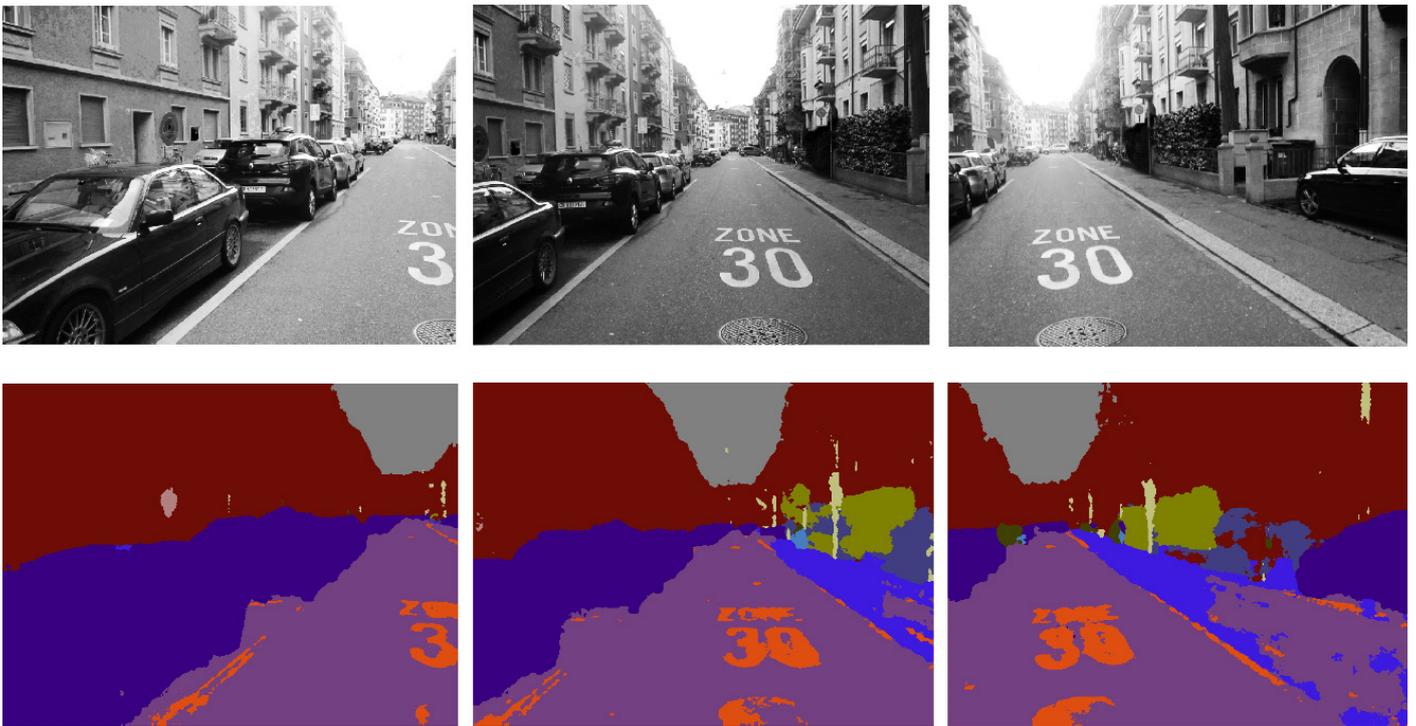


Documentation of the teaching results from the spring semester 2017

Creative Data Mining



Danielle Griego, Dani Zünd and Prof. Dr. Gerhard Schmitt

DARCH

Chair of Information Architecture

Creative Data Mining

Documentation of teaching results

Danielle Griego, Dani Zünd, and Gerhard Schmitt



Chair of Information Architecture

Teaching

Danielle Griego, Dani Zünd and Gerhard Schmitt

Syllabi

<http://www.ia.arch.ethz.ch/category/teaching/fs2017-creative-data-mining/>

Seminar

Creative Data Mining

Students

Biyu Wang, Jiani Liu, Jack Kenning, Hong Liang, Gong Chen, Zhonghau Dai

Published by

Swiss Federal Institute of Technology in Zurich (ETHZ)

Department of Architecture

Institute of Technology in Architecture

Chair of Information Architecture

Wolfgang-Pauli-Strasse 27, HIT H 31.6

8093 Zurich

Switzerland

Zurich, August 2017

Layout

Brigitte M. Clements

Contact

grigod@arch.ethz.ch | <http://www.ia.arch.ethz.ch/griegotd/>

Cover picture:

Front side: Labelled Frames. Jack Kenning

Course Description and Program



Creative Data Mining Uncover and Evaluate

The participants of this course learn how to collect, process, analyze and interpret real spatial and temporal data in order to work with quantifiable qualities in urban planning. This is achieved by using actual data from a recent study and analysing it with different data processing and machine learning techniques.

The goal of the course is to explore a specific research question about the urban environment and test the stated hypothesis using different techniques presented in the course, thus preparing students with a skill-set to further support their design and decision making processes.

The course focuses on creating deeper insights to critically evaluate design alternatives for urban planning projects. Students will work with time-series and geo-referenced data including temperature, relative humidity, illuminance, noise, people density, and dust particulate matter. Subjective impression survey data will also be integrated into the student projects to further explore influencing factors of the urban environment on our perceptual experiences. Non-architectural skills the participants can develop during this course are 1) an introduction to programming 2) how clustering methods like PCA or K-Means could be applied in an architectural context.

Where

HIT H 34.1 (Video Wall)

Supervision

Danielle Griego
Daniel Zünd
Artem Chirkin

griego@arch.ethz.ch
zuend@arch.ethz.ch
chirkin@arch.ethz.ch

Prof. Dr. Gerhard Schmitt
Chair of Information Architecture
Information Science Lab
Wolfgang-Pauli-Strasse 27, 8093 Zurich
www.ia.arch.ethz.ch

- 20.02.2017 **Course Introduction**
- 27.02.2017 **Introduction to Python & Programming I**
- 06.03.2017 **Introduction to Python & Programming II**
- 13.03.2017 **Data Processing**
- 20.03.2017 **Seminar week (No lecture)**
- 27.03.2017 **Intro to time-series data analysis**
- 03.04.2017 **Time series data analysis ctd. & Machine learning**
- 10.04.2017 **Machine learning ctd.**
- 17.04.2017 **Holiday (No lecture)**
- 24.04.2017 **Programming tutorial applications**
- 01.05.2017 **Holiday (No lecture)**
- 08.05.2017 **Q&A Feedback Workshop I**
- 15.05.2017 **Final iA critique**
Combined critique with the other iA courses
(13:00 - 18:00)

Requirement Former knowledge of any digital tool or coding language is most welcome but NOT required. You only need to provide a reasonable amount of motivation and of course a notebook.

* Total 60 h = 2 ECTS
Ungraded Semester Performance

The most recent outline will be found on www.ia.arch.ethz.ch

Content

Does Distance Affect our Feelings? p.9

Student: Biyu Wang & Jiani Liu

Urban Perception: A deep learning approach p.32

Student: Jack Kenning

Street Cross Section p.49

Student: Hong Liang

Security Analysis p.103

Student: Gong Chen & Zhonghau Dai

Objectivity, Subjectivity, Colour

Student: Andrea Panzeri

Motivation

As two geography students, Tobler's first law of geography is one of the most important rules we learned about spatial relationship:

Everything is related to everything else, but near things are more related than distant things.

This applies to many issues in real world. For example, if you map population density of a country, it is highly possible that a dense city is surrounded also by dense cities. We want to examine if this also applies to the cognition world.

In the ESUM (Analyzing trade-offs between Energy and Social performance of Urban Morphologies) project, 32 participants were asked to go through a path with 14 checkpoints and give feedback about their perception. The checkpoints were specifically selected so that they differ from each other. However, we wonder if the participants could feel the difference as planned instead of relating their current perception with their impression from last several checkpoints.

Hypothesis & Research Question

In our project, we investigated if distance affects people's feeling. Based on Tobler's first law of geography, we assume that people's feelings towards a specific place are related to feelings of places nearby. Thus, people's attitudes towards near space are similar.

The concept of distance we used here is not Euclidian distance, but the length of walking distance in the experiment.

Approach & Methods

Similarity is something hard to measure precisely. We used clustering as a reference: objects in one cluster are more similar than those in two different clusters.

We used data from the ESUM project and divided them into 3 groups: survey results which reflect people's subjective perception, biofeedback data which measures people's physical reaction, and environment data which reflects the objective difference. Without offering geometry data, we checked if the cluster distribution conforms to certain spatial patterns.

We did clustering for each group participant by participant rather than aggregating 32 people's data together, mainly because cognition among people as well as dynamic features of environment are different. For example, some people might find it noisy, but others might not evaluate sounds with same decibel value. Some participants walked in the morning, while others might do the same task in the afternoon with significant higher temperature. We also standardized data, since different variable varies in different scale.

For survey results, which are small in sample size, K-means clustering was used and 14 checkpoints were clustered into three classes for each participant. To compare the result, we used a metric to compare if two checkpoints are more similar than another two: count of participants to recognize each two checkpoints into one cluster. For all 91 (14*13/2) checkpoint pairs, the higher the count is (maximal 32, minimal 0), the more similar they are. We checked if checkpoints closer to each other show higher similarity.

For biofeedback data and environment data, which are almost continuous along the path with 666 points, we used DBSCAN with same parameters so the count of clusters would not be predetermined and the clusters are comparable between different dataset. We mapped the results and compared how these two clustering pattern differ from each other for the same participant.

Results & Discussion

□ Subjective feeling

In the first impression, it looks like the first several checkpoints are more into one cluster, and the last several in another, while checkpoints in the middle show more uncertainty. See example of participant No.12 in figure 1. Though, there are also a few exceptions.



Figure 1 Cluster result of participant No.12

Then we calculated the similarity parameters explained before and the results are shown in figure 2. 14 checkpoints were placed on a circle and the width of the chords connecting two checkpoints represents the similarity metric of these two. Note that checkpoint 1 and 14 are the starting and ending points which is the furthest pair although they were placed close to each other. Similarly, 1 and 13, 2 and 14 are also very far away. Then it shows a tendency that checkpoints closer to each other give out a higher

similarity in general. Take checkpoint 11 for example, it's clear that chords by two sides are much wider than chords in the middle.

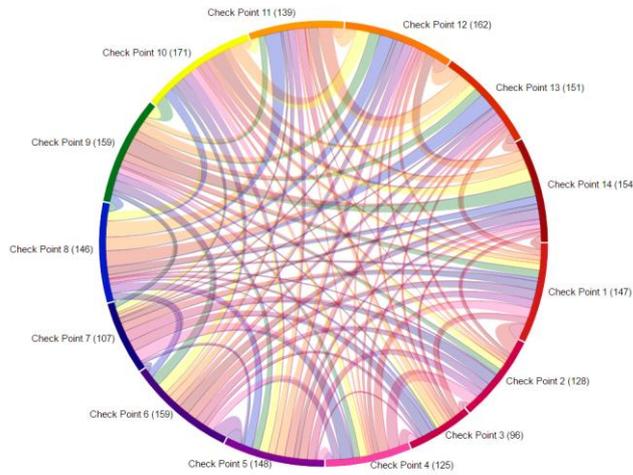


Figure 2: Similarity of subjective feelings

Following the Moran's I index measuring two-dimensional autocorrelation, we also calculated the overall autocorrelation for the similarity metric using following formula:

$$I_i = \frac{\sum_{j=1}^{14} C_{ij} * \exp(-\frac{D_{ij}}{\sum_j D_{ij}})}{\sum_{j=1}^{14} C_{ij}/13}, j \neq i$$

C_{ij} is the similar metric and D_{ij} is the walking distance. It actually uses distance-weighted average of similarity metric divided by arithmetic average. When similar metric is evenly distributed, the I_i would equal 1. I_i larger than 1 illustrates that in general closer points have higher similar metric. The result shows that for most checkpoints it confirms our assumption but only check point 7 and 8.

Table 1 Overall autocorrelation Index for each check point

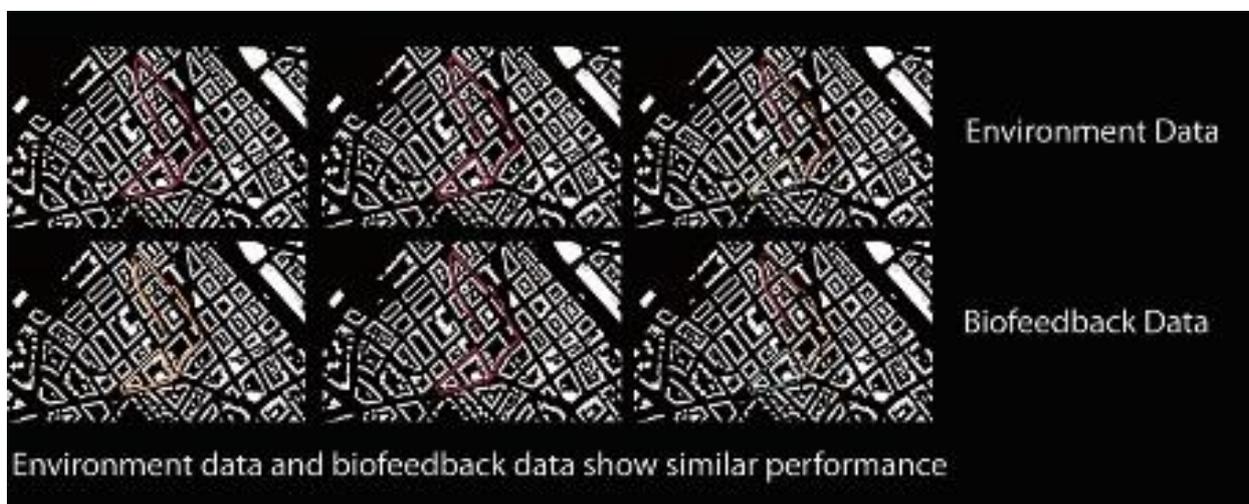
Check Point	1	2	3	4	5	6	7
Overall Correlation	1.87	1.99	1.33	1.35	1.33	1.12	0.49
Check Point	8	9	10	11	12	13	14
Overall Correlation	0.81	1.07	1.15	1.40	1.47	1.47	1.15

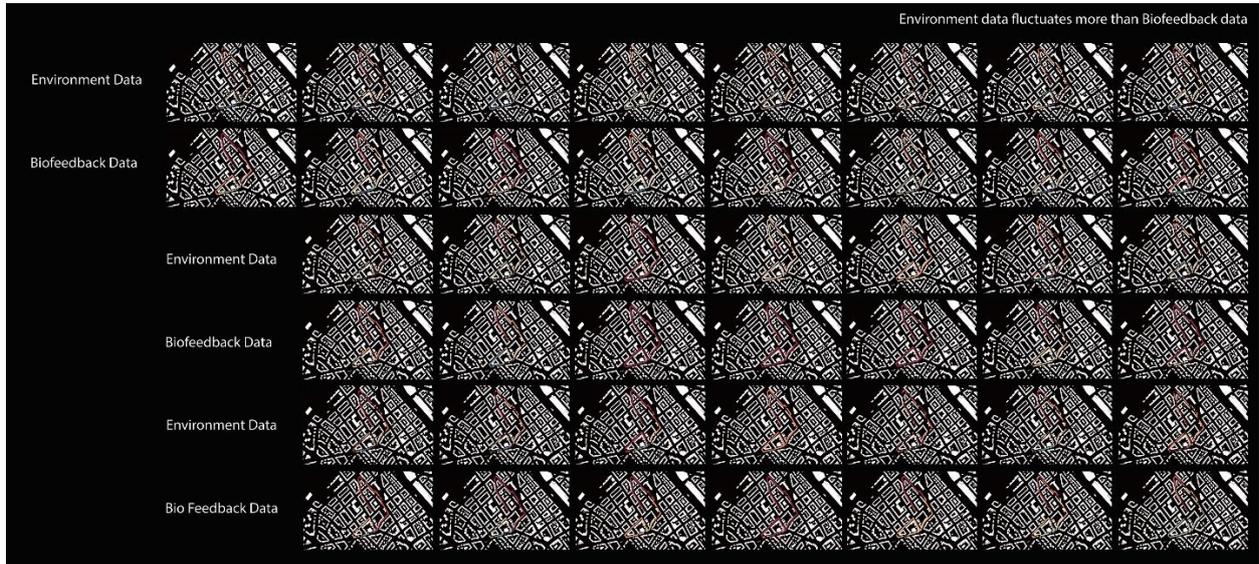
One more thing that worth our attention: checkpoint 9 and 14 are actually the same places (see figure 1 on the map). However, the similar metric is only 24, meaning that only 2/3 respondents feel it similarly when they come back to the exact same place several minutes later. So there must be something affecting people's feelings except the environment of the place itself.

□ Physical reaction and environment difference

Generally Speaking, biofeedback data follows specific spatial pattern, as the closer points tend to be in the same cluster as well as there are less clusters for each participant, even though we do not take spatial information into consideration during clustering; while participants' environment data is opposite, which seems to be more randomly distributed. It makes sense since environment data such as the level of sound, dust, etc. can be suddenly high when a truck pass by and can go back to normal level immediately after it runs away. Therefore the result can be interpreted as people's feelings are relatively stable and change gradually in spite of the sudden change in environment.

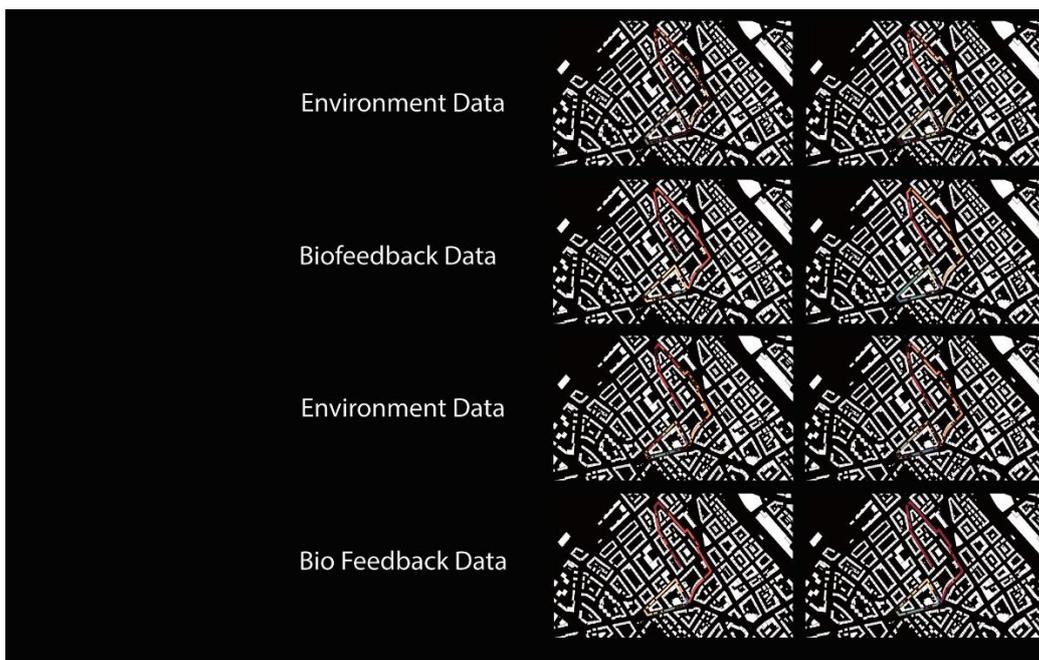
Furthermore, the comparison between participants' biofeedback data and environment data can be summarized as three groups, namely, biofeedback data fluctuates more than environment data, biofeedback data and environment data conform to similar pattern, and biofeedback data fluctuates less than environment data. As we expected, the first two pattern exist but only for minority. Interesting is that, even in these cases, actually biofeedback data still conforms to specific spatial pattern. The possible reason is probably the too stable environment rather than sudden changes in biofeedback.





The pattern “environment data fluctuates more” is dominant, as 22 out of 29 participants’ show this pattern, which proves our hypothesis more or less that biofeedback data possess spatial similarity. As some examples shown below, closer space is highly likely to be classified in the same cluster even though the corresponding environment does not show spatial correlation.

Therefore, we can confirm that people react to environment changes in a more smooth way. Besides, there are always more outliers, which is represented as black point along path in the graph, in environment data which also proves its instability.



Conclusions

Through our three groups of clustering, we concluded that in general, people's feelings towards a specific place are related to feelings of places nearby. People tend to feel that close places are similar both physically, which is proved by biofeedback data, and mentally, which is proved by subjective survey data. Not only your brain feels that way, so do other parts of your body.

However, the conclusion does not apply to all participants in the experiment. The validity is also limited by the sample size. We only confirmed that based on the ESUM project data, most people's feelings are also affected by distance besides the actual environment. Further study on larger sample is needed to validate our conclusion.

References

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1), 234-240.

Scikit-learn(2011): Machine Learning in Python, Pedregosa et al., *JMLR* 12, pp. 2825-2830

Urban Perception: A deep learning approach

Student: Jack Kenning

Motivation

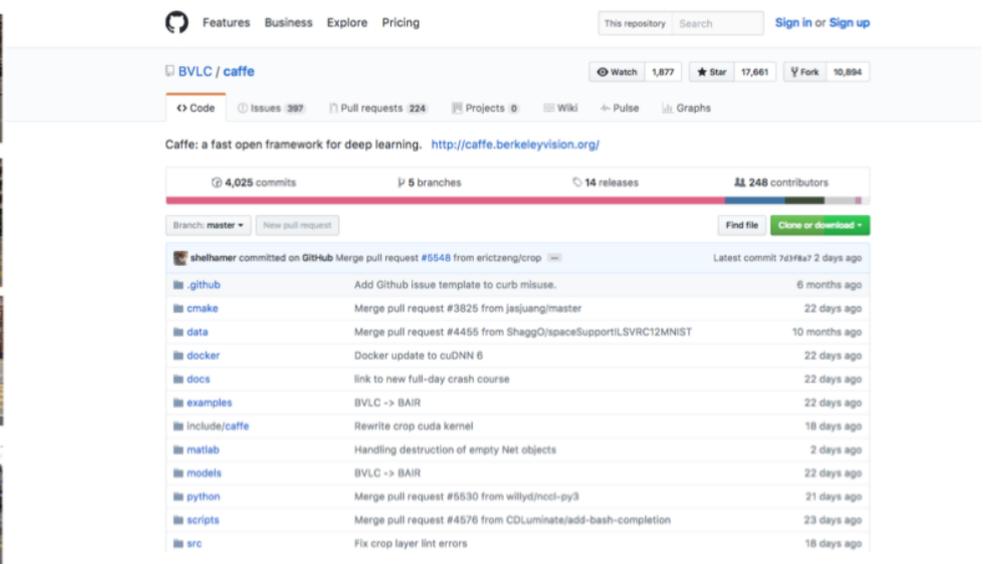
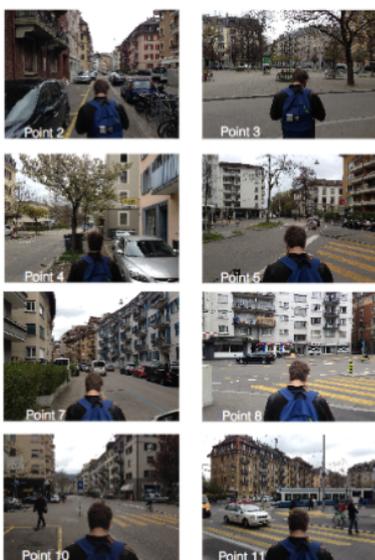
The creative data mining course leads students through the process of data mining with a specific question-based approach. The collection, cleaning, selection of relevant data is followed by a transformation process and ends with a visualization step, making the results visible. From the ESUM Data Set (Energy and Social Performance of Urban Morphologies) collected by the Information Architecture chair of ETH Zürich, there were survey responses from specific path-points in the city of Zürich. The survey asked respondents at each path-point a series of questions, such as: “How beautiful is this location?” or “How ordered do you find this location?”. One popular approach is to compare the subjective survey responses to an objective measure that could guide a more global urban design process.

Hypothesis & Research Questions

One urban element that seemed to be under evaluated was the presence of vegetation and green areas. We can compare this to how beautiful, interesting or ordered that people found a specific point. “Trees make an urban situation more beautiful.” Do they also make it more interesting and/or more ordered?

Approach & Methods

The ESUM Project Data was a valuable source, but it was necessary to collect other data to evaluate the presence of vegetation in an urban context.

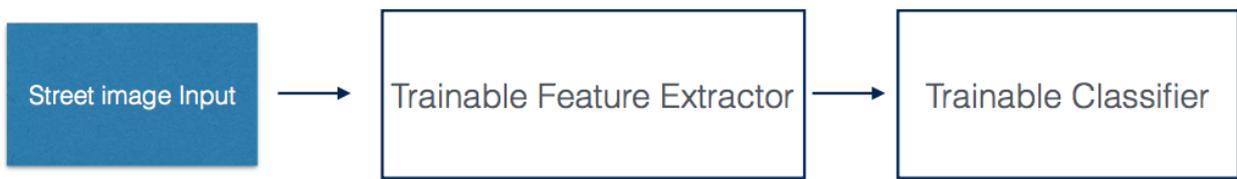


Source: Caffe - Berkeley Vision and Learning Centre <http://caffe.berkeleyvision.org/>

The presence of vegetation is not something that is easily measured, but a logical approach is to use an image-based method. This also makes sense as the research question directly relates to an individual's perception. However, it is not possible to simply identify vegetation by the color of a pixel for example, because many things in an urban context are green and this would result in a large error.

Deep learning based on convolution neural networks makes it possible to quite easily accomplish this task. Through an established framework, we can train the framework to recognize certain features in an image - in this case vegetation (trees, bushes, grass, etc.).

Caffe is a framework developed by the Berkeley Vision and Learning Centre and was used in this example. The advantage is not requiring much coding knowledge to implement the framework. Below is the proposed workflow.



The extractor can learn to identify certain features corresponding to what we are trying to identify from the training set that we will use and then the classifier makes it possible to identify vegetation on any urban image. These can both be set up in Caffe.

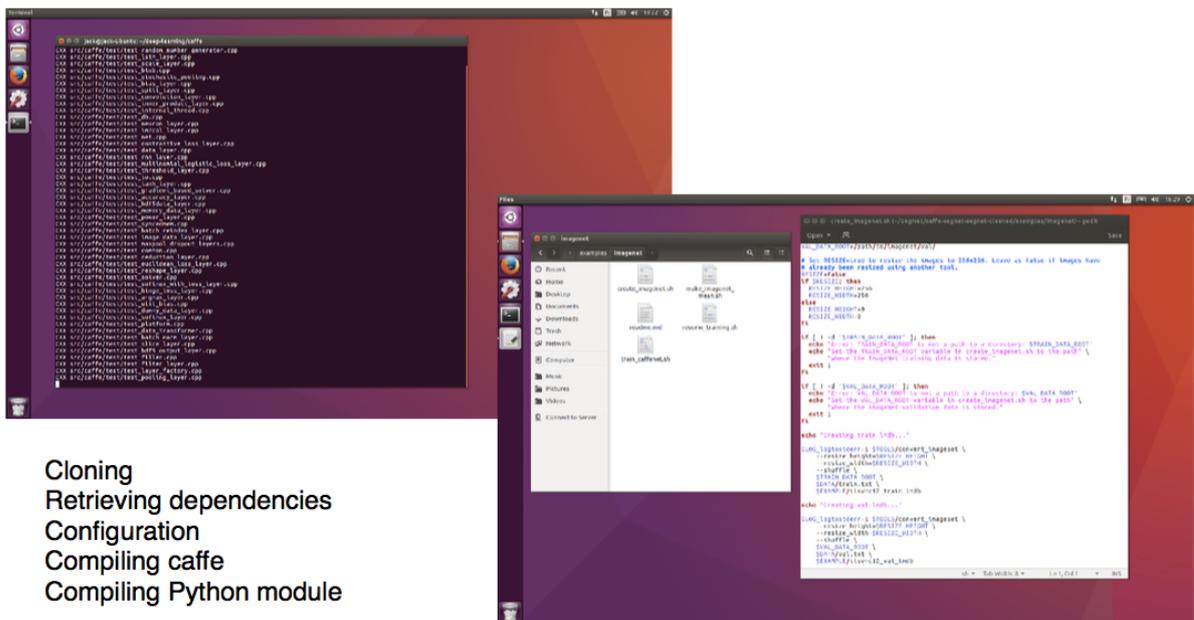


Figure 3: Caffe setup

In this case, the project relied largely on SegNet, a modified version of Caffe, developed by Cambridge University. Hand-labeled images are used to generate a predictive model that can segment an image into different urban elements such as buildings, roads, pavements, trees, vehicles etc. There are 12 different categories that SegNet can be trained to identify from the labeled set, these were the categories used for the analysis, more specifically the tree elements.

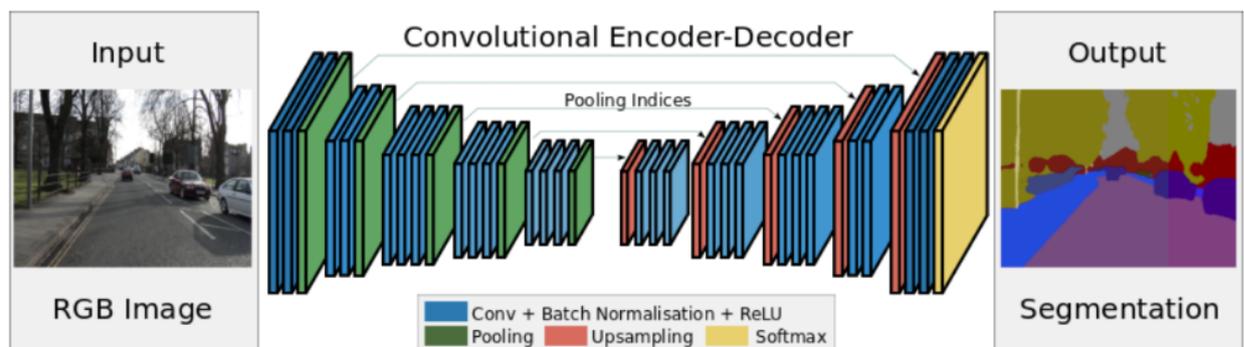
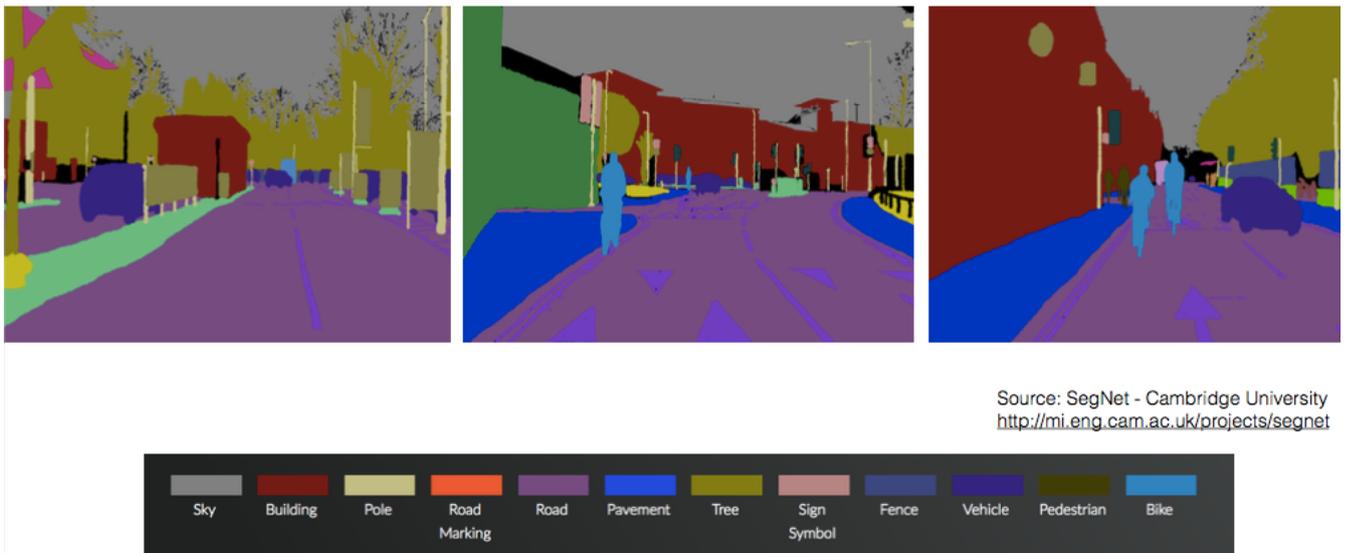


Figure 4: SegNet, Cambridge University

The pre-trained model runs the classification pixel by pixel based on a multitude of input factors, but ultimately outputs a segmented image that can then be used as an objective source of urban data.

Results & Discussion

In order to take into account peripheral vision as someone walks through a neighborhood, we ran the segmentation on three images, taken at eye-level from a vertical position.

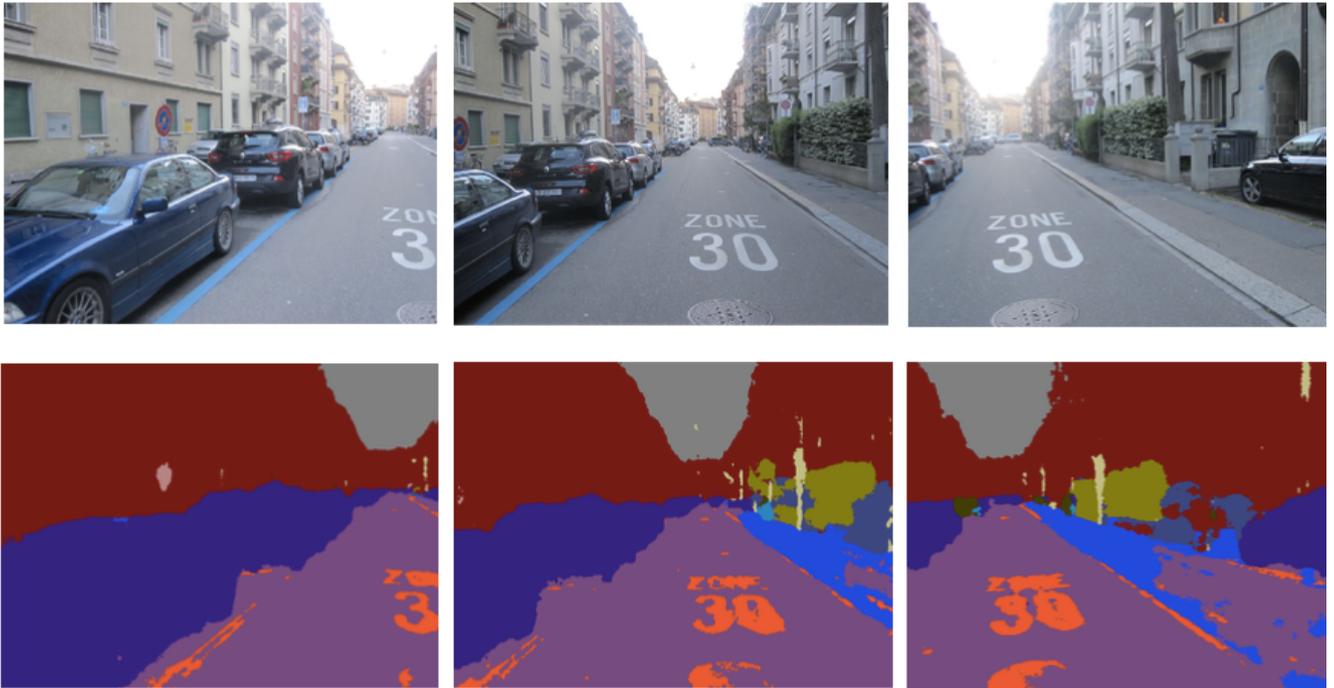


Figure 5: Segmented images from Zürich survey path

After this segmentation was done, it was relatively simple to read the number of pixels of each color with a python script and establish the percentage of image corresponding to a given category. The first assumption was to plot the percentage of vegetation to different survey responses and to analyze the output:

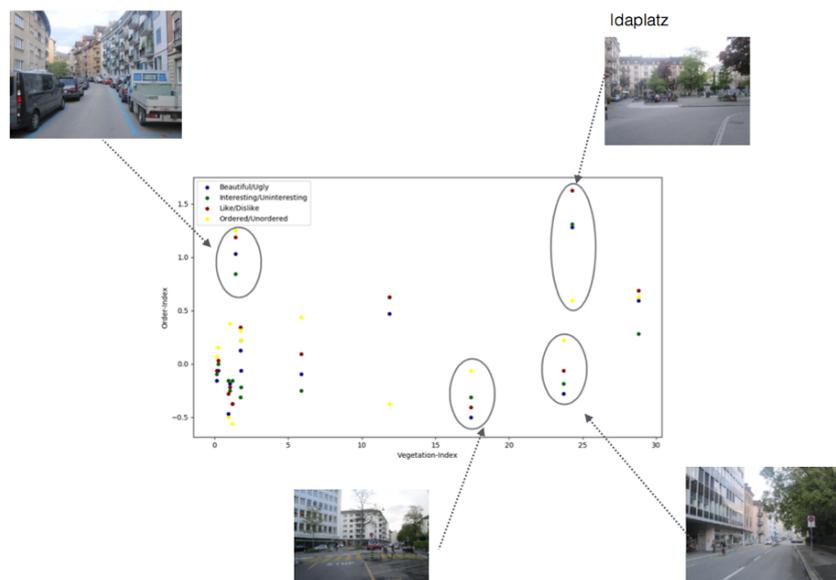


Figure 6: Survey Responses plotted against the segmented tree percentage

The output had a general coherence but a lot of specific cases which didn't fit to the trend of a place being more beautiful with a larger presence of vegetation. For this reason, it made more sense to move back towards a more general analysis where all the different categories could be seen. The expected output would be to see an increase in the "vegetation" part of the image in places that the survey respondents found more beautiful. As the graph below shows, that was not necessarily the case.

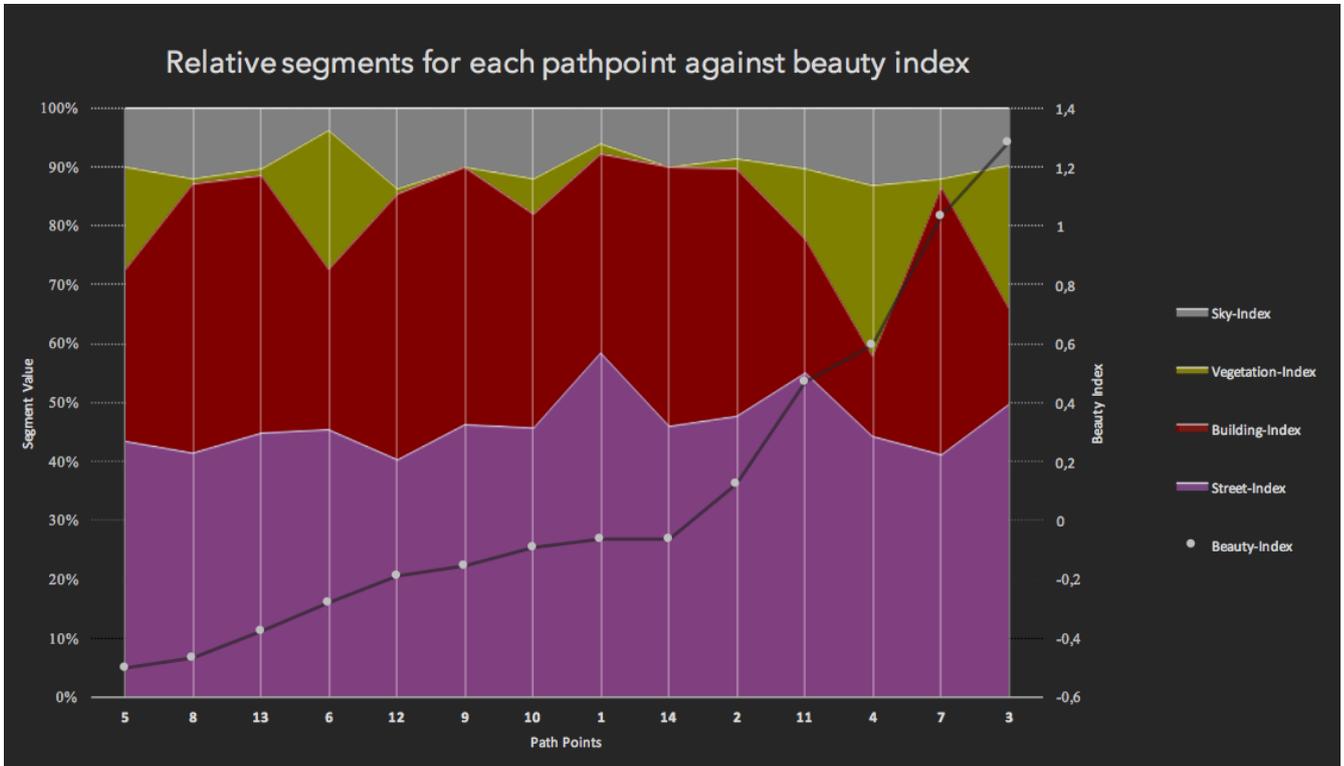


Figure 7: Relative importance of urban elements plotted to survey beauty index

The stacked surfaces represent the relative presence of the street, the buildings, the trees and the sky, but it doesn't appear that any of these have a direct influence on how beautiful people found a given location. To be sure of the output, the same graph was generated relative to the questions of "Like-Dislike", "Interest-Boring", "Order-Disorder".

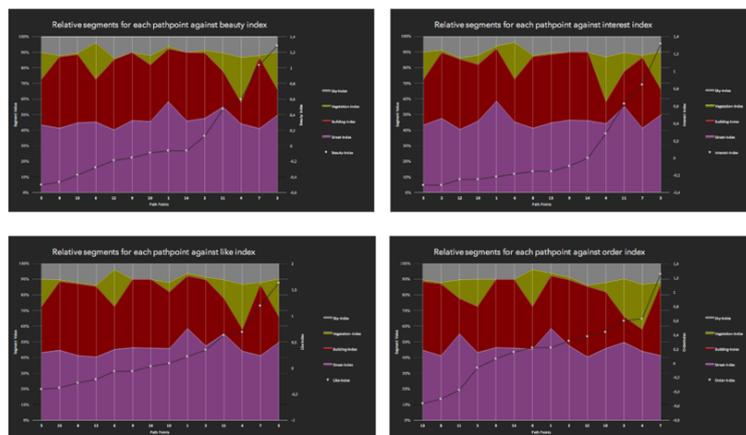


Figure 8: Relative importance for different survey questions

Conclusions

The hypothesis and research questions seem to indicate that the factors leading to a place being described as “beautiful”, “ordered”, or “likeable” are more complex than just the presence of vegetation in a given place. Multiple factors may have influenced these results, such as the specific location of the photos compared to the position of the survey respondents which were not always very precise. There were also minor (and major) modifications to the urban environment with one tree having been chopped down between the survey and the photo analysis. The sample-size may also have induced error in these findings, as only around 30 survey responses were used. With a larger data set it may be possible to extract more meaningful results.

It seems reasonable to imagine that the presence of trees makes a place-point more beautiful but this does not directly influence how beautiful a place is. The characteristic appears to be a lot more complex than solely quantitative. However, the first data used from the segmented image was an average of the number of pixels representing vegetation in the photo. This was considered quite a poor measurement in relation to the other information contained in a segmented image, so in the second visualisation the relative importance of segments is taken into account. For further analysis, maybe an indicator such as the distribution across the image can be looked at.

References

Caffe - Berkeley Vision and Learning Centre <http://caffe.berkeleyvision.org/>

SegNet - Cambridge University <http://mi.eng.cam.ac.uk/projects/segnet>

NVIDIA Caffe Course <http://on-demand.gputechconf.com/gtc/2015/webinar/deep-learning-course/gettingstarted-with-caffe.pdf>

ESUM Project - ETH Zürich <http://www.ia.arch.ethz.ch/esum/>

Street Cross Section

Student: Hong Liang

Motivation

Street cross section design is an important part of urban planning. In most cases, it depends on the road grade, and the property and function of the road. Once the road grade and capacity are confirmed, the type of street cross section is roughly decided. Street cross section offers functional services to pedestrians, on the other hand, different fractions of various street cross section elements may create different atmospheres, which can also affect a pedestrian's feeling and behavior. But how does it work?

In this project, with the ESUM data set and some suitable analysis methods, the relationship between street cross section and a pedestrian's reaction will be explored and discussed.

Hypothesis & Research Question(s)

The main question of this project is to explore the effect of street cross section on pedestrians. The elements of street cross section mainly include drive lane, parking lane, bike lane and side walk. So the more specific question is: which element plays a more important role in affecting a pedestrian's feeling?

For example, parking lane could be a negative factor. The wider the parking lane is, it could make pedestrians feel more insecure and chaotic.

Approach & Methods

Data Collection and Preparation

In order to describe the street cross section along the path, I measured every street cross section element on 14 street points on Google Maps, and compared them with the photos. The collected data and street cross section diagrams are as follows:

Table 1 street cross section data of 14 points

point	street width	sidewalk width	sidewalk rate	parking lane width	parking lane rate	trees	crossing wa
1	12	3	0.2500	4.2	0.3500	1	none
2	12	3	0.2500	4.2	0.3500	1	none
3	4.2	1.2	0.2857	0	0.0000	1	none
4	12.6	3.6	0.2857	0	0.0000	1	none
5	12.4	4.3	0.3468	0	0.0000	0	zebra crossing
6	13.5	4.4	0.3259	2.1	0.1556	0	zebra crossing
7	11.3	3.6	0.3186	4.2	0.3717	1	none
8	16.8	3	0.1786	0	0.0000	1	zebra crossing
9	13.8	3.6	0.2609	4.2	0.3043	0	none
10	11.7	3.6	0.3077	2.1	0.1795	1	zebra crossing
11	18.5	4	0.2162	0	0.0000	0	zebra crossing
12	19.3	3.6	0.1865	0	0.0000	0	zebra crossing
13	15	3.6	0.2400	2.1	0.1400	0	zebra crossing
14	13.8	3.6	0.2609	4.2	0.3043	0	none



Figure 1 street cross section diagram of 14 points

As shown in the diagram above, due to the different street grade and function, the street cross section elements at 14 points are different. In particular, street width and parking lane are the primary differences among 14 points. At the same time, side walk is the most important part of street for pedestrian. Therefore, I chose street width (width between two buildings), parking lane fraction (parking lane width / street width) and side walk fraction (side walk width / street width) to describe the property of street cross section. Some binary variables, such as whether there are trees or zebra crossings are also taken into consideration.

To describe pedestrian's feeling, I chose 3 questions from the survey result of ESUM: like or dislike, ordered or chaotic, secure or insecure. ESUM experiment was taken by 37 participants. In order to compare the results with the street cross section's property at 14 points more easily, first, I calculate the average value for 3 questions at 14 points.

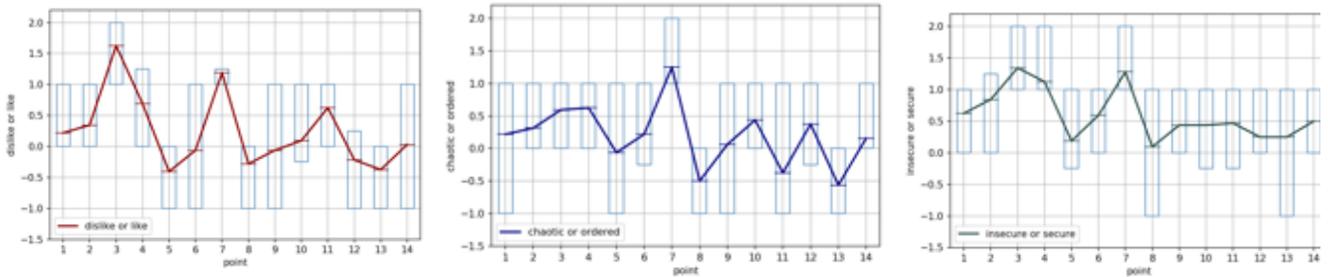


Figure 2 survey result

As shown in the diagrams above, the results of 3 survey questions present similar tendency. It can be also concluded from the analysis, that the deviation at some points (e.g. point 7) is relatively small, which means the average value is more credible, while the feelings of pedestrian at point 5 and 6 are more disperse.

Data Analysis Method

Data analysis is divided into 4 steps.

1. Average value

In first step, I compare the calculated 3 average values of survey results with 3 street cross section properties (street width, parking lane fraction and side walk fraction) in turn, to get the first impression of the relationship between them.

2. Cluster

I choose k-mean as the cluster method to analyze the relationship between 3 average values of survey results with 3 street cross section properties. In this step, some semiquantitative results can be concluded.

3. Multi-variable linear regression

In multi-variable regression, two more binary variables (trees and zebra crossing) are also taken into consideration. For each survey result, we will get a formula consisting of 5 street cross section variables. And according to its coefficient, we can get which variable's effect is more significant.

4. Simple linear regression

Once we get the most important factor for each survey result, we can do the simple linear regression to test its significance and credibility again.

Results & Discussion

1. Average value

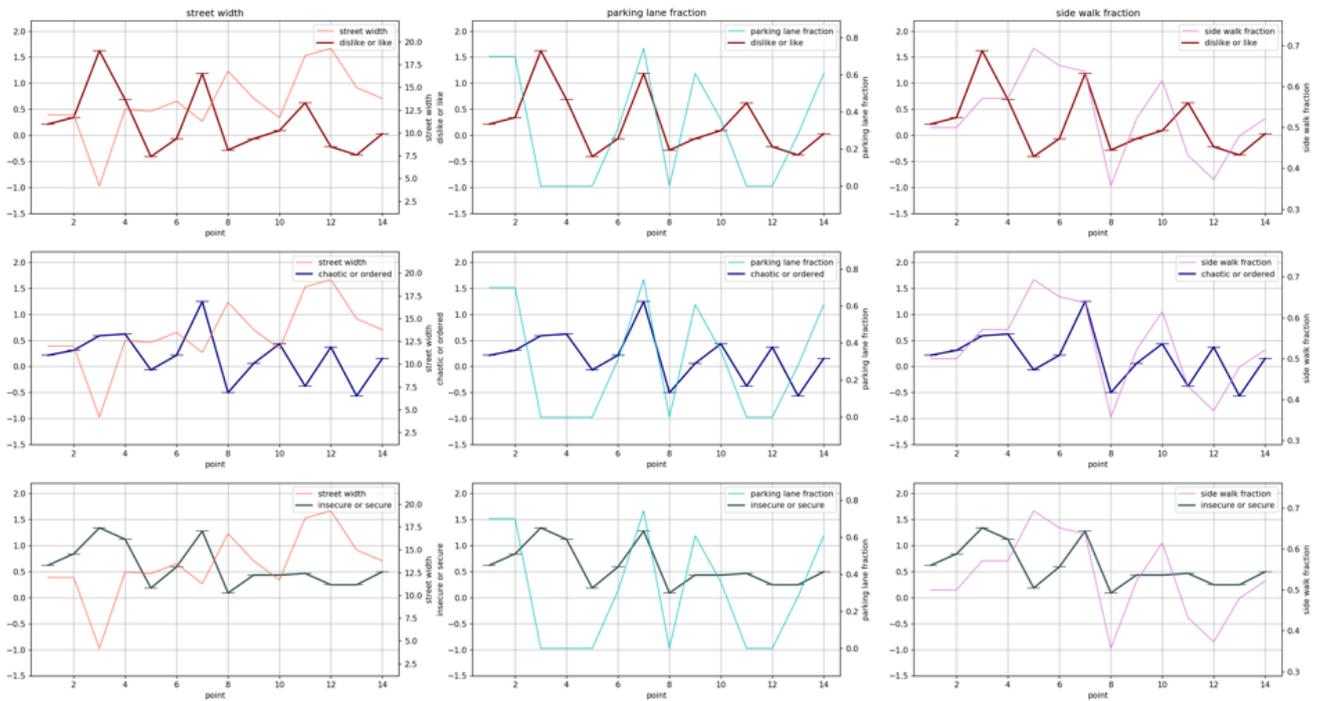


Figure 3 comparison between survey result and street cross section property

According to the diagrams, it is more or less clear that there is correspondent relationship between side walk fraction and pedestrian's feeling in all 3 survey results (like, ordered and secure). The biggest difference occurs at point 5. But point 5 is an intersection, and the street cross data can only describe on street. That could be the reason to explain that big gap. Additionally, the deviation of survey results at point 5 is relatively big, it could also result in this difference.

2. Cluster

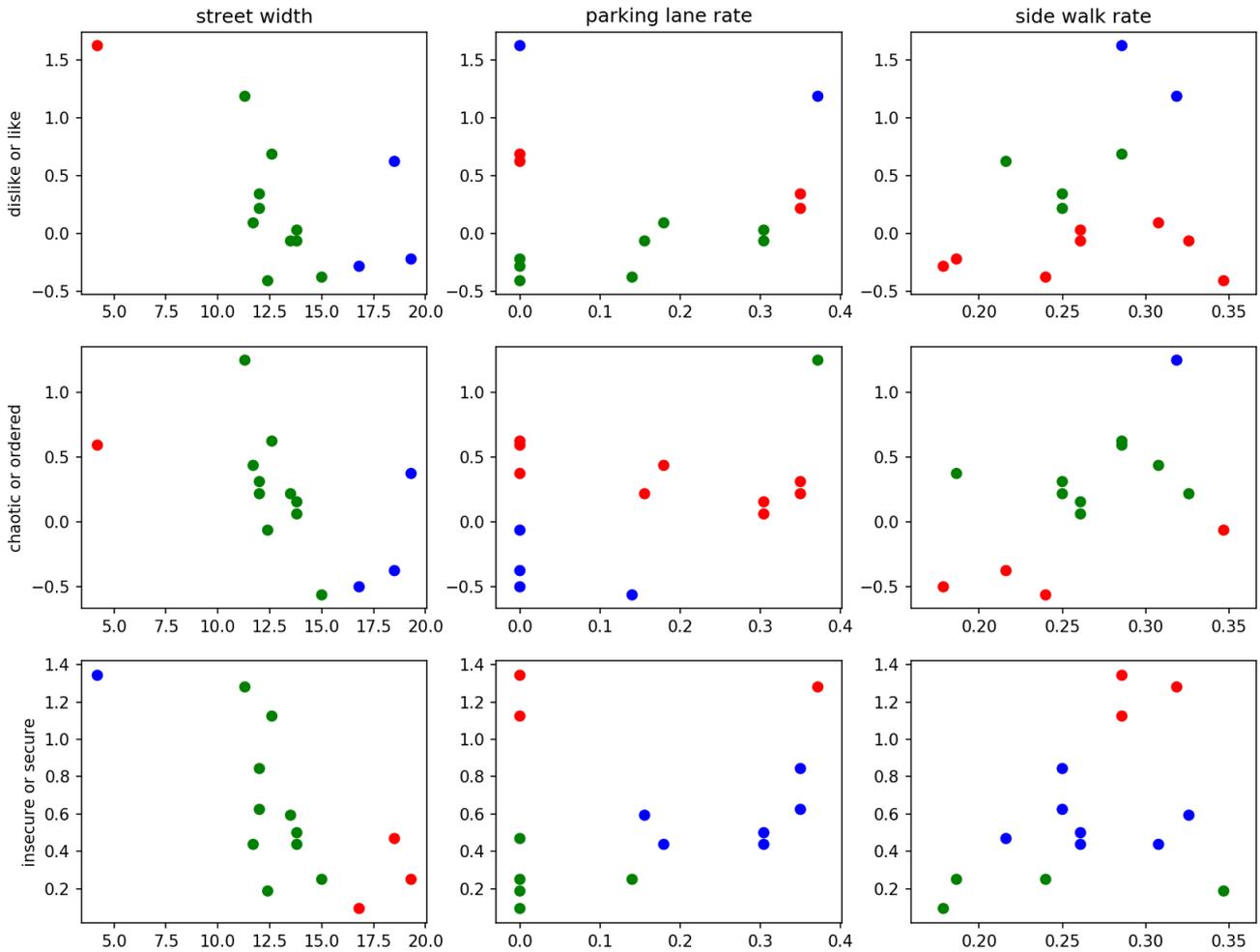


Figure 4 cluster analysis

Figure 4 cluster analysis

Due to the limitation of the amount of data set, the results of cluster analysis are not very significant. But we can still get the similar semi-quantitative result that low side walk fraction will result in more negative feeling, while the pedestrian is more likely to have a positive evaluation for those points which have high side walk fraction.

In additional, there is also some obvious cluster between parking lane fraction and secure. It seems that higher parking lane fraction leads to an insecure feeling for pedestrian. If the experiment could have more participants and cover more streets, cluster analysis will have more credible conclusion.

3. Multi-variable linear regression

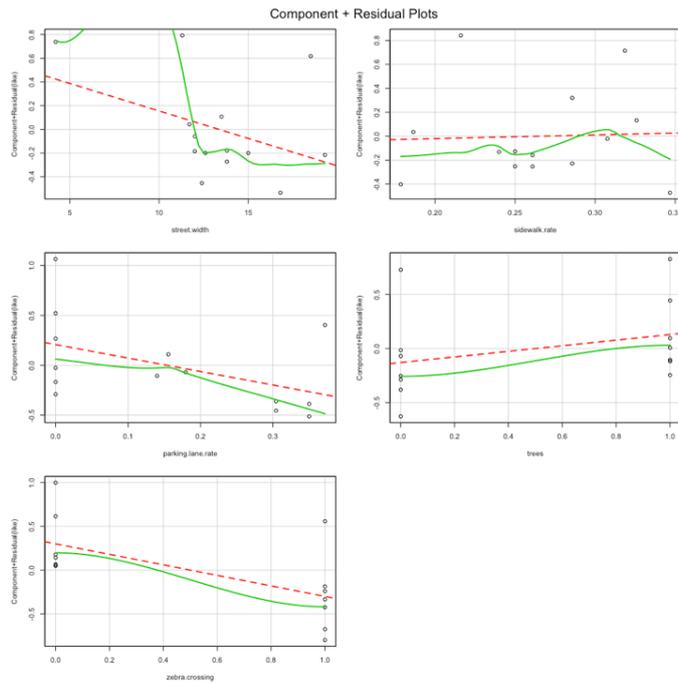


Figure 5 multi-variable linear regression for 'like or dislike'

Like =

$$\begin{aligned}
 & - 0.046 * \text{street width} \\
 & + 0.312 * \text{sidewalk fraction} \\
 & - 1.350 * \text{parking lane fraction} \\
 & + 0.259 * \text{trees} \\
 & - 0.600 * \text{zebra crossing}
 \end{aligned}$$

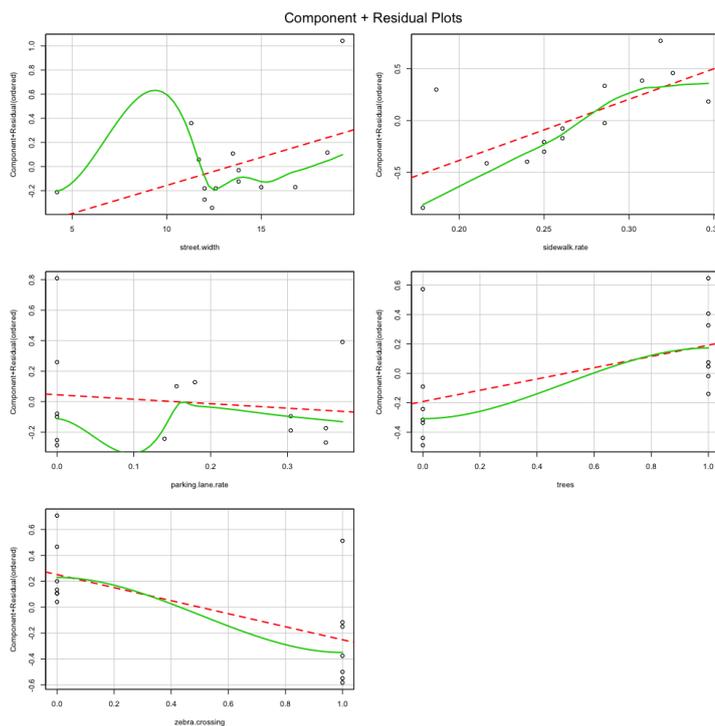
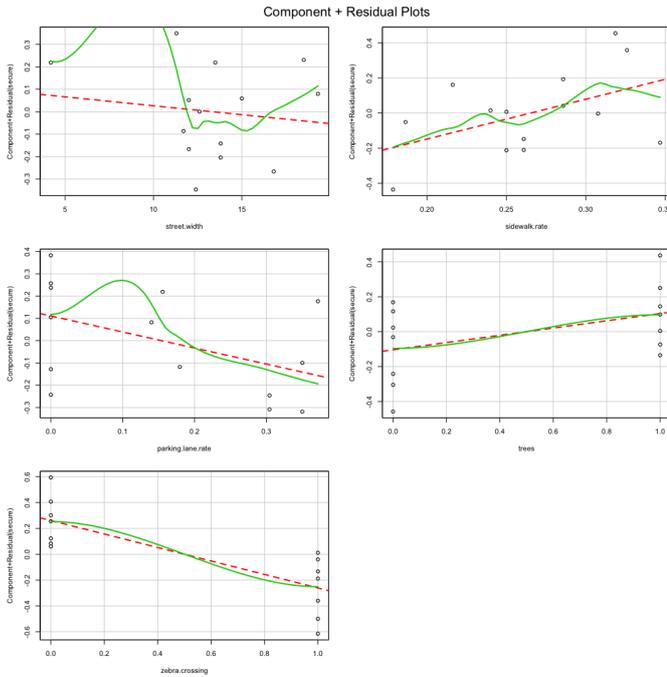


Figure 6 multi-variable linear regression for 'ordered or chaotic'

Ordered =

$$\begin{aligned}
 & + 0.046 * \text{street width} \\
 & + 5.898 * \text{sidewalk fraction} \\
 & - 0.293 * \text{parking lane fraction} \\
 & + 0.383 * \text{trees} \\
 & - 0.503 * \text{zebra crossing}
 \end{aligned}$$



$$\begin{aligned}
 \text{Secure} = & \\
 & - 0.008 * \text{street width} \\
 & + 2.269 * \text{sidewalk fraction} \\
 & - 0.720 * \text{parking lane fraction} \\
 & + 0.208 * \text{trees} \\
 & - 0.521 * \text{zebra crossing}
 \end{aligned}$$

Figure 7 multi-variable linear regression for 'secure or insecure'

Table 2 coefficients of multi-variable linear regression

	street width	sidewalk fraction	parking lane fraction	trees	zebra crossing
like	-0.046	0.312	-1.350	0.259	-0.600
ordered	0.046	5.898	-0.293	0.383	-0.503
secure	-0.008	2.269	-0.720	0.208	-0.521

From the table and diagrams, we can get the conclusion that side walk fraction and trees have and positive effect on pedestrian, because all the coefficients are positive. On the other hand, parking lane fraction and zebra crossing effect pedestrian in a more negative way.

4. Simple linear regression

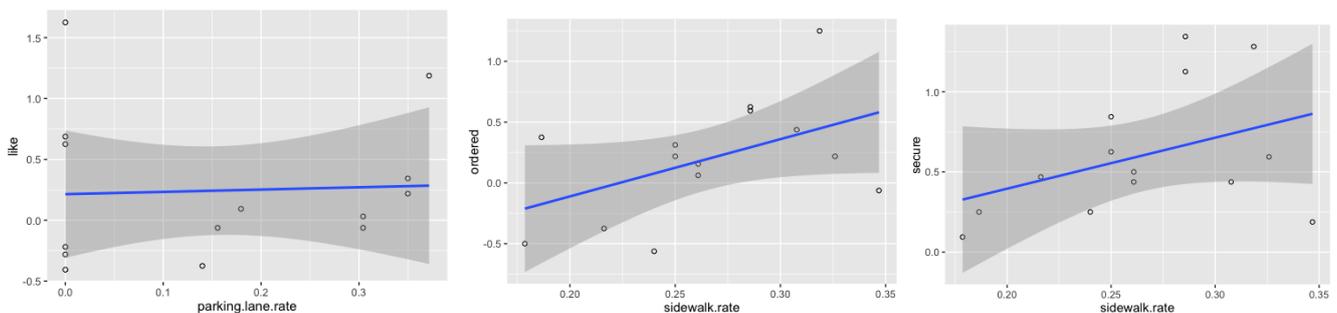


Figure 8 simple linear regression

From last analysis, we find the most important factor for each question, and test it with simple linear regression. According to the diagrams above, the negative effect of parking lane fraction on 'like or dislike' is not very significant. But the positive effect of side walk fraction is still credible.

Conclusions

After all these analysis, we can now answer the questions at the beginning. The different elements of the street cross section effect pedestrian in a different way. We can find that parking lane fraction has a negative effect on pedestrian, while side walk fraction affect pedestrian in a more significantly positive way.

But to get a more meaningful and credible results, we still need more data. Additionally, the binary variable should be avoided in multi-variable regression. For example, 'trees' variable could be replaced by green area. Or there could be some other methods more suitable for binary variable analysis. That's could be the further step of this project.

References

Marc Schlossberg, John Rowell, Davd Amos, and Kelly Sanford, Rethinking Streets: An Evidence-Based Guide to 25 Complete Street Transformations, 2013.

Security Analysis

Student: Gong Chen & Zhonghau Dai

Motivation

Security level is one important index to evaluate urban planning. Security level can be treated as a citizens' comprehensive subjective reflection on urban surroundings. Therefore, based on the project ESUM- Analyzing trade-offs between Energy and Social performance of Urban Morphologies, various datasets are available. Among there datasets, survey questions are of most interest, which covers twelve questions (Figure 1), including security, noise, beauty evaluation and etc. Hence, there should be a best combination of these parameters to provide us a high urban security.

Checkpoint ID:.....
Atmosphere

dislike like

chaotic	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	ordered
noisy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	quiet
private	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	public
boring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	interesting
crowded	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	empty
insecure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	secure
ugly	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	beautiful
narrow	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	spacious
enclosed	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	open
dark	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	light

Unfamiliar Familiar

Figure 1: Survey questions

Hypothesis & Research Question(s)

Following are questions conveyed during ESUM project, dislike/like, chaotic/ordered, noisy/quiet, private/public, boring/interesting, crowded/empty, insecure/secure, ugly/beautiful, narrow/spacious, enclosed/open, dark/light, unfamiliar/familiar, among which insecure/secure is response variable focused during this project, all these variables are evaluated in -2, -1, 0, 1, 2.

Based on intuition, when a place is quite, it will make people feel secure while when it goes noisy, people tend to feel upset and become feeling insecure. In this case, one hypothesis is made that secure/insecure has a negative (because the high level in questionnaire means quiet) relationship with noisy/quiet in a wide range. And, by searching such relationship, a best range of noise are intended to be found.

Besides, another hypothesis made in project is that people's evaluation truly reflects noise level in decibel.

Approach & Methods

Since dataset is continuous and labeled, regression method is applied in project. Main procedures carried in project are listed following:

1. Decided whether to use noise in decibel level or in evaluation level in survey questions, i.e. does people's evaluation on noise level truly corresponds to decibel level
2. Searching potential variables related with security, both from visualization of data and linear regression method, i.e. single variable linear regression
3. Do multivariable linear regression to acquire a simple linear regression model to quantify security level
4. Check if the above linear model can be simplified further to get the most crucial parameters

Results & Discussion

1. Noise in decibel level versus evaluation level Boxplot of decibel level and evaluation level are both shown below (Figure 2 and Figure 3). Although boxplot provides more information e.g. outlier, percentile, it is hard to tell relationship.

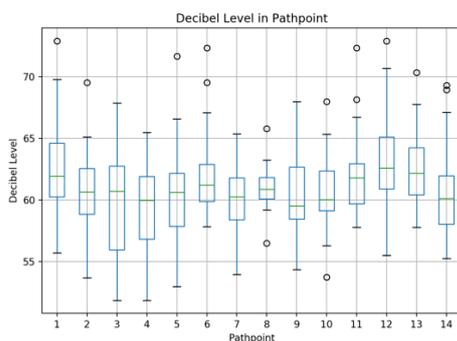


Figure 2: Decibel level

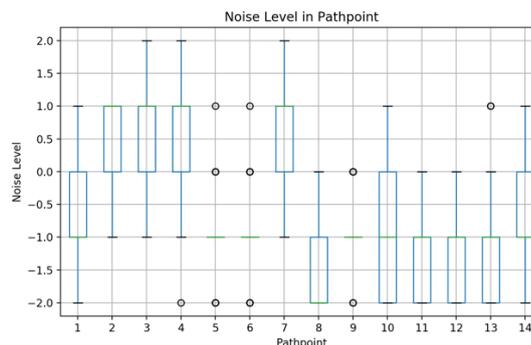


Figure 3: Noise level

Then two boxplots are shown in Figure 4 in the same axis, and mean values of these two data are shown in Figure 5. Important thing here is that in question survey, -2 means noisy and 2 means quite, so noise level is replaced with its opposite number. Basically, from Figure 4 & 5, evaluation on noisy/ quite level corresponds to decibel level obtained by sensor.

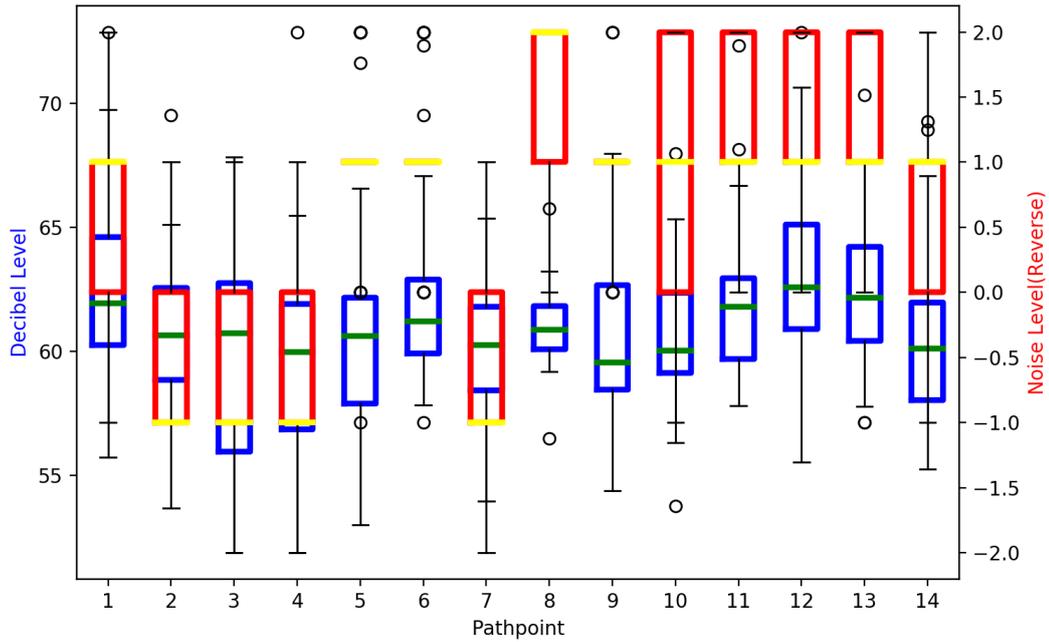


Figure 4: Decibel versus noise level

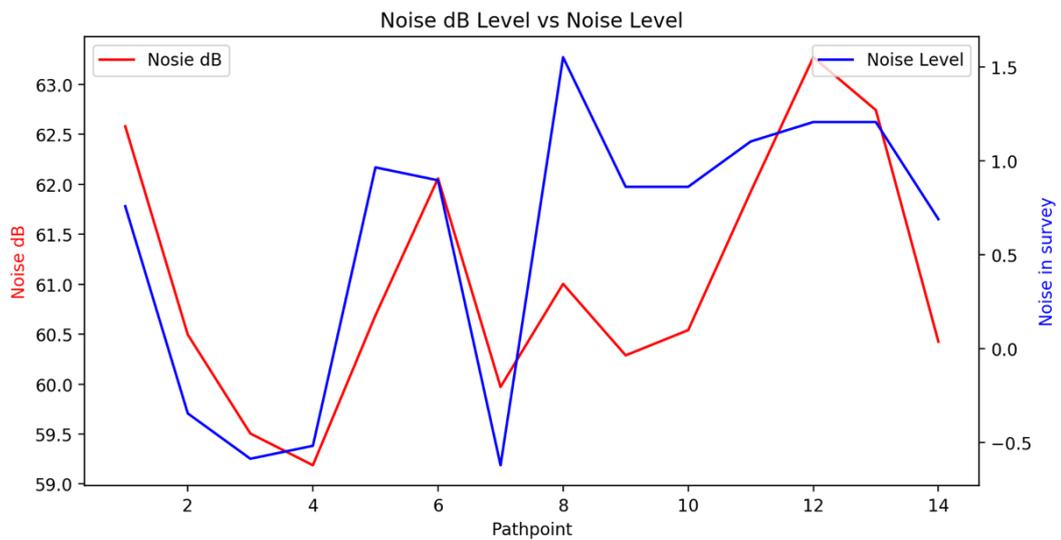


Figure 5: Mean value

2. Potential variable searching

2.1 Visualization

Following figures show mean value of 11 questions mentioned above versus security evaluation level at 14 path points (Figure 6), from which the most related variables are selected: dislike/like, chaotic ordered, noisy/quiet, boring/interesting, crowded/empty, ugly/beautiful, which means these variables determines insecure/secure level significantly.

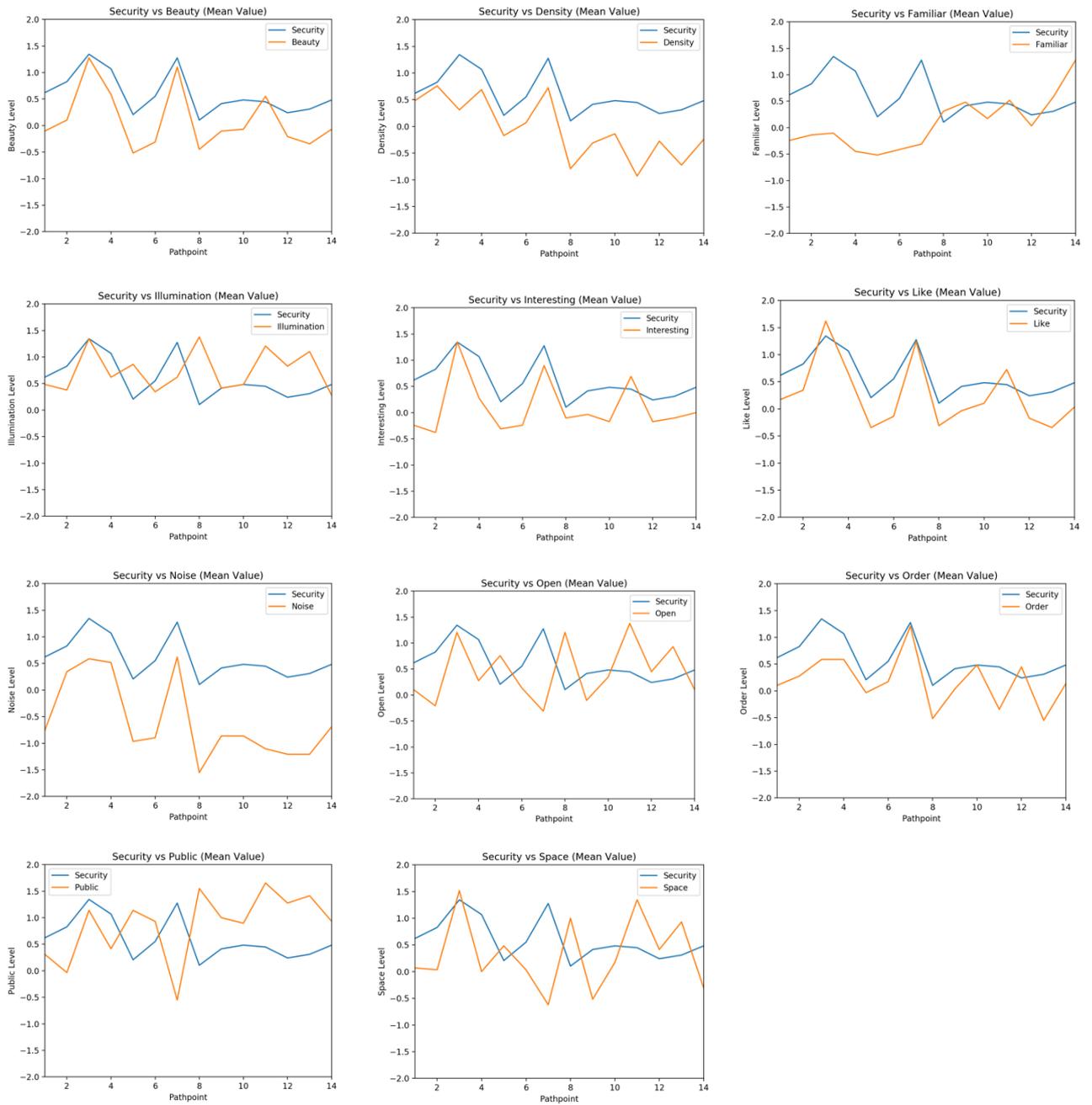


Figure 6: Mean value of each path point

Also, scatter plot methods can be employed here to show linear relationship between these variables and security more clearly, for example shown in figure 7.

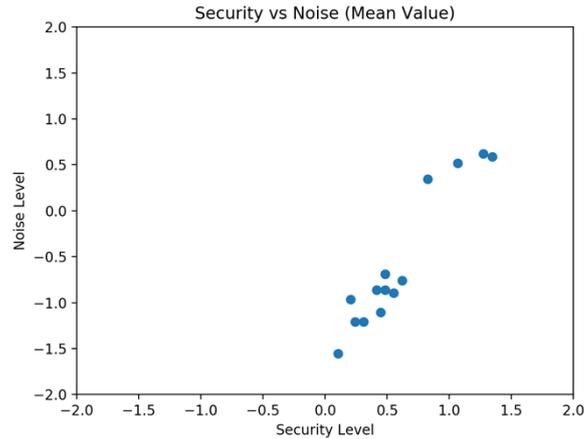


Figure 7: Scatter plot

2.2 Single variable linear regression

Apart from mean value visualization of 14 path points, single variable linear regression is also applied here to find the coefficient between each of these 11 survey questions and security level. In these case, all datasets are employed instead of mean values only in section 2.1. The results together with coefficient and R-square values are shown in Figure 8.

The same result is given: dislike/like, chaotic/ordered, noisy/quiet, boring/interesting, crowded/empty, ugly/beautiful, these 6 variables are most likely to determine insecure/secure level.

Notes: Since there are only 5*5 combinations in each subfigure, overlaps are inevitable, thus bigger the circle is, more overlaps there are.

3. Multivariable linear regression

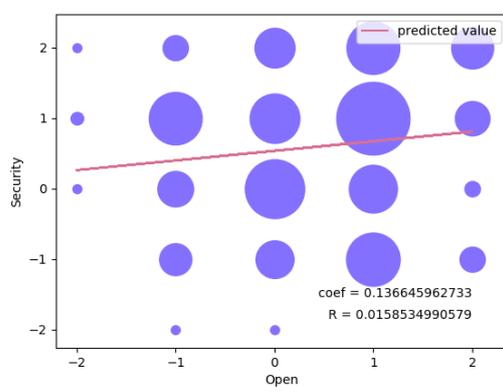
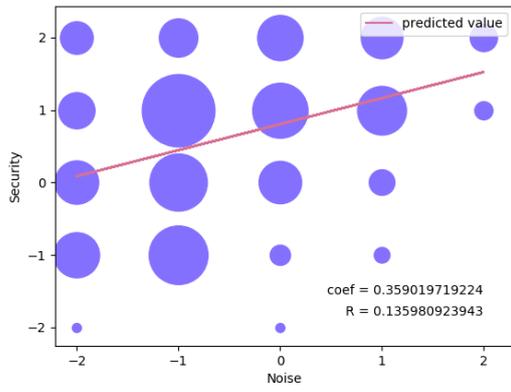
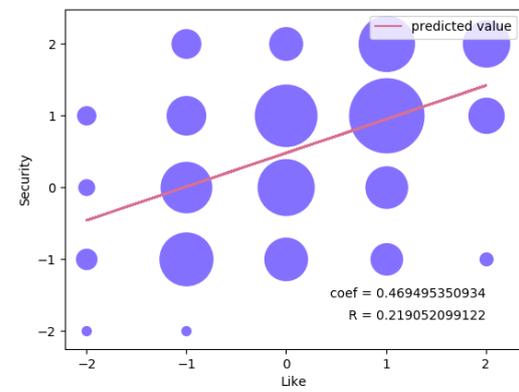
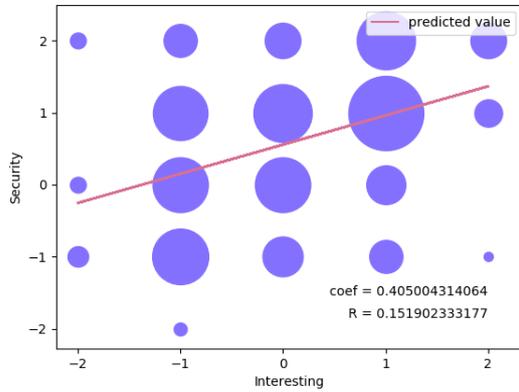
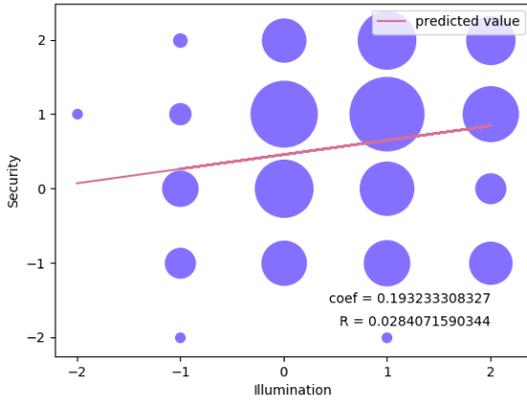
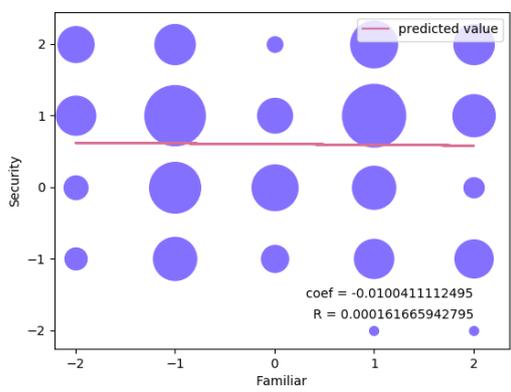
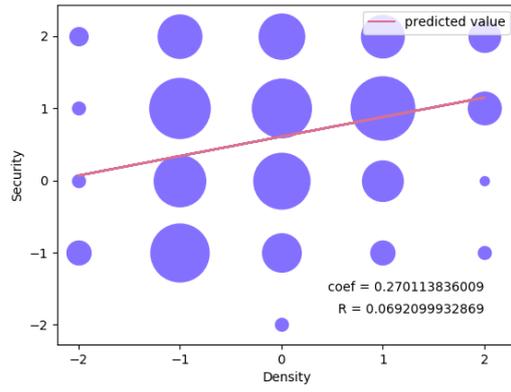
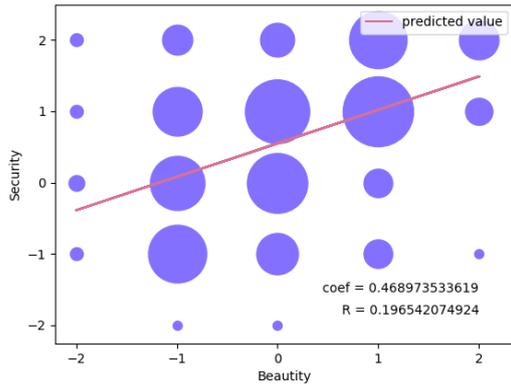
First, linear regression with 6 dependent variables and security level as response variable is accomplished with its corresponding model:

Security level

$$= 0.53 + 0.0685 * \text{noise} + 0.2211 * \text{order} + 0.1121 * \text{beauty} + 0.0975 * \text{density} + 0.1670 * \text{like} + 0.1473 * \text{interesting} \quad (1)$$

Note: Abbreviation is used here: like = dislike/like, order = chaotic/ordered, noise = noisy/quiet, interesting = boring/interesting, density = crowded/empty, beauty = ugly/beautiful.

In this case, the R-squared value is 0.323, which is satisfying since when do linear regression of all other parameters i.e. all questions except security level, the value is 0.339. The accuracy is not destroyed while 5 parameters are dropped, making the model much simple and clear.



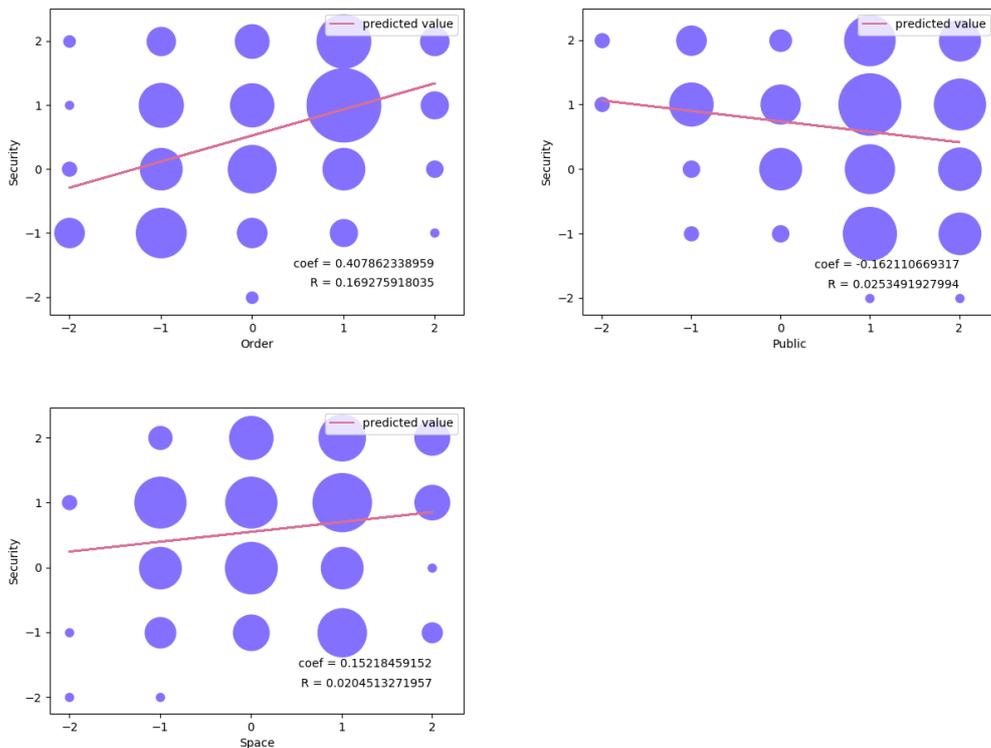


Figure 8: Single variable linear regression

From this equation, noise level does not have much to do with security compared with resting parameters, in next step, feasibility of model simplification will be checked.

4. Model simplification

In this section, the model is checked to see if the above linear model can be simplified further to get the most crucial parameters, basically, R-squared mean is applied to check the accuracy of model. In this case, the noise is dropped, and the new model is followed:

Security level

$$= 0.49 + 0.2285 \cdot \text{order} + 0.1312 \cdot \text{beauty} + 0.1209 \cdot \text{density} + 0.1312 \cdot \text{like} + 0.1469 \cdot \text{interesting} \quad (2)$$

Meanwhile, R value drops only by 0.002 i.e. 0.625%, so this model is still effective.

Conclusions

According to equation (1), the security level does have relation with noise, but a positive coefficient is shown in equation (2) which means people feels more secure when it is quiet. So a wrong hypothesis is made although security and noise level are correlated, but the relationship is negative. However, compared with other five parameters in equation (2), noise level is negligible and order level, i.e. the place is chaotic or ordered, is dominates in security level. If security level is intended to be increased, one should make the place more ordered.

