Integrated Data Analysis for Parametric Design Environment

\textit{mineR: a Grasshopper plugin based on R}

Abdulmalik Abdulmawla\textsuperscript{1}, Sven Schneider\textsuperscript{2}, Martin Bielik\textsuperscript{3}, Reinhard Koenig\textsuperscript{4}
\textsuperscript{1,2,3}Bauhaus-University Weimar \textsuperscript{4}Bauhaus-University Weimar and Austrian Institute of Technology
\textsuperscript{1,2,3,4}\{abdulmalik.abdulmawla|sven.schneider|martin.bielik|reinhard.koenig\}@uni-weimar.de

In this paper we introduce mineR- a tool that integrates statistical data analysis inside the parametric design environment Grasshopper. We first discuss how the integration of statistical data analysis would improve the parametric modelling workflow. Then we present the statistical programming language R. Thereafter, we show how mineR is built to facilitate the use of R in the context of parametric modelling. Using two example cases, we demonstrate the potential of implementing mineR in the context of urban design and analysis. Finally, we discuss the results and possible further developments.

Keywords: Statistical Data Analysis, Parametric Design

Introduction

Design is a cyclic process aimed on finding an optimal solution for a given problem. This process can be broken down into two stages: First, finding an acceptable initial design plan that might provide a baseline for the desired solution. Second, further development and optimisation of the initial design plan into an implementable design solution (Karimi 2012). In this process, the design generation is informed by design evaluation in a series of successive steps - the design cycle. Since each step aims to further improve the design, it is important that the design workflow promotes ease of transition between the generation and evaluation. During the recent years, parametric design has emerged as a quantitative design approach that connects the parameters for generating and evaluating a design via algorithms to shorten the optimisation cycle and providing the designer with an access to the data throughout the design stages (Motta 1999). Parametric design approach is usually connected to programmable tools that have the capacity to generate and manipulate complex data. These tools help assign values to the design parameters as well as optimize the design to satisfy specific criteria.

To test and analyse a design, designers apply evaluation methods for different design criteria. Some of these criteria are more qualitative and can be
better evaluated intuitively. For example, to evaluate the aesthetics of a building, architects support the human intuition with graphical representations that convey the experience of being inside that space. Other design criteria are more quantitative, therefore, intuition can be hard to employ and sometimes even misleading. Therefore, computational analysis tools offer valuable support, however in cases where numerous criteria are used, the evaluation of this computed analysis results can become complex. For example, for the optimisation of the economic, ecological and social performance of a city the evaluation of a multitude of criteria needs to be considered (e.g. density, accessibility, visibility, costs, energy demand). Therefore, data analysis methods are useful in order to efficiently and effectively explore quantitative design data. In this paper we will focus on the quantitative evaluation of parametric design data using computational statistics.

**Integrating statistics into parametric modelling**

There are different tools that can be used for data analysis like Stata, SPSS and R. These tools exist as professional software independent from the design environment. Therefore, for designers to use these tools in parallel to the design environment the data needs to be exported from the design environment to the statistical environment. Thereafter, the analysis results need to be transferred back to the design environment. Here it is critical to highlight that these statistical tools sometimes have complex interfaces, and mostly require a sound knowledge of programming to access basic functions. Consequently, using statistical tools in the design cycle becomes a time-consuming task. Thus, both statistical tools and parametric design need to be connected more tightly.

The idea of bridging the gap between both the parametric design environment and the statistical data analysis tools is not entirely new. Some tools already exist to provide such a connection. For example, the machine learning tool in Lunchbox by Proving Ground (1), the data visualization tool Conduit by the same developer, and the excel dependent tool Bumblebee by NeoArchaic Design (2). These tools used the node-based parametric design environment Grasshopper to construct a set of pre-configured functions (grasshopper components). But they either provide basic data visualization, connect to a non-professional statistical software like excel, or lack some essential and advanced methods of statistical data analysis -descriptive and inferential statistics- that are necessary for data exploration.

**Method & Implementation**

Integrating statistical tools into parametric design could lead to discovering knowledge about the data what would not be possible otherwise. Moreover, since this knowledge would exist as statistical data output inside the parametric environment, it could be used further in the design process as input parameter. This would open the door for more advanced design generation and exploration possibilities. Therefore, choosing a flexible statistical analysis tool would be an essential part of the integration process.

For the tool we present in this paper, we selected R (R Core Team 2017) - the popular opensource language for statistical computing - as a perfect tool to be integrated inside the parametric design environment Grasshopper. R provides statistical methods for the Exploratory Data Analysis (EDA) that were introduced by Tukey (1977). EDA consists of two main steps: descriptive and inferential statistics.

First, one starts with descriptive statistics to clarify what the data shows and to understand its basic characteristics. This step uses data summaries and simple graphics to offer a quick overview about the types and distribution of the data. This facilitates deciding the appropriate way to examine the data further. R provides methods to graphically display different data variables such as histograms for understanding the distribution of numerical data, Box plots to show the spread of the data and Bar plots for comparing categorical data.
Second, one uses inferential statistics to explore the relation between data and to support explaining the behaviour of this relation through different metrics. This provides the possibility to generalize and extrapolate beyond the current data sample. Moreover, it can direct the designer to different methods on how to improve the performance of a design. The use of both methods in R - Descriptive and Inferential Statistics- is mostly a linear process inside the design cycle; one uses descriptive statistics to decide how to proceed further to the inferential statistics. However, accessing each stages’ functions independently should provide more flexibly platform. Therefore, choosing R would facilitate a richer environment that host a variety of methods that could benefit the design process once integrated.

**mineR**

Our tool mineR is built for the node-based parametric design plugin Grasshopper3D for Rhino (3). The software interface RDotNET(4) - an opensource CSharp library - provides an interface for various programs to connect with R kernel (a term for the core functions of a software). mineR can be integrated in the workflow as seen in Figure 1. It can receive data from Grasshopper and pass them to R, which then can either provide the response for this data with summaries and matrices or by plotting them into charts in a separate window. Since the whole concept of the parametric design environments is wrapped around the idea of flexible customization, this also counts for mineR as well. Thus, mineR is structured in multiple components that provide the essential functions of R, and still provide the possibility of accessing the parameters of these functions from Grasshopper.

mineR consists of three different types of pre-built functions (components); descriptive, inferential and custom. Descriptive components in mineR deal with different types of quantitative data variables like numerical and categorical variables (Diez et al. 2015). These components display the data in plots like Histogram, Bar Plot, Box Plot and the regular Curve Plot. They can as well provide related data summary including Standard Deviation(SD) and Interquartile Range(IQR). Inferential components in mineR deal with exploring the relation between nu-
numerical data and provide summaries of the result, like Linear Regression (with correlation coefficient r and the p-value) or Heat Map of correlations. The last type of components are custom components that provide direct access to Rscript (the programming language of R) but still can receive and send data while remaining inside the parametric design environment.

The design of each component in mineR is driven by both environments: how Grasshopper works and how R is being implemented. The input of each component can receive the corresponding data type inside Grasshopper. Attached to them are related constraints to control how R should deal with this data input. Moreover, each plotting component has inputs to modify the data printed on the plot. This includes text describing what the plot shows, and the custom colors for categorical data if required. Furthermore, each plot component can export the charts either to be displayed on the users’ screen in a parallel window to Grasshopper, or to be directly stored as PDF document on the user device.

Each mineR component has specific functions to deal with different types of data variables and provide the corresponding output to these functions. For example, the BarPlot component automatically transforms an unsorted list of categorical data into organized lists. It will also provide a summary of their category names, the frequency with which they appeared in the list, and their relative ratios inside the whole list. The Correlation component calculates a linear regression test on two datasets to describe the nature of this relation by the correlation coefficient (Person or Spearman) with values between -1 and +1 (indicating the strength and direction of the relation). Moreover, with the p-value it also provides a metric to describe how high the probability is that the respective relation is found by chance.

**Application**

To show how mineR can be used inside the parametric design environment, in the following we present two exemplary case studies related to urban design and analysis. The first example will show how the evaluation of an existing dataset about a city could be done using the essential components. The second example uses mineR to create an evaluation criterion for generative street network design. We will use in the first example the essential descriptive statistical methods in mineR, while in the second we will focus more on the use of inferential statistics. Moreover, we will be using these examples to describe more information about mineR components, and how their output could be used inside Grasshopper.

**Case 1: Urban Data Analysis**

The first example demonstrates the functionality of the different components using a case for analysing urban form and its relationships to the distribution of building uses (case is taken from the paper Schneider et al., 2017). We use the descriptive statistics components and simple inferential statistics components for analysing urban data of a small city in Thuringia, Germany (Zella-Mehlis). The data, collected in this example, is stored in a single Rhino file, which contains geometrical information about the city, including the street network, street blocks, plots, buildings in simple level of details (LOD1). Moreover, it contains empirical data collected from our visit to the city, such as movement countings and information about the uses of each building.

The exploration of the data was done in the following steps; to gather knowledge about the city, we used Grasshopper to first organize data about the available geometry. This led to creating lists of data that contains information about street segments length, area of each plot, area of each buildings in footprint and in total, number of storeys in each building, and finally the number and type of functions inside each building. To understand how the morphology affects the performance of the city, we analysed the city geometry using grasshopper analysis tools. We measured the density of the city for each plot and street block (Floor Area Ratio FAR and Ground Space Index GSI). And analysed the street network accessibility using different centrality metrics (on different radii to for both the angular and...
Figure 2
Simple Descriptive and Inferential Statistics for the city Zella-Mehlis using mineR for Grasshopper:
(Histogram) distribution of street segment lengths. (BoxPlot) right: the same data about street segment length, left: Density (FAR) of each function. (BarPlots) left: number buildings per storeys count, right: number of buildings categories by the function. (Heatmap) correlation matrix between different centrality radii and number of functions.

metrical distance) using the CityGraph component of the DeCodingSpaces Toolbox (Fuchkina, 2016). This resulted in two types of data; data that describes the geometrical properties of the city, and data that describes the performance of this geometry.

In the following we exemplary show, how the mineR components can be used to quickly derive further information from the existing urban data. First, regarding the street segment lengths, on the one hand, a histogram provides information about all the groups of lengths and how frequent each group is in the city network. This shows that the city has very few number of segments longer than 120 m (Fig. 2, first row). On the other hand, a Box Plot of the street segment lengths more clearly shows that most segment lengths are between 35 - 90 m (Fig. 2, second...
Second, using the types of building uses, we computed the number of buildings of each use and visualized them using the Bar Plot. This shows that almost 85 percent of buildings were residential buildings, and almost 8 percent of the buildings were commercial (Fig. 2, third row). Third, the data about the buildings was categorized by the number of storeys using the Bar Plot. This shows that most of the buildings are 3 storeys or less with almost 60 percent of the buildings having two storeys. Visualizing the density measure FAR for differently used plots, the Box Plots show that the plots of certainly used buildings have a FAR of around 0.3. Thereby industrially used plot have the highest values and private clinics the least. Plots for educational and public use have highest spread in FAR, indicating very different building forms for these uses. Simple inferential statistics was done using the Heat Map component to discover relations between the centrality measures of the street segments and the number of certain uses attached to them (e.g. commercial and vacant buildings). This showed a strong correlation ($r = +0.75$) between the number of commercial functions and the local angular choice radius 200, but no considerable correlation to the vacant buildings (see Fig. 2, bottom).

**Case 2: Optimizing the synergy of a Street Network Design**

In the second example, we generate a street network that exhibits a certain characteristic, which is a high synergy. Synergy is a measure, describing the relationship between local and global street network centrality (Hillier, 1996, p. 99 - 101). High synergy, thus describes, how well local centers in the city connect to globally central streets. Using the correlation component in mineR we will leverage the output of the correlation coefficient($r$) to create the evaluation criteria synergy for the generated design.

The study was done using Grasshopper with the following steps; to provide the initial design constraints for the city, we drew the city boundary using a hexagon with a diameter of 1.4 kilometer with 4 existing external roads. This provided the inputs for the street network generation (NetworkSynthesis) component (Koenig, Treyer, & Schmitt 2013; Koenig 2015). This component generates variants of street networks based on different input parameters such maximum segment length, random angle between segments, maximum number or connections from each street, the depth of these connections and a random value to control the location of the intersection points. The outcome of this component was connected to the CityGraph Component to calculate local and global closeness centrality of the street network. In order to calculate the synergy measure, we used the correlation component of mineR. The goal of this test was to achieve the highest possible correlation between the local closeness centrality (Radius 300) and global closeness centrality.

Fig. 3 displays five typical outcomes of this process. Thereby the centrally shown variant exhibited the highest synergy value. For each variant the street network is colored according to global closeness centrality. The information under each output expresses the input parameters used to generate the network, as well as mineR output that we used to test the performance of this network.

The change of variables was in few cases inconsistent. For example, increasing the maximum segment length did not always provide the best result, but values around 200 produces the highest $r = +0.52$ (Fig. 3, upper left corner). Furthermore, values lower than 150 were most decreasingly lower under $r = +0.300$. Another example, when increasing the value of the random angle between segments was inconsistent as well, but values between 25 and 40 were marked as medium around $r = +0.458$ (Fig. 3, lower left corner). Changes in other variables were more consistent, but due to their effect on the increase of computation time they were kept at lower values. Increase in maximum number or depth of branching arms was mostly corresponded with higher correlation. Manipulating the random seed values that controls the location of the points input provided a wide range of results, with the highest at 54, $r = +0.589$ (Fig. 3, center).
Figure 3
Parametric design with high Synergic correlation test using mineR for Grasshopper: show multiple stages of optimization while manipulating the data inputs to a street network generation algorithm.

**Conclusion & Outlook**
Quick access to statistical methods and data could improve the design cycle because it supports the interpretation of design data and thus improve decision making. In this paper we presented how mineR implements statistical analysis methods from R into easy-to-use components for Grasshopper. Then using two examples, we demonstrated the application of mineR for design evaluation and used the evaluation result as optimization objective.

In the two exemplary cases shown in this paper; mineR provided typical methods used in both design and research to explore and test the design data. In the first case, the statistical methods we explained were used to gain generalizable knowledge of how cities function based on their urban form. The second example showed how an evaluation criterion can be created and optimized to govern the design generation process.

In both cases we evaluated the available design and performance data while remaining the whole time inside grasshopper interface. Moreover, once such analysis is set up inside grasshopper, this evaluation process then could be accessed and manipulated to satisfy any further development. This reduces the effort of thinking about the tools and focusing more on the main objective of the design or research.

Integrating statistics into parametric design can probably lead to an improved design workflow and new approaches to design. This is due to not only the decrease in the number of steps designers need to take to access statistical methods. But also due to the nature of the parametric design environment itself. Since the commands are passed back and forth as input and output variables, they can be flexibly used in the whole process of designing.

Integrating R as a professional data analysis tool in a parametric modeling environment like Grasshopper provides a direct access to vast amount of meth-
ods and applications that could be explored in many directions. There are still more advanced methods and plots in R that could be included, such as the Chi-squared test, Multivariate analysis. Moreover, since both R and Grasshopper have gathered a large community of users from many fields, we are planning to publish mineR as an opensource library which could open the door for further developments.

REFERENCES

Diez, DM, Barr, CD and Cetinkaya-Rundel, M 2012, OpenIntro statistics, CreateSpace
Hillier, B 1996, Space is the machine: A configurational theory of architecture, Cambridge: Cambridge University Press
Tukey, JW 1977, Exploratory data analysis, Addison Wesley Publishing
Woodbury, R 2010, Elements of parametric design, Taylor & Francis Group