_type	distance
mpg	horsepower	cylinders	3	Linear Regression	0.2202895729
mpg	horsepower	cylinders	4	Linear Regression	0.044827793
mpg	horsepower	cylinders	5	Linear Regression	0.1408230328
mpg	horsepower	cylinders	6	Linear Regression	0.101901999
mpg	horsepower	cylinders	8	Linear Regression	0.0625959606
mpg	horsepower	model year	70	Linear Regression	0.0440369053
...	...	...	...	...	...

dependent	independent	splitby	subgroup	trend_type	distance
mpg	horsepower	cylinders	6	Linear Regression	0.101901999
mpg	acceleration	cylinders	6	Linear Regression	0.9208861583
horsepower	acceleration	cylinders	6	Linear Regression	0.100538338

	independent	
dependent	acceleration	horsepower
horsepower	0.100538338	NaN
mpg	0.9208861583	0.101901999

cylinders : 6

Figure 4. Internal pipeline to generate a distance matrix heatmap from a result table. (A) Records with the same trend type, splitby, and subgroup are selected to build a subset table. (B) Distance values are mapped into dependent variable rows and independent variable columns to form a distance matrix. (C) The matrix is visually encoded as a distance heatmap.

The detail view for **rank trend** contains two sub-views: a grouped bar chart and a parallel coordinate plot. As shown in Figure 5, when users click a cell in the distance matrix heatmap, the cell is highlighted with a solid black border, and the detail view provides details on rates and counts for aggregate data and subgroups. A grouped bar chart at the top displays the counts of records for all combinations of the values of the independent variable and the splitby variable. Bars are grouped by position above the axis for the aggregate and each subgroup. Bars are color-encoded to represent each ranked group in the independent variable. The parallel coordinate plot allows comparison of the aggregate and subgroups to identify trend differences among them. The detailed information for the aggregate is always plotted on the leftmost axis, with the user-selected subgroup on the second axis. After users click a cell for another subgroup, the order of the vertical axis adjusts correspondingly. The order of the remaining axes follows the subgroup order in the distance heatmap view. To better satisfy the need to observe the change of rates, each axis has the same scale. The colors of the connecting lines represent the ranked group.

A trend plot—a scatterplot with trend enhancements—is a useful tool for observing relationships in bivariate data, and supports interpretation of the correlation coefficient or a linear regression model of a **regression trend**. As shown in Figure 6, the user selects a subgroup trend by clicking a cell in the distance matrix. Wiggum gets two columns of the data from the column names, which are indicated by the cell's *dependent* and *independent* values, then plots all the data points in the trend plot. The vertical axis represents the *dependent* variable, and the horizontal axis represents the *independent* variable. Wiggum uses the *splitby* variable to color points in the trend plot. Regression lines for the selected subgroup and aggregate are plotted for users to observe their trends. A dotted black line represents the aggregate trend, and each colored line indicates the corresponding subgroup trend. The slope of the regression line indicates the direction of the relationship between dependent and independent variables. The positive slope of the aggregate trend and negative slope of the subgroup trend in Figure 6 indicates a trend reversal. The degree of the slope allows users to track changes in trends even without a trend reversal.

At the top of the trend plot, Wiggum provides a checkbox for users to switch between two different axis range settings: same axis range and different axis ranges. The same axis range setting helps users more accurately observe slope and the angle between two regression lines. In contrast, angle and slope are not well-preserved when different axis ranges are used. It is sometimes preferable to use different axis ranges to have a finer view of the data when two variables have different ranges. In addition, the trend plot provides an interactive legend that indicates the subgroups in the *splitby* variable. Clicking legend cells selects subgroups and highlights their data.

The Control Panel Area

Wiggum provides a control panel to support efficient exploration of mix effects. The control panel allows **filtering**, **detecting**, **ranking**, **saving**, and **resetting** to support this process.

Users can **filter** the result table via the combo box at the top of the result table and the Filter button in the control panel (**FG5**). Each combo box displays all unique values from the corresponding column. The result table only keeps the rows with the same values based on the combo boxes' selections. Users can select multiple values from the combo boxes in multiple columns (i.e., *dependent*, *independent*, *splitby*, subgroup, and trend type) at the same time. The distance heatmap view is updated based on the filtered result table. A Reset button restores the original rows of the table.

To start **detecting**, users can adjust the thresholds for subgroup trend strength, aggregate trend strength, and distance using the sliders in the result table. The table and heatmaps update correspondingly. Wiggum suggests default values for those thresholds. When users click the Detect button, the thresholds are sent to a filtering method in the data flow pipeline. By adding filtering by strength and distance, users can discover interesting patterns. For example, a distance threshold setting of 1 for a regression trend detects all reverse subgroup trends, since the distance method returns 1 for a reverse trend and 0 for a same trend.

Furthermore, Wiggum supports a comprehensive **ranking** function for users to examine the detected results. Wiggum provides three levels of ranking, as follows.

Viewpoint-level ranking. We define a viewpoint as a set of subgroup trends having the same *dependent* and *independent* variables. We refer to a pair of variables (x_1, x_2) that are used for computing trends as a viewpoint of the data, considering that two variables allow us to plot the data. To go deeper into the data, the next level of ranking is based on summary statistics by grouping the subgroup trends in the same viewpoint. Each viewpoint is ranked by the viewpoint aggregate score.

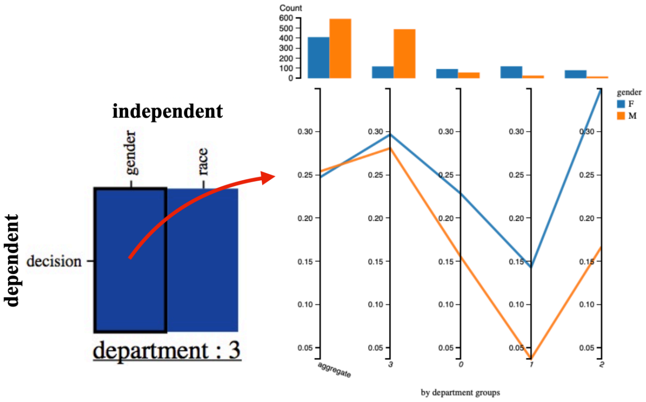


Figure 5. A screenshot of a distance heatmap and its detail view. The detail view shows a rank trend. When the user clicks a decision by gender cell in the distance heatmap, the parallel coordinates plot shows a line for each gender through decision rates on each axis to represent a rank subgroup trend in the department 3 subgroup. A grouped bar chart provides details about the counts of each subgroup.

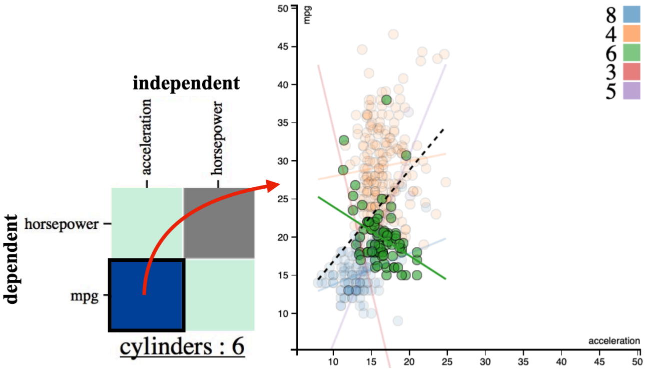


Figure 6. A screenshot of a distance heatmap and its detail view. The detail view shows a regression trend. When the user clicks the mpg by acceleration cell in the distance heatmap, the trend plot shows the relationship between mpg and acceleration to represent a regression subgroup trend in the cylinders 6 subgroup. The legend in the trend plot is a list of the cylinder numbers.

Colored Viewpoint-level ranking. In this level, we define a colored viewpoint as a set of subgroup trends having the same *dependent*, *independent*, and *splitby* variables. We refer to a 3-tuple of variables (x_1, x_2, x_3) that are used for computing subgroup trends as a *colored viewpoint* of the data, with encoding of the *splitby* variable (x_3) as a point color in the plot. The colored viewpoint-level measure of ranking can be obtained by aggregating over all subgroup trends within the same colored viewpoint. Users can choose which aggregation method (sum, mean, max, min) on the distance column to use to compute a colored viewpoint score. The options for colored viewpoint and viewpoint ranking are shown by clicking the select button. The colored viewpoint aggregate scores are added to the result table and used to rank each colored viewpoint.

Subgroup-trend-level ranking. In the analyzing stage, users are shown a result table comprising potential reverse subgroup trends. Each row represents a subgroup trend. Each subgroup trend has four elements (x_1, x_2, x_3, y), in which x_1 represents the *dependent* variable, x_2 is the *independent* variable, x_3 is the *splitby* variable, and y represents subgroup value. A user can rank the subgroup trends by clicking the header of the distance column in the result table, in which case the ranks of all the subgroup trends are based on the results of a descending sort on distance in the result table.

Wiggum provides users with a button for saving the original data, the metadata, and the result table. In the preparation page, users can load prior saved files into Wiggum from a chosen folder. This approach can avoid repeating the process of data labeling, and save time by reusing a result table instead of computing it again in the data flow pipeline. Wiggum also allows users to retrieve the initial result by clicking a Reset button. After resetting is triggered, the result table is set to the original result table without any filtering, detecting, or ranking. The distance heatmap view is redrawn from the original result table.

Implementation

The Wiggum system is a web app. It uses D3[58] for visualization and a Flask[59] server to connect the Wiggum Python library to JavaScript-powered visual analytics in

Table 2. Summary of the three use cases.

Case	Domain	Type	#Records	#Attributes
1	Auto MPG	Regression	392	9
2	Graduate Admissions	Binary Rank	12	4
3	Adult's Income	Multiple Rank	32,561	8

the browser. The Python Wiggum library includes all of the computational features and runs as a back-end server to the Wiggum app. Wiggum has been designed as a modular framework, allowing each individual component to be modified or extended.

Use Cases

In this section, we describe how Wiggum can be applied in practice to two real-world data sets from the UCI Machine Learning repository[57] and a famous example of Simpson's paradox[1]. The first use case shows how Wiggum can be applied to examine trend reversal in partitioned data for linear regression analysis of multidimensional data. The second use case explores a binary ranking of rates which ranks two groups through a known example of Simpson's paradox from the UC Berkeley admissions data set. A third scenario highlights how Wiggum can be generalized to more advanced usage for examining mix effects relative to trends based on both binary and multiple ranking cases. Multiple ranking analysis considers rank trends with respect to the rank of three or more groups. Table 2 summarizes the domain, trend type, number of the records, and number of attributes for each use case.

Regression Analysis

In the first use case, we demonstrate how Wiggum can be used to examine mix effects in a regression trend type data set. We apply Wiggum to the Auto MPG data set. After removing rows with missing data, 392 records each represent a car model with 9 attributes: *mpg*, *cylinders*, *displacement*, *horsepower*, *weight*, *acceleration*, *model year*, *origin*, and *car name*. Data preparation begins after loading the data. The *mpg* attribute is set to *continuous* type and *dependent* role. The *horsepower* attribute is set to *continuous* type and role as both *dependent* and *independent*. The *acceleration* attribute is set to *continuous* type and *independent* role. The other (discrete, multi-valued) attributes are set to *categorical* type and *splitby* role. The *displacement*, *weight*, and *car name* attributes' roles are set to *ignore* to exclude them from trend computation in regression calculations. To study the relationship between two continuous variables in the data, we also set the trend type to *linear regression*.

Upon clicking the *Visualize Trends* button, Wiggum switches to the visualization page. The result table view shows 63 rows, one for each combination of 21 subgroups and 3 *splitby* variables. The distance heatmap view displays a heatmap for each of the 21 subgroups and chosen *splitby* variable. Each heatmap shows a 2x2 matrix corresponding to the chosen two *dependent* variables and two *independent* variables. Inspecting the distance heatmap view immediately reveals some surprising patterns. Two of the 13 heatmaps for model year subgroups contain dark blue cells located in the same cell position, indicating a relationship between *mpg* and *acceleration*. Clicking the dark blue cell in the heatmap of model year 75, we see a positive relationship between *mpg* and *acceleration* for all data points. One might conclude that higher *mpg* corresponds to higher *acceleration*, yet this inference is not supported in the dis-aggregated data. A negative relationship is obvious in model years 75 and 79. The inconsistency of patterns between aggregate data and subgroups may lead a user to hesitate to draw conclusions. Furthermore, we observe an interesting pattern in the heatmap for the 6 cylinders subgroup, as shown in Figure 3, in which the relationship between *mpg* and *acceleration* shows a trend reversal. This finding could influence the decision making process for a car consumer who considers a high MPG and high acceleration car. Once they recognize that MPG will decrease as acceleration increases for the cars with 6 cylinders, other cylinder models tend to be more attractive to them.

There are more surprising trend reversals in the Auto data set that we can identify and study through Wiggum. Although space precludes describing more of them here, the example above is evidence of Wiggum's capability for both examining and explaining mix effects in linear regression analysis.

Binary Ranking Analysis

We illustrate how Wiggum can support visual data exploration through the gender bias case in the UC Berkeley graduate admissions data set. A study of graduate admissions to the University of California, Berkeley was conducted to investigate gender bias[1]. The binary ranking analysis is based on the ranking of two ranked groups, *men* and *women*. The admission rates for Fall 1973 showed that 44% of men and 35% of women were accepted overall, yet in the data for the six largest departments, four departments show a lower acceptance rate for men than women.

The admissions data contains 12 records with 4 attributes: *department*, *gender*, *number of applicants*, and *rate of admission*. After loading the data set, we annotated data types and roles for each variable in the data configuration page. To study the ranking of gender in admission rate, we chose the *rank trend* type, then clicked the *Visualize Trends* button. In the visualization page, each of the six 1×1 distance heatmaps (Fig. 7A) represents one department. The dark blue cell indicates a high trend distance between the aggregate data and the subgroup data. After clicking a cell in the distance heatmap for Department A, the detail view (Fig. 7B) displays the corresponding grouped bar chart and parallel coordinate plot for examining rank trend type. In the parallel coordinate plot, the first axis shows men have a higher admission rate than women in the aggregate data (F:35% vs. M:44%). Nevertheless, the second axis reveals a reverse trend in which the admission rate for women exceeds that of men in Department A (M:62% vs. F:82%). Since the relative rank of the genders by admission rate is flipped (from [*women*, *men*] to [*men*, *women*]), the clicked cell in the distance heatmap is color-encoded dark blue. The parallel coordinate plot represents the trend distance by visualizing the rankings for gender on each axis. The grouped bar chart shows the gender distribution in aggregate and in each department. The Berkeley admission case clearly

Figure 7. The visualization page for the UC Berkeley Graduate Admissions data set. After clicking a dark blue cell in the distance heatmap (A) for department A, the detail view (B) displays the corresponding grouped bar chart and parallel coordinate plot for examining the trend reversal between the aggregate data and the subgroups.

illustrates mix effects, since four out of six departments have the reverse trend.

Multiple Ranking Analysis

The adult data set[57] contains 32,561 records with 8 attributes: *workclass*, *education*, *marital-status*, *occupation*, *relationship*, *race*, *sex*, and *income*. A multiple ranking analysis can be applied to the ranking based on any attribute (e.g., race) with more than two values (e.g., *White*, *Black*, *Asian-Pac-Islander*). In this use case, we aim to find surprising trends in the rate of people making an income of more than 50K per year. We again load the data file and select variable types and roles. All attributes except *income* are set to *categorical* type and both *independent* and *splitby* roles. Thus, those 7 attributes are either the *independent* or *splitby* variable of a subgroup trend. The *income* attribute is a binary attribute containing two values: *>50K* and *<=50K*. We set its type to *binary* and role to *dependent*, making it the *dependent* variable of a subgroup trend. Since we use this data set to study the trend based on the rate of adults' income exceeding 50K per year, we set the trend type to *rank trend*.

On the visualization page (Figure 8), Wiggum displays 9 of 60 distance heatmaps, each representing a subgroup. Each heatmap has one row (for the *dependent* variable) and six columns (for the *independent* variables), with the *splitby* variables used to partition the data. In the result table view, there are 360 records (one *dependent* × six *independent* × 60 subgroups). Each one represents a subgroup trend corresponding to a cell in the heatmaps.

The blue cells in the distance heatmap view are mostly related to race and sex. Focusing on the darkest ones, we adjust the distance slider to 0.9 and both strength sliders to 0. Clicking the *detect* button leaves only three subgroup trends in the result table. Meanwhile, we see three 1 × 1 distance matrix heatmaps in the overview area. Several interesting observations arise upon checking the detail view of the three subgroup trends. For instance, in the *Married-civ-spouse* subgroup, we can observe a reverse trend in terms of *sex*. The trend in the aggregate data is that the rate of over 50K annually is higher for males (30.6%) than females (10.9%). After dividing the data set based on *marital-status*, we see there is a reverse trend with females (45.5%) at a higher rate than males (44.6%) in the *Married-civ-spouse* subgroup.

pictures/2021_example_berkeley.pdf

In addition to a binary ranking, we observe another interesting case in the detail view, in which the trends are seen among different racial groups based on the rate of over 50K income. We select *race* in the *independent* menu selection and pick *occupation* in the *splitby* menu selection, then click the *Filter* button. Looking within racial groups, the rate of income over 50K among *Asian-Pac-Islander* is 26.6%, followed by *White* (25.6%), *Black* (12.4%), *Amer-Indian-Eskimo* (11.6%) and *Other* (9.2%). However, *Asian-Pac-Islander* is not always the highest income earning race. Clicking the cell belonging to the *Exec-managerial* subgroup partitioned by the *occupation* variable reveals that *White* experiences the highest rate (49.9%) of income over 50K, followed by *Asian-Pac-Islander* (45.2%), *Black* (34.4%), *Other* (18.2%), and *Amer-Indian-Eskimo* (10%). Similarly interesting patterns exist and can be readily explored for each occupation's subgroups.

This example demonstrates the capability to explore trends in both binary and multiple ranking cases. A more comprehensive understanding of the data emerges from observing how surprising ranking trends of subgroups are inconsistent with aggregate ranking trends.

Evaluation

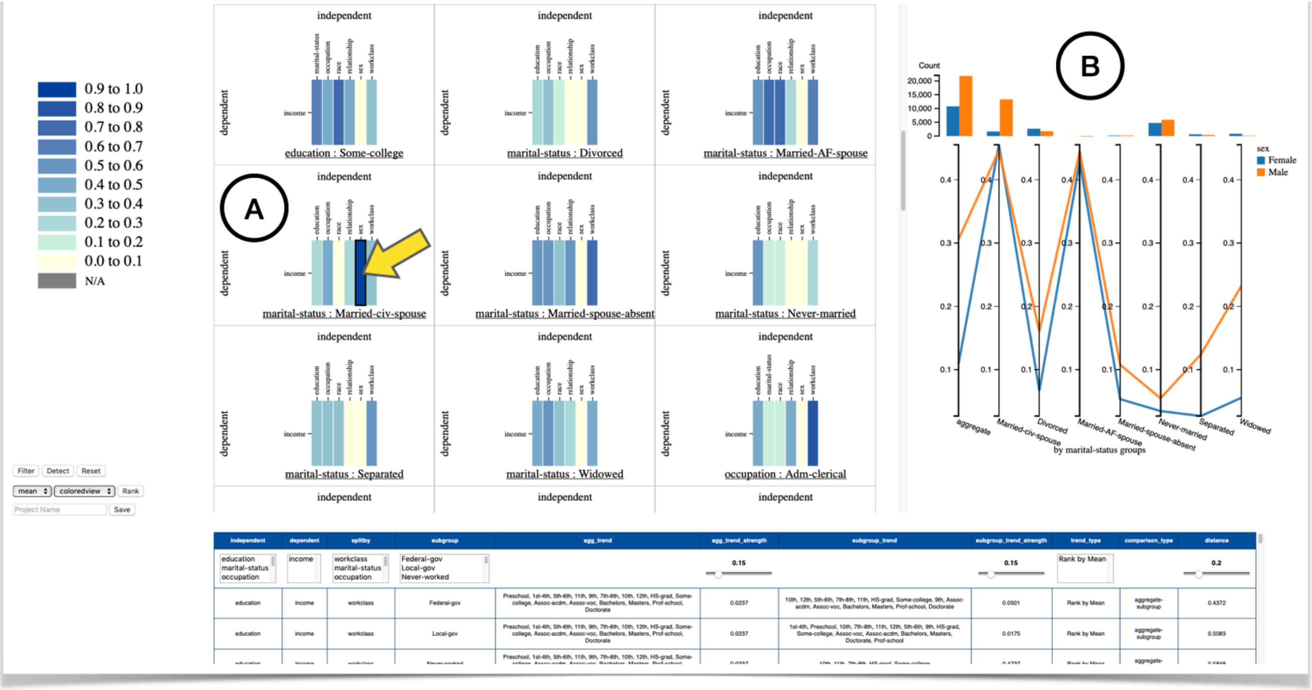
We conducted a formal user study to evaluate how well Wiggum helps target users examine and understand mix effects (**AG4**), and also to solicit ideas for enhancement motivated by their individual data analysis wants and needs. In this study we focused specifically on examination of reverse regression trends using Pearson correlation.

Participants

We did a pilot study with 2 participants—one library staff member with a biology background, one undergraduate in computer engineering—in order to test feasibility, duration, and refine the study design. For the main study, we recruited 37 other participants: 14 Ph.D. students, 8 faculty members, 8 M.S. students, 4 staff members (3 library, 1 senior research associate), 2 postdocs, and 1 undergraduate student. Participants' majors or current primary academic disciplines included physics, economics, geography, meteorology, computer science, psychology, data science, library information science, geology, biology, mathematics, and management information systems. All participants had at least basic knowledge of statistics, such as linear regression or Pearson correlation. Most participants had completed at least one full semester statistics course. Participants were not compensated.

Procedure

We conducted the study with participants one-on-one via Zoom video conferencing. With the consent of all participants, we recorded audio and video with screen sharing. Participants were encouraged to think aloud and to ask questions. The procedure consisted of the following steps: an *introduction*, to give a quick overview and explain the purpose of the study; a *questionnaire*, to gather demographic and educational background details; a *knowledge assessment*, to evaluate participants' knowledge of mix effects and associated concepts; a six minute *video tutorial*, to give an overview of Wiggum and demonstrate its features; performance of a set of data exploration *tasks* (within-subjects) to assess utility and identify usability issues with Wiggum; and a *post-survey*, to gather subjective feedback about its design and features. We set up a local web



independent	dependent	splitby	subgroup	agg_trend	agg_trend_strength	subgroup_trend	subgroup_trend_strength	trend_type	comparison_type	distance
education marital-status occupation	income	workclass marital-status occupation	Federal-gov Local-gov Never-worked		0.15		0.15	Rank by Mean		0.2
education	income	workclass	Federal-gov	Preschool, 1st-4th, 5th-6th, 11th, 9th, 7th-8th, 10th, 12th, HS-grad, Some-college, Assoc-acdm, Assoc-voc, Bachelors, Masters, Prof-school, Doctorate	0.0237	10th, 12th, 5th-6th, 7th-8th, 11th, HS-grad, Some-college, 9th, Assoc-acdm, Assoc-voc, Bachelors, Masters, Prof-school, Doctorate	0.0501	Rank by Mean	aggregate-subgroup	0.4372
education	income	workclass	Local-gov	Preschool, 1st-4th, 5th-6th, 11th, 9th, 7th-8th, 10th, 12th, HS-grad, Some-college, Assoc-acdm, Assoc-voc, Bachelors, Masters, Prof-school, Doctorate	0.0237	1st-4th, Preschool, 10th, 7th-8th, 11th, 12th, 5th-6th, 9th, HS-grad, Some-college, Assoc-voc, Assoc-acdm, Bachelors, Masters, Doctorate, Prof-school	0.0175	Rank by Mean	aggregate-subgroup	0.5083
education	income	workclass	Never-worked	Preschool, 1st-4th, 5th-6th, 11th, 9th, 7th-8th, 10th, 12th, HS-grad, Some-college, Assoc-acdm, Assoc-voc, Bachelors, Masters, Prof-school	0.0237	10th, 11th, 7th-8th, HS-grad, Some-college	0.4737	Rank by Mean	aggregate-	0.5848

Figure 8. The visualization page for the adult data set. The distance heatmap view displays 9 out of 60 1×6 distance heatmaps; vertical scrolling reveals the others. After clicking a cell in the distance heatmap for a subgroup (A), the detail view displays the corresponding grouped bar chart and parallel coordinate plot for examining rank trend type (B).

Table 3. The predefined roles—dependent, independent, and splitby—of variables, the number of distance heatmaps, and their internal dimensions, generated in the user study.

Data Set	Size	Dependent Variables	Independent Variables	Splitby Variables	Distance Heatmap	
					Count	Dimensions
Iris	150	sepal length, sepal width	sepal length, petal length, petal width	class	3	2 × 3
Auto	392	mpg, acceleration	displacement, horsepower, weight, acceleration	cylinders, model year, origin	21	2 × 4
Facebook	495	page total likes, reach, consumers, num_comments, num_shares	page total likes, consumers, num_comments, num_likes, num_shares	type, category, month, weekday, hour, paid	44	5 × 5

server running Wiggum. Participant access was remote over public internet through a secure tunnel (*ngrok*[60]).

Data and Tasks

Our study employed three open-source data sets available from the UCI machine learning repository[57]: Iris, Auto, and Facebook (short for *Facebook metrics*). We chose these data sets because they give the exploration tasks three levels of complexity: easy, medium, and hard. These levels are driven by the number and dimensions of distance heatmaps. For example, the easy data set (Iris) generates three 2×3 distance heatmaps, whereas the hard data set (Facebook) generates 44 5×5 distance heatmaps. We preset the roles for all variables (see Table 3), allowing participants to get started with minimal data preparation. After loading the data, the participants chose *Pearson correlation* for the trend type, then clicked the *Visualize Trends* button. The number and dimensionality of distance heatmaps vary from data set to data set. (Some dimensions have too few records or attribute values to include. Heatmaps are not generated for such dimensions.)

Amar, et al. describe ten kinds of low-level analysis tasks used to evaluate the design of visualizations for data understanding[61]. We focus on tasks that involve retrieving values, finding anomalies, and assessing correlations. As our objective was to observe how Wiggum helps target users examine and understand mix effects, we designed tasks to capture an interaction process that includes input, output, and analysis steps[62]. We designed three tasks to afford participants an opportunity to gradually learn key concepts needed to understand mix effects: subgroup trend vs. aggregate trend, same trend vs. reverse trend, and Simpson's paradox vs. mix effects. Table 4 summarizes the goals, questions asked, and scoring rubrics applied for each task.

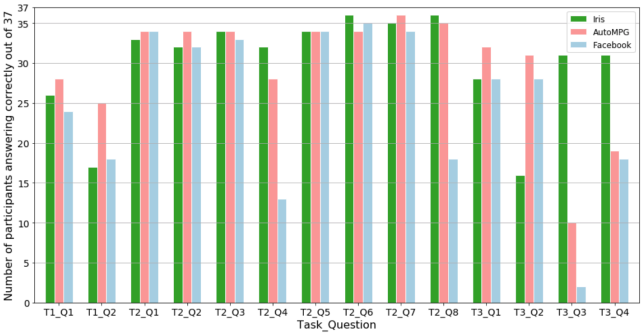


Figure 9. Correct answers for each task-question-data set combination.

Table 4. Tasks and questions in the user study.

Task	Question	Task Score Rubrics
Identify reverse trends in heatmaps	T1_Q1: How many reverse trends do you find? T1_Q2: Describe a reverse trend.	Score = average correctness per question (0.5 partial credit for simple miscount)
Identify subgroup or aggregate trends (for two different subgroups and pairs of variables: Q1–Q4 and Q5–Q8)	T2_Q1/Q5: What is the trend between variables for the aggregate data? T2_Q2/Q6: What is the trend between variables for the subgroup? T2_Q3/Q7: Are the two trends the reverse or the same? T2_Q4/Q8: What are the other subgroups' trends between the variables?	Score = average correctness per question
Confirm Simpson's paradox or mix effects	T3_Q1: How many instances of Simpson's paradox? T3_Q2: Describe one instance of Simpson's paradox. T3_Q3: How many instances of mix effects? T3_Q4: Describe one instance of mix effects.	Score = average correctness per question (0.5 partial credit for simple miscount)

Correctness

Figure 9 shows the number of participants who answered correctly for each {task, question, data set} combination. For correctness, scores were binary. Of the 37 participants, the majority (Iris: 26/37; Auto: 28/37; Facebook: 24/37) were able to identify all reverse trends (T1_Q1). For the Iris data set, most incorrect answers (8/11) arose from misconception about how reverse trends are represented in heatmaps; in those cases participants interpreted cells in the same position in heatmaps as a single reverse trend. For the other two data sets, most of the participants who gave incorrect answers (Auto: 6/9; Facebook: 8/13) miscounted reverse trends due to difficulty counting over the increased number of heatmaps and the need to scroll over them. Participants performed well on most questions when asked to identify a trend by giving specified information (T2_Q1–Q8). For the Facebook data set, participants performed poorly on T2_Q4 and T2_Q8 due to the large number of subgroups and their trends. In T3, the Iris data set exhibits only Simpson's paradox, and the Auto and Facebook data sets exhibit only mix effects. Participant responses offer insight even without a data set that exhibits both Simpson' paradox and mix effects. When asked to confirm mix effects (T3_Q3), the number of correct answers declined sharply between data sets (Auto: 10/37; Facebook: 2/37). This suggests that data set size substantially degrades (and ultimately limits) one's ability to visually examine mix effects in the Wiggum design. Evaluation on a wider range of data set sizes is needed to assess this hypothesis.

Time vs. Task Score

The average session duration was 109.05 minutes (s.d.=29.64). Mean completion times for T1 in the Iris, Auto, and Facebook data sets were 7.81 (s.d.=4.38), 6.92 (s.d.=3.62), and 5.95 (s.d.=2.40) minutes, respectively. For T2 these were 6.30 (s.d.=3.53), 5.70 (s.d.=2.36), and 5.92 (s.d.=3.24) minutes. For T3 these were 7.19 (s.d.=4.32), 7.19 (s.d.=3.37), and 9.73 (s.d.=6.73) minutes. For timing, T1 and T3 scores could earn half credit, to account for the prevalence of mental addition errors even when participants counted items correctly (see Table 4). Half credit produced the distribution of fractional average task scores shown in Figure 10. Task scores (color) are widely distributed over task durations (vertical position). They are only weakly correlated in a few of the nine task-data set combinations, such as higher scores on T1 taking longer particularly on the Iris data set, and much longer times to correctly count all mix effects in T3 on the Facebook data set. The former suggests that some participants were still learning Wiggum, with more or different training possibly needed. The latter suggests that task duration is closely coupled with counting effort, and thus that adding counting aids to the Wiggum design, such as some kind of selection highlighting to remember counted items, could significantly increase both efficiency and correctness of task performance.

Qualitative Results

We administered a post-survey, asking participants to rate Wiggum on four aspects. Average Likert ratings were high all around on a 5-point scale: easy to learn (3.9/5.0, s.d.= 0.66), easy to use (3.9/5.0, s.d.= 0.85), fast to use (4.1/5, s.d.= 0.95), and comfortable to perform tasks (3.9/5.0, s.d.= 0.97). While this feedback is encouraging, it indicates no obvious directions for design or feature improvement.

Some participants clearly found Wiggum helpful for learning mix effects. P10 wrote: *"I feel like after using the tool I have a more practical sense of what the paradox looks like, and the formal definition makes more sense to me now that I've seen examples and can put it in my own words."* Similarly, P9 commented, *"seeing it visually helped my understanding of the concept."* Unfortunately, learning to interpret a new type of visualization is not an easy task, and participants often struggled to identify instances of mix effects. Some participants were not able to understand the transition between heatmaps and trend plot, and did not realize that clicking on the same cell position in any heatmap of a given *splitby* variable highlights different subgroups in the same trend plot. In the distance heatmap view, we observed that participants were hindered by how Wiggum shows all distance heatmaps together in one scrolling layout rather than grouping them by their *splitby* variables. P15 wrote: *"Looking where a particular category ends and another begins was difficult too."*

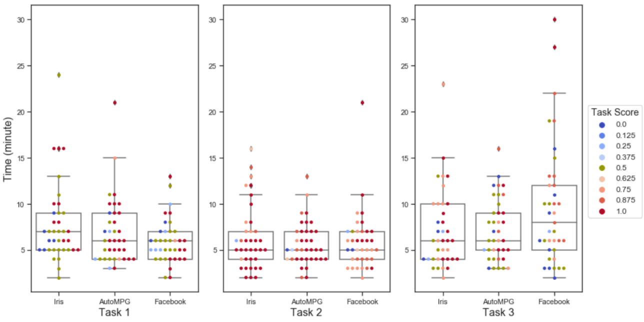


Figure 10. Time versus task score in each task level for the different data sets.

We found the distance heatmap view to help participants' awareness of different patterns. In general, participants enjoyed using it. P33 commented, *"I really liked the heat maps, and the fact that all subgroup trends for a particular relationship were shown at once."* P22 said, *"I preferred to use the heat maps for detecting Simpson's paradox, especially for the larger data sets, because the trendlines [in the trend plot] were difficult to see for larger sets."* However, some participants noted how working with larger data was a struggle. P29 wrote, *"The volume of data increased the difficulty to look and see if there was a Simpson's paradox."* P20 said, *"When there was a great number of variables to address in the Facebook data set, that feeling of comfort with the interface started to decline."* Participants sometimes overlooked the scroll bar, which hides itself in some browsers, thus missing some heatmaps, or accidentally zoomed in on individual heatmaps when trying to scroll.

Participants especially enjoyed the overall ease and efficiency of using Wiggum. P3 said, *"I liked it. It was easy to use and understand what was going on/what I was doing."* P24 pointed out that *"Visual feedback showing selected data sets and trendlines was very quick, essentially instantaneous."* Similarly, P16 commented: *"It was simple to use. It definitely helps you look at data in an easy manner."*

Discussion and Future Work

In this section, we discuss the strengths and limitations of the current version of Wiggum and highlight opportunities for future development.

Trends Types and Comparisons

Wiggum is currently designed to analyze for rank and regression trends (**AG2**), which are most commonly used in practice for studying mix effects and Simpson's paradox. However, exploring other trend types is also valuable. Wiggum can be extended to accommodate additional statistical measures, such as those relevant to the performance of binary classification tasks. For example, the difference between precision for an entire population and conditioned on a specific subgroup ($\frac{TP}{TP+FP}$) is a promising measure to support fairness forensics[10,63–65]. Exploring non-linear trends is also a promising direction, and further effort will be needed to develop and investigate non-linear metrics, such as Spearman's correlation, in future research. Although integrating novel metrics and visualizations for new trend types is a challenging task, it has the potential to benefit users across a broader range of fields (**AG3**). In addition, beyond the brute force discovery of clusters currently implemented in Wiggum, future work on more flexible proxy techniques to recover the omitted *splitby* variable could enable Wiggum to further support fairness forensics. Wiggum's current support for comparison between aggregate and subgroup trends could be extended to interesting and practical comparison between subgroups themselves.

Visualization Design

Participants performed poorly at identifying and counting mix effects in larger data sets. Techniques to increase visual scalability for larger and more complex data sets (**AG1**) is a clear direction for improvement. In practice, we find it helps to report summary information such as the number of mix effects. In future work we will explore more guided (less ad hoc) approaches to selecting dimensional subsets. It can also be difficult to differentiate the subgroups of different *splitby* variables in the current heatmap layout (**AG3**). This could be addressed by more clearly organizing the heatmaps by their subgroups. The heatmaps also become denser as rows and columns increase in number. An overview heatmap over all subgroups might help with checking trends in individual heatmaps. The same issue arises in the parallel coordinate plot as the number of distinct values of the *splitby* variable increases. Although parallel coordinates can help users observe changes of ranking between aggregate and subgroup data, the distance and strength of a subgroup trend are relatively hard to read visually. Alternative visualization techniques could support more precise reading and hence easier interpretation of those statistical measurements. (The bump chart[66] is a candidate, but it poorly shows relative rate difference, which is crucial for rank comparison.) The parallel coordinates plot favors comparing the aggregate to the user-selected subgroup over comparing it to the other subgroups. Observing line crossings between the aggregate axis and the remaining axes can be challenging, but the transitivity of the crossing can help perceive it from the previous axes. For instance, if users notice a crossing to the second axis from the privileged axis and observe no further crossings, they can infer that all the others have crossings with the privileged axis. In the trend plot, the angle between two regression lines is hard to interpret and compare across axis scale changes, which happens when users select a different cell in a heatmap. We plan to explore visual techniques to help users read angles; simple radial plots may be a promising option for this task.

Causality

Wiggum demonstrates the ability to identify information, relationships, and patterns within the foraging activity of the analytical process. While it offers statistical tools to examine mix effects and Simpson's paradox, deriving clear visual inferences from these phenomena can be challenging (**AG4**). Causal inference provides a practical approach for re-evaluating and explaining these phenomena. For instance, identifying variables strongly associated with mix effects and conducting a causality analysis on those variables could support the sense-making process. Additionally, incorporating methods to visualize the network of causal relationships may further improve the interpretation of results. Developing new techniques for interactive visualizations to explore confounding bias, interventions, and counterfactuals could also improve data interpretation and support decision-making. However, automated algorithms for causal discovery are not always entirely reliable. As a result, visual analytic approaches present a viable alternative to fully automated methods and may be effectively integrated with automated causal detection techniques within a human-in-the-loop framework to create hybrid solutions (**AG3**).

Scalability

Moving forward, we will investigate the scalability of computation and visualization in Wiggum in terms of data size and dimensionality (**AG1**), particularly in terms of the number of independent, dependent, and splitby variables. We conducted a preliminary study that produced three findings. First, varying data sizes doesn't slow down the computation significantly. Second, as the number of dependent/independent variables increases, we see a trend showing a quadratic increase in the running time. Lastly, as the number of total subgroups increases, there exists a trend showing a linear increase in the running time. The current visualization design (see Section 5) anticipated and addressed some scalability issues. Moreover, filtering low correlation pairs of dependent and independent variables after the computation of aggregate data can alleviate visual complexity by reducing the number of rows and columns in the heatmap view. As the number of subgroups increases, Wiggum might display only the top-k heatmaps that have the highest average distance scores. Followup studies could be conducted to investigate multiple optimization methods with Wiggum with the goal of supporting data analysis at larger scales.

Opportunities for Automated Approaches

Wiggum provides a comprehensive overview of all detected instances and allows users to explore mixed effects. This method is generally effective for small datasets when it is necessary to inspect the context of specific instances of interest. However, when applied to large datasets, the efficiency may be compromised due to the visualization of numerous graphical elements. To alleviate the cognitive burden on users, automated approaches could be employed. For instance, automating the detection of instances with significant trend distances and then presenting these instances with relevant details for further analysis could streamline the process. This automation may also assist in prioritizing a subset of dimensional combinations and enable further queries based on that subset. Additionally, these automated approaches may aid in creating schemas and generating hypotheses, thereby extending Wiggum's role in the sense-making process (**AG4**). Moreover, if pairs of attributes with spurious relationships can be automatically identified and filtered before being presented to users, it could enhance the efficiency of the exploration phase and reduce the likelihood of incorrect interpretations (**AG1, AG4**). Nevertheless, users may not always prioritize identifying the most prominent cases of mixed effects with respect to trend distance and strength. In such cases, the automated approach becomes less helpful, especially when users need to focus on or filter specific cases based on prior knowledge. The flexibility in dimension selection within Wiggum supports the exploration of hypotheses related to specific dimensions (**AG1**). Lastly, while interactions in Wiggum initiate computation for three levels of ranking for instances of mixed effects, the heatmap overview is not effective at displaying ranks. Automatically selecting the appropriate visualization to show the ranking information will be crucial for effective exploration and interpretation of ranks (**AG3, AG4**). These improvements in automated approaches have the potential to enhance both sensemaking and the overall analysis. They inspire further investigation in future research.

Conclusion

Wiggum is an integrated system for detecting, ranking, and visualizing mix effects and Simpson's paradox. A distance heatmap view provides an overview of subgroup trends and offers clues to find potentially important ones. Trend plots and parallel coordinate plots support detailed examination of mix effects and Simpson's paradox. Coordination of multiple views through a variety of supporting interactions helps users browse and examine patterns related to trend reversal. Users can filter, rank, and compare instances on trend strength and distance to eliminate spurious subgroup trends. Overall, Wiggum appears effective for underpinning deep visual analysis of multiple trend types in real-world multidimensional data sets. The user study uncovered needed design improvements and additional features for ongoing development and study of Wiggum and its capabilities.

Appendix: Glossary of Terms

There could be some variation in the words about what they mean from different communities, but the appendix provides the definitions of the terms that we're adopting.

Trend A tendency of a variable in which another variable is changing in the data set. For example, an upward trend of admission rates when gender changes from female to male.

Regression Trend A linear relationship between two variables.

Rank Trend Ranking groups defined by one variable according to a statistic calculated on a second variable.

Subgroup A group of data created by grouping instances under a certain condition.

Mix Effects A statistical phenomenon occurs when some subgroups of the data partitioned by a certain condition exhibit trends opposite to the data as a whole.

Simpson's Paradox A statistical phenomenon occurs when all subgroups of the data partitioned by a certain condition appear the opposite trend of the aggregate data, and it is a more special case of mix effects.

Dependent and independent variables Variables used to compute a trend. For a regression trend, Wiggum attempts to model the relationship between them. For rank trends, the aggregation method is applied on the dependent variable. The independent variable partitions a population into ranked groups based on the aggregate result.

Splitby Variable A variable used to group rows that have the same values into multiple subgroups.

Subset A feature subspace of a data set consisting of two columns (i.e., a pair of dependent and independent variables).

Trend Distance A measurement of the discrepancy between two trends.

Trend Strength A metric for assessing how well a trend represents data being examined.

Acknowledgements

We thank Charles D. Nicholson, Mark Laufersweiler, and the OU Data Analytics, Visualization & Informatics Syndicate for constructive feedback. We thank Darby Cobb for data cleaning.

References

1. Bickel PJ, Hammel EA, O'Connell JW et al. Sex bias in graduate admissions: Data from Berkeley. *Science* 1975; 187(4175): 398–404.
2. Pearl J. Comment: understanding Simpson's paradox. *The American Statistician* 2014; 68(1): 8–13.
3. Armstrong Z and Wattenberg M. Visualizing statistical mix effects and Simpson's paradox. *IEEE Transactions on Visualization and Computer Graphics* 2014; 20(12): 2132–2141.
4. van Wijk J. The value of visualization. In *VIS 05. IEEE Visualization, 2005.* IEEE, pp. 79–86.
5. Pirolli P and Card S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, volume 5. McLean, VA, USA, pp. 2–4.
6. Cook KA and Thomas JJ. Illuminating the path: The research and development agenda for visual analytics. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2005.
7. Gahegan M, Wachowicz M, Harrower M et al. The integration of geographic visualization with knowledge discovery in databases and geocomputation. *Cartography and Geographic Information Science* 2001; 28(1): 29–44.
8. Gahegan M. Beyond tools: Visual support for the entire process of giscience. In *Exploring Geovisualization*. Elsevier, 2005. pp. 83–99.
9. Sacha D, Stoffel A, Stoffel F et al. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 2014; 20(12): 1604–1613.
10. Buolamwini J and Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, pp. 77–91.
11. Cabrera ÁA, Epperson W, Hohman F et al. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, pp. 46–56.
12. Blyth CR. On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* 1972; 67(338): 364–366.
13. Norris F. Median pay in us is stagnant, but low-paid workers lose. *New York Times* 2013; .
14. Freitas AA. On objective measures of rule surprisingness. In *European Symposium on Principles of Data Mining and Knowledge Discovery*. Springer, pp. 1–9.
15. Salimi B, Gehrke J and Suciu D. Bias in OLAP queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, pp. 1021–1035.
16. Wardrop RL. Simpson's paradox and the hot hand in basketball. *The American Statistician* 1995; 49(1): 24–28.
17. Pearl J and Mackenzie D. *The Book of Why: The New Science of Cause and Effect.* Basic Books, 2018.
18. Kievit RA, Frankenhuis WE, Waldorp LJ et al. Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology* 2013; 4.
19. Xu C, Brown SM and Grant C. Detecting Simpson's paradox. In *Florida Artificial Intelligence Research Society Conference (FLAIRS)*. pp. 221–224.
20. Hernán MA, Clayton D and Keiding N. The Simpson's paradox unraveled. *International Journal of Epidemiology* 2011; 40(3): 780–785.
21. Tu YK, Gunnell D and Gilthorpe MS. Simpson's paradox, Lord's paradox, and suppression effects are the same phenomenon–the reversal paradox. *Emerging Themes in Epidemiology* 2008; 5(1): 2.
22. Alipourfard N, Fennell PG and Lerman K. Can you trust the trend?: Discovering Simpson's paradoxes in social data. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, pp. 19–27.
23. Alipourfard N, Fennell PG and Lerman K. Using Simpson's paradox to discover interesting patterns in behavioral data. *Twelfth International AAAI Conference on Web and Social Media* 2018; 12(1): 2–11.
24. Lerman K. Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Science* 2017; : 1–10.
25. Day SM. Simpson's paradox and Major League Baseball's hall of fame. *Red* 1994; 4: 8.
26. Charig CR, Webb DR, Payne SR et al. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br Med J (Clin Res Ed)* 1986; 292(6524): 879–882.
27. Julious SA and Mullee MA. Confounding and Simpson's paradox. *BMJ* 1994; 309(6967): 1480–1481.
28. Froelich W. Mining association rules from database tables with the instances of Simpson's paradox. In *Advances in Databases and Information Systems*. Springer, pp. 79–90.
29. Arah OA. The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect: covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology* 2008; 5(1): 5.
30. Pearl J. Simpson's paradox: An anatomy. *Department of Statistics, UCLA* 2011; .
31. Salimi B, Cole C, Li P et al. HypDB: a demonstration of detecting, explaining and resolving bias in OLAP queries. *Proceedings of the VLDB Endowment* 2018; 11(12): 2062–2065.
32. Wang J and Mueller K. The visual causality analyst: An interactive interface for causal reasoning. *IEEE transactions on visualization and computer graphics* 2015; 22(1): 230–239.
33. Wang J and Mueller K. Visual causality analysis made practical. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, pp. 151–161.
34. Pearl J. *Causality*. Cambridge university press, 2009.

35. Bandyoapdhyay PS, Nelson D, Greenwood M et al. The logic of Simpson's paradox. *Synthese* 2011; 181(2): 185–208.
36. Pavlides MG and Perlman MD. How likely is Simpson's paradox? *The American Statistician* 2009; 63(3): 226–233.
37. Fabris CC and Freitas AA. Discovering surprising patterns by detecting occurrences of Simpson's paradox. In *Research and Development in Intelligent Systems XVI*. Springer, 2000. pp. 148–160.
38. Rücker G and Schumacher M. Simpson's paradox visualized: the example of the rosiglitazone meta-analysis. *BMC Medical Research Methodology* 2008; 8(1): 34.
39. Chen A, Bengtsson T and Ho TK. A regression paradox for linear models: Sufficient conditions and relation to Simpson's paradox. *The American Statistician* 2009; 63(3): 218–225.
40. Yan JN, Gu Z, Lin H et al. Silva: Interactively assessing machine learning fairness using causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. pp. 1–13.
41. Xie X, Du F and Wu Y. A visual analytics approach for exploratory causal analysis: Exploration, validation, and applications. *IEEE Transactions on Visualization and Computer Graphics* 2020; 27(2): 1448–1458.
42. Jin Z, Guo S, Chen N et al. Visual causality analysis of event sequence data. *IEEE Transactions on Visualization and Computer Graphics* 2020; 27(2): 1343–1352.
43. Choudhry A, Sharma M, Chundury P et al. Once upon a time in visualization: Understanding the use of textual narratives for causality. *IEEE Transactions on Visualization and Computer Graphics* 2020; 27(2): 1332–1342.
44. Elmqvist N and Tsigas P. Causality visualization using animated growing polygons. In *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No. 03TH8714)*. IEEE, pp. 189–196.
45. Guo Y, Binnig C and Kraska T. What you see is not what you get!: Detecting Simpson's paradoxes during data exploration. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. ACM, p. 2.
46. Crotty A, Galakatos A, Zgraggen E et al. Vizdom: interactive analytics through pen and touch. *Proceedings of the VLDB Endowment* 2015; 8(12): 2024–2027.
47. Wexler J, Pushkarna M, Bolukbasi T et al. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics* 2019; 26(1): 56–65.
48. Ahn Y and Lin YR. FairSight: Visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics* 2019; 26(1): 1086–1095.
49. Wang Q, Xu Z, Chen Z et al. Visual analysis of discrimination in machine learning. *IEEE Transactions on Visualization and Computer Graphics* 2020; .
50. Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 2018; 16(3): 31–57.
51. Saha D, Schumann C, Mcelfresh D et al. Measuring non-expert comprehension of machine learning fairness metrics. In *International Conference on Machine Learning*. PMLR, pp. 8377–8387.
52. Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: Machine learning in Python. *the Journal of Machine Learning Research* 2011; 12: 2825–2830.
53. Allison JS, Santana L and (Jaco) Visagie I. A primer on simple measures of association taught at undergraduate level. *Teaching Statistics* 2022; 44(3): 96–103.
54. De Winter JC, Gosling SD and Potter J. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods* 2016; 21(3): 273.
55. Croux C and Dehon C. Influence functions of the spearman and kendall correlation measures. *Statistical Methods & Applications* 2010; 19: 497–515.
56. Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. In *The Craft of Information Visualization*. Elsevier, 2003. pp. 364–371.
57. Lichman M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.
58. Bostock M, Ogievetsky V and Heer J. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics* 2011; 17(12): 2301–2309.
59. Flask web framework. http://flask.ocoo.org.
60. Ngrok: Secure introspectable tunnels to localhost. https://ngrok.com.
61. Amar R, Eagan J and Stasko J. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*. IEEE, pp. 111–117.
62. Lam H, Tory M and Munzner T. Bridging from goals to tasks with design study analysis reports. *IEEE Transactions on Visualization and Computer Graphics* 2017; 24(1): 435–445.
63. Crawford K. The trouble with bias. *Conference on Neural Information Processing Systems, Keynote* 2017; .
64. Holstein K, Wortman Vaughan J, Daumé III H et al. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. pp. 1–16.
65. Dwork C, Hardt M, Pitassi T et al. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. pp. 214–226.
66. Tufte E. *Envisioning Information*. Cheshire, CT: Graphics Press, 1990.