QUANTITATIVE ANALYSIS ON ARCHITECTS USING CULTUROMICS

Pattern Study of Prizker Winners Based on Google N-gram Data

NAN BAI\(^1\) and WEIXIN HUANG\(^2\)
\(^{1,2}\)Tsinghua University, Beijing, China
\(^1\)bn16@mails.tsinghua.edu.cn  \(^2\)huangwx@mail.tsinghua.edu.cn

Abstract. Quantitative studies using the corpus Google Ngram, namely Culturomics, have been analyzing the implicit patterns of culture changes. Being the top-standard prize in the field of Architecture since 1979, the Pritzker Prize has been increasingly diversified in the recent years. This study intends to reveal the implicit pattern of Pritzker Winners using the method of Culturomics, based on the corpus of Google Ngram to reveal the relationship of the sign of their fame and the fact of prize-winning. 48 architects including 32 awarded and 16 promising are analyzed in the printed corpus of English language between 1900 and 2008. Multiple regression models and multiple imputation methods are used during the data processing. Self-Organizing Map is used to define clusters among the awarded and promising architects. Six main clusters are detected, forming a 3×2 network of fame patterns. Most promising architects can be told from the clustering, according to their similarity to the more typical prize winners. The method of Culturomics could expand the sight of architecture study, giving more possibilities to reveal the implicit patterns of the existing empirical world.

Keywords. Culturomics; Google Ngram; Pritzker Prize; Fame Pattern; Self-Organizing Map.

1. Introduction

Being the top-standard prize in the field of Architecture since 1979, the Pritzker Prize presents the award to an outstanding architect each year for his or her special contribution. In this honor hall, there are theoretical giants such as Aldo Rossi (1990) and Robert Venturi (1991), productive stars like Zaha Hadid (2004) and Jean Nouvel (2008), people as Frank Gehry (1989) who hasn’t even built his most important works in the year when he won, and people like Frei Otto (2015) who was proclaimed given the award several days earlier because of his passing away. As time passes to recent years, more and more architects from outside of the mainstream have been appreciated and given the prize, among the examples there are Wang Shu (2012), Alejandro Aravena (2016) and RCR (2017).
this abundant variety, it has been a question in the public: who can indeed win the prize? Especially these years when the prize misses the highly-promising starchitects repeatedly, the discussion has been rather popular over the criteria for it.

If taking the methods and perspectives from Culturomics as a reference, is there a certain ‘pattern of fame’, or to be precise, the time series change of name frequency in all the publication in the human culture, that the architects share around the period of their winning the prize? If there is, can this pattern work as some analysis and prediction for those architects who haven’t yet been presented the prize? Is it possible to explain the result of the winner each year using the pattern in some way? These are the main issues that this research wants to discuss with.

2. Previous Research

2.1. CHARACTERISTICS STUDIES OF PRITZKER WINNERS

Because of its humanistic and social properties as a prize for individual architects, the common factor of Pritzker Prize is hard to tell. Strictly speaking, every winner from 1979 to 2017 is heterogeneous to each other, with several vital characteristics that the others don’t have. Therefore, many researchers and critics can only comment on a specific winner each year discussing his or her properties that drive him or her to be awarded. As British architect Richard Rogers got his prize in 2007, Some critic (Liu C., 2007) believed the prize had returned to the mainstream and started to present honor to the commonly-believed famous architects. However, only 5 years later, the prize has started leaning to the ‘uncommon’ architects. Commenting on winning for RCR, Song W., etc. (2017) used the title of ‘Unexpectedness and Inevitability’ to express their analysis, yet still basing on the specific characteristic only, failing to give explanation to why the traditional principles have been discarded.

Criticizing and analyzing on individuals can be much specified, but will cause the problem of over-fitting in statistics. Some critics tried to use the term of transnational capitalism class (Leslie, 2005), the introducing of digitalized architecture form (Xu and Liu, 2011) or the personality and altruism of architects themselves (Liu, 2014) to give some conclusions, while still seem pale to be applied to all the winners.

To give further explanation on the variation of Pritzker Prize, Zhou (2017) thinks this shows the division of its original principle. Before 2001, the winners rarely aroused controversial discussions. However, with Glenn Murcutt as separatrix, Pritzker Prize has been divided into the Fundamentalism for pursuing the intrinsic nature of architecture, and the Self-Organizing System for pursuing the possibility of public consumption through the mass media. Whenever some member of the former pattern who is not selected by the media consumption wins the prize, the public would be lead into an inevitable discussion, or even argument. This idea of dividing the winners into different patterns and groups is appreciable.
2.2. CULTUROMICS STUDIES BASED ON GOOGLE NGRAM

Researchers from Harvard, Google, MIT, etc. in 2011 constructed a corpus of digitized texts containing about 4% of all books ever printed between 1500 and 2008, basing on the method of N-gram as the segmentation algorithm, with eight languages and 8,116,746 books included, namely the Google Ngram. The significance of the corpus is that it succeeded to code and organize all the important cultural products of human being, thus believed to be the genomics of culture. The initial researchers proposed the methods based on Google Ngram to analyze and visualize the development and change in use of words in human history, to demonstrate and explain the significance of the word in different historic periods. Source of the word choice is abundant, leading to various research approaches, including language rules represented by the change of irregular verbs (Lieberman E., Michel J.B., etc., 2007), cultural individualism property represented by changes in pronoun use (Yu F., Peng, T., Peng, K., etc. 2016; Takeshi H., 2015), society development represented by change in daily word use (Greenfield, P.M., 2013), or even political events or Zeitgeist represented by the fluctuating of the names of famous people (Michel J. B., etc., 2011).

Drawing on lessons from the method, we could analyze and explain the group of architects. To be mentioned, the traditional Culturomics works as the demonstration of frequency of words based on the huge corpus, the observation for the appearance of the peaks and the entire trend, and responsive explanations in relationship with events in the reality, forming a result-oriented method. However, to solve our problem, merely using descriptive analysis as explanation is not enough, thus we introduced some new tools in statistics based on classical methods and neural networks.

2.3. IMPUTATION METHODS FOR MISSING VALUE

The dataset from Google N-gram reaching only before 2008, the analysis on the fame of recently awarded architects could be problematic. As we analyze the fame pattern from comparing the frequency of name in the scale of year towards winning, rather than the actual year, as will be discussed later, there will be several values missing due to transformation of independent variables.

As reviewed by several researchers (Little R. J., 2011; Yuan Z., 2008; Mao Q., 2005), there are several methods for imputing the missing values according to the pattern of the dataset. With Rubin first proposing the concept of Multiple Imputation in 1970s (Rubin D. B., 1976, 1996), this Bayesian-based method has been broadly used for much complicated pattern of missingness (Horton N. J., Lipsitz S. R., 2001 Cao Y., Xie J., Zhang L., 2003). However, simple methods such as Ad Hoc, Mean Substitution, and Expectation Maximization are still powerful in most common cases (Yuan Z., 2008; Mao Q., 2005).

The number of samples in this research being relatively small, each sample has a significant importance in the analysis, thus roughly discarding samples or drawing means from existing values as suggested by the simple methods mentioned above might harm to the result of analysis. As the missing values follow a monotone pattern (as shown in Figure 1), we decide to take reference from the
Multiple Imputation methods and use Multiple Regression Models to impute the missing value gradually, based on the assumption that various architects have some correlations in their fame.

2.4. SELF-ORGANIZING MAP

As suggested by Kohonen T. (1982, 1998), Self-Organizing Map (SOM) is a neural network approach, in which multidimensional signals can be received and automatically mapped onto a set of 1- or 2-dimensional output responses showing with topological features. Being a self-applied unsupervised learning system, SOM method has been broadly used to visualize the multidimensional data and functions as a competitive alternation for traditional clustering method like Hierarchical, Two-Step or K-means Clustering (Liu F., 2012). Several systems can achieve the aim of SOM method (Stefanovic P., Kurasova O., 2011), such as NeNet, SOM-Toolbox, Databionic ESOM and Viscovery SOMine. In this research we use Viscovery SOMine as the tool for clustering and visualizing the fame pattern of the architects along the time.

3. Method

3.1. MATERIALS AND DATA RESOURCES

Using the Python-based Web Crawler Bazhuayu, the database of all the Pritzker Winners through 1979-2017 including variables of the winning year, nationality, birth year, representative works and its built year, word count in the on-line article is collected from Wikipedia. Because of the limited sample size of winners only, we also use the same method for the promising names of architects who are frequently mentioned through the mass media when the Pritzker Prize is announced each year. The dataset then consists of 62 individual or group of architects, including 40 awarded (with 2 winners in 1988), and 22 promising. Each name of individual or group in the database is searched through Google Ngram in the time range of 1900-2008 in the English corpus, the frequency of each name in each year appearing in all the digitized publications in English language of that period is collected by checking the source code of the engine, which will be later analyzed as their fame, as suggested by Michel (2011).

During the process of data collecting, two issues make the fame pattern for some of the architects unachievable. The first is called Homonymity Conflicts as suggested by Michel J. B. etc. in 2011, in which the specific name cannot be separated from people in other fields with the same name, causing abnormal peak in the frequency. Using the method suggested by Michel, five of the architects failed the combat in homonymity conflicts, i.e. James Stirling (1981), Gottfried Böhm (1986), Norman Foster (1999), Richard Rogers (2007) and Wang Shu (2012). The second is that some of the architects failed to reach the threshold of being FAMOUS, which requires the peak of fame being higher than 10^-9 in the entire corpus, also as suggested by Michel, i.e. Ieoh Ming Pei (1983), Alejandro Aravena (2016), RCR Architects (2017) and several promising architects.

Thus, the final dataset for analysis consists of 48 architects, including 32 awarded and 16 promising, with their name frequency in human publication history
3.2. TRANSFORMATION OF THE DATASET

SPSS 22.0 is used during the analysis process of the dataset. The frequencies are taken the logarithm in order to be comparable, and all the values lower than $10^{-9}$ are set -9. Factor Analysis and ANOVA are tested among the awarded and promising architects through the years.

The original independent variable of the actual year is translated to the form of the year towards the winning of Pritzker Prize, from 25 years before winning to 10 years after winning for each architect, with the winning year of the promising architects all set on 2017, making a 48×36 matrix with empty values due to unavailable data in years after 2008, which occurs for the architects awarded after 1998 and all the promising ones. The transformation is shown left in Figure 1.

![Figure 1. The work flow of dataset transformation and imputation for missing values. The green dots represent the complete values before and after transformation. The grey dots represent the missing values after transformation. The orange-sided dots represent the missing values being predicted by regression models. The yellow dots represent the value accepted and imputed for the missing values. The red dots mark the position for data of name frequency right the year of Pritzker winning in different procedure.](image)

3.3. REGRESSION IMPUTATION FOR MISSING VALUES

The Matrix with architects as samples and time as variables is transposed, forming a new matrix with time as samples and architects as variables as shown middle in Figure 1.

As discussed in Section 2.3, the pattern of missing values here being monotone, and with the assumption that the architects have several similarities and correlation in their patterns of fame around the year of winning based on the previous discovery, multiple regression models are used as an imputation method gradually for several times. Each time in the multiple imputation process, the annual fame of a new architect with missing value forms the objective variable in the multiple regression model, and all the architects without missing value are included as potential explanatory variables. Stepwise method is used in the models as a tool for automatic selection of explanatory variables that finally enter the regression models. Imputation process goes along the hierarchy for the quantity of missing values. Each time when prediction of one specific architect is made, the existing values and the prediction is compared. If the correlation is significantly high, the predictions for the missing values are accepted, thus forming a new complete variable which is then added to the regression model for the next architect. This gradual imputation method completes all the missing values, making estimations...
for the trend on one hand, and giving support for the later clustering analysis of architects on their fame patterns on the other hand. Afterwards the matrix is transposed again. The whole working flow is shown in Figure 1.

3.4. CLUSTERING AND PATTERN ANALYSIS

The awarded architects are clustered with SOM method in the program Viscovery SOMine. Pattern of fame around the time of Pritzker winning is then analyzed using the result of clustering. Then the promising architects are added in the clustering SOM model, the distribution of which shows the discrimination within each pattern, thus giving clue for the most promising ones according to their similarity to the more typical prize winners.

4. Results

4.1. DISTRIBUTION AND ANOVA

Factor Analysis is first taken with all the frequency values from 1900-2008 for all the architects to reduce the dimension of variables. Two factors with eigenvalue greater than 10 is kept, explaining 79.394% of the total variation, among which Factor 1 mainly represents the variation before 1950, and Factor 2 the variation after 1950.

ANOVA is tested using the both factors between the awarded and promising architects. In both factors, the F values between both groups of architects are not significant with \( p > 0.05 \), meaning that in a general period, there is no significant difference in the fame between the awarded architects and promising ones, giving the necessity and support to take later analysis and comparison.

4.2. MULTIPLE REGRESSION FOR MISSING VALUE IMPUTATION

Thirty rounds of Multiple Regression are made following the process discussed in Section 3.3. All the regression models during the process are successful, with the prediction models all explain more than 90% of the \( R^2 \) of the variations, and the scatter points plots of prediction values and observed values all distributing along the line of \( y = x \), indicating the reliability and validity of the method.

Taking Rem Koolhaas who won the Pritzker at 1999 with one missing value as an example. A significant regression equation is found, \( F(4,33)=506.78, p<0.001 \). The variables successfully added into the regression model are Robert Venturi, Alvaro Siza, Gordon Bunshaft and Fumihiko Maki, all coefficients of regression model are significant, with \( p<0.05 \), explaining 98.4% of \( R^2 \). The statistic result of the regression is shown in Table 1.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient</th>
<th>SE</th>
<th>Standard Coefficient</th>
<th>t Value</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>8.317</td>
<td>1.803</td>
<td></td>
<td>4.612</td>
<td>.000</td>
</tr>
<tr>
<td>Robert Venturi</td>
<td>-1.216</td>
<td>.088</td>
<td>.761</td>
<td>16.203</td>
<td>.000</td>
</tr>
<tr>
<td>Alvaro Siza</td>
<td>.402</td>
<td>.054</td>
<td>.396</td>
<td>7.482</td>
<td>.000</td>
</tr>
<tr>
<td>Gordon Bunshaft</td>
<td>.697</td>
<td>.240</td>
<td>.777</td>
<td>2.858</td>
<td>.008</td>
</tr>
<tr>
<td>Fumihiko Maki</td>
<td>-.223</td>
<td>.111</td>
<td>-.140</td>
<td>-1.909</td>
<td>.058</td>
</tr>
</tbody>
</table>

Table 1. The regression result for Rem Koolhaas.
Worth mentioning is that during each round of regression, the variables selected by the Stepwise method is differentiated, indicating that each architect is independent and heterogeneous; while there is no single prediction without any variable selected, indicating that there exist some correlations among the fames of architects, which can be either positive or negative. Also, the result suggests that there could be some different patterns of fame among different architects, giving the source to and support for later clustering analysis.

4.3. PATTERN CLUSTERING AND PREDICTION

Using the SOM method discussed in Section 2.4, awarded architects could be made into different clusters. The process of the automatic topology mapping of different samples in each layer of variables in the neural network is shown in Figure 2, showing in distribution of all architects in each time slice. Six clusters are generated by the SOM system, as shown in figure 3 left.

![Figure 2](image)

Figure 2. Different variable layers of Self-Organizing Map showing topology of samples in each variable. The year of Prize winning is highlighted with the black frame. In each layer, the colder part, namely the blue and green represents smaller values of the mapped sample, while the warmer part represents the larger values.

![Figure 3](image)

Figure 3. Left: Clusters of Awarded Architects through SOM; Right: Clusters of All Architects through SOM. With awarded architects faded in grey, promising architects highlighted in black, and architects changed in cluster marked in different colors.

By drawing the line chart of fame values of different architects within each cluster, as shown in figure 4, certain patterns could be recognized through observation.

Three patterns each with two groups occur in the awarded architects, forming a $3 \times 2$ network. **Pattern A** contains 9 awarded architects including *Robert Venturi, Aldo Rossi* and *Rem Koolhaas*, which could be named as **Fame-Remaining Pattern**, since the architects in this pattern gained their fame early in their careers, their name frequency being higher than $10^{-8}$ in most of the time from 25 years before getting awarded till 10 years after. They could be regarded as the real starchsitects through their time, no matter in a theoretical way or in a practical way. **Pattern B** contains 14 awarded architects including *Frank Gehry, Rafael Moneo*, and *Toyo*
Ito, which could be named as *Fame-Fluctuating Pattern*, since along the career period, the fame of the architects in this pattern varied around the threshold of $10^{-8}$, with almost half above and half below the marked line in Figure 4. **Pattern C** contains the other 9 awarded architects including Philip Johnson, Sverre Fehn and Eduardo Moura, which could be regarded as *Fame-Descending Pattern*, since in most of the career life analyzed in the study, the fame of the architects failed to reach the threshold of $10^{-8}$, even during their peak of fame around the time of winning. The architects here are most commonly believed to be untypical winners, either from out of mainstream culture or seldom recognized by the public media beforehand as discussed in Section 2.1. In each pattern, there could still be two groups that are greatly heterogeneous, namely the *Ascending* and *Recovering*, the former showing the fame increasing all along the timeline, and the latter showing at least one trough period before the winning year.

The same method and process can be applied with the full database of both awarded and promising architects, SOM clusters have some mild changes as shown in Figure 3 right. The promising architects are added into all three patterns, with 7 in pattern A, 4 in pattern B and 5 in pattern C, respectively, the line chart of these new groups in all patterns are also visualized in Figure 4. As discussed above, the new group in pattern A should be regarded as the most promising ones as they are similar to the starchitects. The names in this pattern include Santiago Calatrava, Daniel Libeskind, Peter Eisenman and Steven Hall, not surprisingly being just those who are mostly talked about every year during the Pritzker time.

**5. Discussion**

The result of fame analysis here suggests that there exist some implicit patterns within the architects, whether he or she has been given Pritzker Prize or not. We coincidently find out the most promising not-yet-awarded architects have a strong relationship with the most popular nominees mentioned by public. However, as we
point out in earlier sessions, the Pritzker Prize is a complicated system containing various information, even reaching out of the boundary of the architects, the jury, the media and the public. Any trail to make a general analysis or a solid prediction with only one factor is not at all convincing, which is also the case in our research. The fame pattern clusters indicated by our result are only showing how similar or typical a promising architect is to those who have already been awarded in the aspect of their name frequency in written publications, which could be regarded as an indicator of their reputation. For further analysis, some more aspects of various architects should be considered, including their similarities and differences of design philosophy, their orientation in theoretical publications, and their development in architecture practice. Those aspects are also available by using the method of Culturomics with the corpus of Google Ngram, which could add more reliability and validity while studying and comparing the biography of one specific architect or a certain group of architects who are homogeneous or heterogeneous with each other.

Other than analysis on architects, Culturomics is also a practical tool of text analysis within the context of big data. The method can work as further assistance for architecture design and urban planning in several promising fields, including but not limited to site analysis, historical context synthesis, and cultural development classification.

Back to our own proposal, right now it depends only on the database in Google Ngram in a corpus of printed English books before 2008. With a further study comparing the similar case in different languages, reputation of architects in a cross-culture context could be detected. And with consideration of recent information explosion caused by mass media of Internet, the attention of public on the architects and their architecture works could also show more instant adaptations or vigorous mutations nowadays.

6. Conclusion

Using the fame data of different architects in different years from Google Ngram, we can find three basic patterns for Pritzker winners, typical or not. The most promising not-yet-awarded architects discriminated by this model are coincidently highly related to the ones being popular in public media, making the result more persuasive.

As it has since long been confirmed that we are already in a new age of data and information, case studies and critic in architecture design can also be revolutionarily updated and supplemented by learning from some other disciplines, expanding the concept of Computer Aided Architecture Design into broader aspects with digitalization, visualization and analysis of enormous open data. Culturomics could be a starting point.

Acknowledgement

We would like to show our gratitude to Professor Peng Kaiping from Department of Psychology, Tsinghua University, who gave us the insight of developing this research, and to all the anonymous reviewers who helped improve the quality of
References


