Qualitative analysis of housing demand using Google trends data

Kun-Huang Huarng\textsuperscript{a}, Tiffany Hui-Kuang Yu\textsuperscript{b} and Maria Rodriguez-Garcia\textsuperscript{c}

\textsuperscript{a}Department of Product Innovation and Entrepreneurship, National Taipei University of Business, Taiwan (R.O.C.); \textsuperscript{b}Department of Public Finance, Feng Chia University, Taichung, Taiwan (R.O.C.); \textsuperscript{c}Department of IUDESCOOP, University of Valencia, Valencia, Spain

\textbf{ABSTRACT}
Big data analytics often refer to the breakdown of huge amounts of data into a more readable and useful format. This study utilises Google Trends big data as a proxy for an analysis of housing demand. We employ a qualitative method (fuzzy set/Qualitative Comparative Analysis, fsQCA), instead of a quantitative method, for our estimate and forecast. The empirical results show that fsQCA successfully forecasts seasonal time series, even though the dataset is small in size. Our findings fill the gap in the qualitative and time series forecasting literature, and the forecasting procedure herein also offers a good standard for industry.

\textbf{ARTICLE HISTORY}
Received 20 August 2018
Accepted 29 October 2018

\textbf{KEYWORDS}
Time series models; qualitative forecasting; housing demand

\textbf{JEL CLASSIFICATIONS}
C32; C35; R21

1. Introduction

One potential cause of inefficiency in a housing market is a lack of liquidity (Tsai & Tsai, 2018). Building up a housing demand index could help understand why such liquidity may not arise at certain times. With Internet usage now rather common worldwide, vast amounts of consumer data are now available. A question thus arises: Can we analyse big data to forecast housing demand?

Big data technologies are naturally very useful when it comes to storing and processing huge sums of data (Diaconita, 2015). Wu and Brynjolfsson (2015) showed that queries submitted to Google’s search engine correlated strongly with both the volume of housing sales as well as a house price index released by the Federal Housing Finance Agency in the U.S. Hence, this study uses Google Trends data as the forecasting target, because the data are collected from the Internet and are readily available to the public.

Housing demand exhibits seasonality, as shown, for example, in Bangladesh (Ahmad, 2015) and the U.S. (Wu & Brynjolfsson, 2015). Other studies have applied a seasonal adjustment to model the housing demand problem, such as that in New Zealand (Grimes & Aitken, 2010), in Central, Eastern, and Southeastern Europe

\textbf{CONTACT} Kun-Huang Huarng \textsuperscript{a} khhuarng@ntub.edu.tw

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Fuzzy set/Qualitative Comparative Analysis (fsQCA) is a variant of QCA (Roig-Tierno, Gonzalez-Cruz, & Llopis-Martinez, 2017). It reveals the sufficient conditions that lead to a specific outcome (equivalent to a dependent variable in multiple regression analysis, MRA) (Woodside, 2013; Woodside, Nagy, & Megehee, 2018). The sufficient conditions can be combinations of antecedents or independent variables in MRA. Hence, to tackle seasonal time series like those existing for housing demand (Badun, & Franić, 2015), this study thus (1) uses fsQCA to autoregressively model a time series of order $k$, $AR(k)$, as the antecedent and the next time period’s data as the outcome and (2) employs the model to forecast the AR time series.

The rest of this study is organised as follows. Section 2 reviews the relevant literature. Section 3 introduces how to use Google Trends data as a big data proxy for housing demand, fsQCA, and the model set-up. Section 4 discusses the empirical analyses. Finally, section 5 concludes this study.

### 2. Literature review

FsQCA has been applied to solve various types of problems. For example, Trueb (2013) utilised fsQCA to integrate qualitative and quantitative data to create useful index for economic and policy development. Vis, Woldendorp, and Keman (2013) employed fsQCA to examine the variation in economic performance. Denk and Lehtinen (2014) used both Qualitative Comparative Analysis (QCA) and fsQCA to conduct contextual analysis of mobilisation of minority. Huarng (2015) employed fsQCA to analyse relationships within the development of information and communication technologies in more than 100 countries.

Huarng and Yu (2015) used fsQCA to explore the sufficient conditions for the outcome of healthcare expenditure from various antecedent combinations such as longevity, number of doctors, aging population, etc. FsQCA provided antecedent combinations for each year to show that causal complexities lead to highly consistent outcomes. The analysis in that study also showed strong predictive validities. Rey-Martí, Ribeiro-Soriano, and Palacios-Marqués (2016) also used both QCA and fsQCA to analyse culinary tourism success and entrepreneurial attributes under human capital and contingency factors.

The literature also shows how fsQCA can be employed to solve various time series problems. Huarng (2016), for example, used fsQCA to explore the relationships between energy consumption related antecedents and the outcome of gross domestic product. These regime switches identified by the drastic changes of the values in the antecedent combinations matched the real historical events, such as oil crises. Huarng and Yu (2017) employed fsQCA to analyse the autoregressive relationships of upward and downward regime switches for in-sample Taiwan Capitalisation Weighted Stock Index. The relationships are used to forecast regime switches in out-of-sample data. Taking data on the Taiwan Capitalisation Weighted Stock Index for analysis, the empirical results again show that fsQCA provides strong predictive validities.
3. Data and research method

3.1. Data

This study uses Google Trends to search for keyword ‘591.com’, one of the most popular housing websites in Taiwan and obtain the target data for this study (Google Trends data hereafter). The monthly data spanned the period January 2010 to March 2018. The data for the period January 2010 to November 2016 were used as the in-sample, and the data for the period December 2016 to March 2018 were used as the out-of-sample.

3.2. Seasonality

We define a seasonal time series, $k$, as follows:

$$\text{obs}(t-k), \text{obs}(t-k+1), \ldots, \text{obs}(t-1) \to \text{obs}(t),$$

where $t$ is the time period, and $\text{obs}(t-k), \text{obs}(t-k+1), \ldots$ are the Google Trends data. Because this study is modelling housing demand, we set $k$ to 11. For example, when we want to establish the relationships among the data, we present:

2010-Jan, 2010-Feb, 2010-Mar, … 2010-Nov → 2010-Dec
2010-Feb, 2010-Mar, 2010-Apr, … 2010-Dec → 2011-Jan
2010-Mar, 2010-Apr, 2010-May, … 2011-Jan → 2011-Feb
...
2015-Feb, 2015-Mar, 2015-Apr, … 2015-Dec → 2016-Jan
2015-Mar, 2015-Apr, 2015-May, … 2016-Jan → 2016-Feb
...
2015-Dec, 2016-Jan, 2016-Feb, … 2016-Oct → 2016-Nov

In addition, for our forecast, we have:

2016-Jan, 2016-Feb, 2016-Mar, … 2016-Nov → 2016-Dec
2016-Feb, 2016-Mar, 2016-Apr, … 2016-Dec → 2017-Jan
2016-Mar, 2016-Apr, 2016-May, … 2017-Jan → 2017-Feb
...
2017-Feb, 2017-Mar, 2017-Apr, … 2017-Dec → 2018-Jan
2017-Mar, 2017-Apr, 2017-May, … 2018-Jan → 2018-Feb
2017-Apr, 2017-May, 2017-Jun, … 2018-Feb → 2018-Mar
3.3. FsQCA

Based on fuzzy sets and set theory, fsQCA is a popular qualitative analysis method and tool. There are four major differences between fsQCA and MRA (Ragin, 2008). First, MRA creates coefficients for the dependent variables that are symmetrically designed. Secondly, fsQCA focuses on set relations that are asymmetrical, while MRA conducts analysis on the data directly. However, the data for fuzzy sets must be calibrated before analysis. Thirdly, MRA treats each independent variable individually, whereas fsQCA yields antecedent combinations as the sufficient conditions for an outcome. Fourth and lastly, MRA focuses on net effects, but in fsQCA there may be multiple antecedent combinations for the same outcome.

For fsQCA analysis, we first need to calibrate the data into values between 0.0 and 1.0, where 0.0 means full non-membership, 1.0 means full membership, and 0.5 means neither non-membership nor full membership. Using Google Trends, we search 591.com for the data. Because the values of the data from Google Trends fall into the range between 26 and 100, we set 80, 60, and 40 to be 1.0, 0.5, and 0.0, respectively, for calibration.

This study conversely proposes a new method called de-calibration. In other words, after forecasting, we need to covert the fuzzy values back into real data so that we can compare the forecasting performance. For example, the value of 1.0 is de-calibrated to 80, 0.5 to 60, and 0.0 to 40.

FsQCA analyses the relationships between the calibrated data and renders antecedent combinations, such as:

\[ \sim X \ast Y \rightarrow Z, \]

where X and Y are the antecedents (equivalent to independent variables), and Z is the outcome (equivalent to the dependent variable). The symbol \( \ast \) represents AND; the symbol \( \sim \) represents NOT; and the symbol \( \rightarrow \) means ‘is sufficient for’. The antecedents connected by \( \ast \) and \( \sim \) are called antecedent combinations (AC), such as \( \sim X \ast Y \). The above equation means that the antecedent combination \( \sim X \ast Y \) is a sufficient condition for Z. In other words, \( \sim X \ast Y \) can lead to Z.

We calculate consistency as follows:

\[
\text{consistency} = \frac{\sum \min(\text{comp, } y)}{\sum \text{comp}},
\]

where \( \text{comp} \) represents the calibrated value for each AC, and \( y \) is the calibrated value of the outcome.

4. Empirical analyses

For both the in-sample and out-of-sample data, we first conduct calibration. Based on the calibrated data, we then proceed to the model and lastly to the forecast.
The in-sample data allow us to establish relationships. For example, the in-sample data are:

April:

\[ AC_1 = \text{mar} \cap \text{feb} \cap \text{dec} \cap \text{sep} \cap \text{aug} \cap \text{jun} \cap \text{may} \cap \text{apr} \cap \text{mar} \cap \text{feb} \cap \text{nov} \cap \text{oct} \cap \text{sep} \cap \text{aug} \cap \text{jun} \cap \text{may} \cap \text{apr} \]

All the ACs can be generated for the other months, as seen in Table 1. Note that there can be multiple ACs leading to one month. Table 2 lists the correspondences for all the ACs.
Table 2. In-sample consistency.

| Consistency | Comp1 | Comp2 | Comp3 | Comp4 |
|-------------|-------|-------|-------|-------|
| 2016-Dec    | 0.9741| 1     |       |       |
| 2017-Jan    | 0.9381|       |       |       |
| 2017-Feb    | 0.9286|       |       |       |
| 2017-Mar    | 1     | 1     |       |       |
| 2017-Apr    | 1     |       |       |       |
| 2017-May    | 1     |       |       |       |
| 2017-Jun    | 1     |       |       |       |
| 2017-Jul    | 1     |       |       |       |
| 2017-Aug    | 1     | 0.8182|       | 1     |
| 2017-Sep    | 0.9041|       |       | 1     |
| 2017-Oct    | 0.9257|       |       |       |
| 2017-Nov    | 1     | 0.8693|       | 1     |
| 2017-Dec    | 0.9741|       |       |       |
| 2018-Jan    | 0.9381|       |       |       |
| 2018-Feb    | 0.9286|       |       |       |
| 2018-Mar    | 1     |       |       |       |

Table 3. The values of comp’s.

|        | Comp'1 | Comp'2 | Comp'3 | Comp'4 |
|--------|--------|--------|--------|--------|
| 2016-Dec | 0.32   | 0.32   |        |        |
| 2017-Jan | 0.32   |        |        |        |
| 2017-Feb | 0.16   |        |        |        |
| 2017-Mar | 0.11   | 0.16   |        |        |
| 2017-Apr | 0.68   | 0.26   |        |        |
| 2017-May | 0.11   | 0.16   |        |        |
| 2017-Jun | 0.11   | 0.03   | 0.16   |        |
| 2017-Jul | 0.11   | 0.03   | 0.16   |        |
| 2017-Aug | 0.11   | 0.03   | 0.11   | 0.16   |
| 2017-Sep | 0.16   | 0.11   |        |        |
| 2017-Oct | 0.32   | 0.11   |        |        |
| 2017-Nov | 0.11   | 0.11   | 0.35   |        |
| 2017-Dec | 0.57   | 0.11   |        |        |
| 2018-Jan | 0.43   |        |        |        |
| 2018-Feb | 0.43   |        |        |        |
| 2018-Mar | 0.43   | 0.29   |        |        |

Table 4. Product of comp’ and in-sample consistency.

| Product  | Comp'1  | Comp'2  | Comp'3  | Comp'4  |
|----------|---------|---------|---------|---------|
| 2016-Dec | 0.3117  | 0.32    |         |         |
| 2017-Jan | 0.3002  |         |         |         |
| 2017-Feb | 0.1486  |         |         |         |
| 2017-Mar | 0.11    | 0.16    |         |         |
| 2017-Apr | 0.68    | 0.26    |         |         |
| 2017-May | 0.11    | 0.16    |         |         |
| 2017-Jun | 0.11    | 0.03    | 0.16    |         |
| 2017-Jul | 0.11    | 0.03    | 0.16    |         |
| 2017-Aug | 0.11    | 0.0245  | 0.11    | 0.16    |
| 2017-Sep | 0.1447  | 0.11    |         |         |
| 2017-Oct | 0.2962  | 0.11    |         |         |
| 2017-Nov | 0.11    | 0.0956  | 0.35    |         |
| 2017-Dec | 0.5553  | 0.11    |         |         |
| 2018-Jan | 0.4034  |         |         |         |
| 2018-Feb | 0.3993  |         |         |         |
| 2018-Mar | 0.43    | 0.29    |         |         |
4.2. Out-of-sample forecasting

Because we execute the forecast one by one, there is no summation for the function of consistency. Therefore, the equation becomes:

$$\text{consistency} = \frac{\min(\text{comp}, y)}{\text{comp}}$$

For out-of-sample forecasting, we have $\text{comp}'$ and $y'$, representing the calibrated values of the out-of-sample:

$$\text{consistency} = \frac{\min(\text{comp}', y')}{\text{comp}'}$$

In other words, we have:

$$\min(\text{comp}', y') = \text{comp}' \times \text{consistency}, \text{ whereby}$$

| 1-below0.5 | Comp'1 | Comp'2 | Comp'3 | Comp'4 |
|------------|--------|--------|--------|--------|
| 2016-Dec   | 0.6883 | 0.68   |        |        |
| 2017-Jan   | 0.6998 |        |        |        |
| 2017-Feb   | 0.8514 |        |        |        |
| 2017-Mar   | 0.89   | 0.84   |        |        |
| 2017-Apr   | 0.68   | 0.74   |        |        |
| 2017-May   | 0.89   | 0.84   |        |        |
| 2017-Jun   | 0.89   | 0.97   | 0.84   |        |
| 2017-Jul   | 0.89   | 0.97   | 0.84   |        |
| 2017-Aug   | 0.89   | 0.9755 | 0.89   | 0.84   |
| 2017-Sep   | 0.8553 | 0.89   |        |        |
| 2017-Oct   | 0.7038 | 0.89   |        |        |
| 2017-Nov   | 0.89   | 0.9044 | 0.65   |        |
| 2017-Dec   | 0.5553 | 0.89   |        |        |
| 2018-Jan   | 0.5966 |        |        |        |
| 2018-Feb   | 0.6007 |        |        |        |
| 2018-Mar   | 0.57   | 0.71   |        |        |

Table 6. De-calibrated values.

| De-calibrated | Comp'1 | Comp'2 | Comp'3 | Comp'4 |
|--------------|--------|--------|--------|--------|
| 2016-Dec     | 65     | 65     |        |        |
| 2017-Jan     | 66     |        |        |        |
| 2017-Feb     | 72     |        |        |        |
| 2017-Mar     | 74     | 71     |        |        |
| 2017-Apr     | 65     | 67     |        |        |
| 2017-May     | 74     | 71     |        |        |
| 2017-Jun     | 74     | 83     | 71     |        |
| 2017-Jul     | 74     | 83     | 71     |        |
| 2017-Aug     | 74     | 84     | 74     | 71     |
| 2017-Sep     | 72     | 74     |        |        |
| 2017-Oct     | 66     | 74     |        |        |
| 2017-Nov     | 74     | 75     | 64     |        |
| 2017-Dec     | 61     | 74     |        |        |
| 2018-Jan     | 63     |        |        |        |
| 2018-Feb     | 63     |        |        |        |
| 2018-Mar     | 62     | 67     |        |        |
If $comp_0' \geq y'$, then $\min(comp_0', y') = y'$;

If $comp_0' < y'$, then there is no way to know $y'$ from $\min(comp_0', y')$.

Hence, we suppose $\min(comp_0', y') = y'$. In other words, $y' = comp_0' \times consistency$.

The value of consistency is known from the in-sample estimation. The values of $comp_0'$ are calculated by the out-of-sample data as seen in Table 3. Thus, we are able to calculate the value of $y'$ as Table 4 shows.

The values of $y'$ may be greater than or lower than 0.5. If the future trend is clearly rising or dropping, then we can determine whether the values are greater or lower than 0.5. Suppose the trend is upward (downward); the values therefore need to be greater (lower) than 0.5. If the values are opposite to this, then we need to convert them by $1-y'$.

All the values (except $comp_1'$ for 2017 April and $comp_1'$ for 2017 December) in Table 4 are lower than 0.5. However, we know that the trend is upward. Hence, we take $1-y'$ for all the values as presented in Table 5. As for the values greater than 0.5, they remain the same. As stated previously, we de-calibrate the values back to the real data, listing the de-calibrated results in Table 6.

4.3. Performance evaluation

There are different ways to validate the empirical results. First, as in many fsQCA studies, we use an XY plot to see if the forecasts fall in the upper-left triangle of the chart, which would be considered as good forecasts. Figure 1 presents the forecasts for both 2017 January and 2018 January. The dashed circle indicates the forecast of 2017 January, and the dashed triangle indicates the forecast of 2018 January. Both
are above the diagonal line and are therefore considered as good forecasts. Table 7 shows that all the forecasts are denoted as being good, except that for 2018 February.

Secondly, one can always use conventional root mean squared errors (RMSEs) to test a forecasting performance. There is a de-calibrated value corresponding to each relationship. When there are multiple relationships for the same outcome, there are multiple de-calibrated values. Hence, the choice of which de-calibrated value as the representative value to the outcome becomes an issue. There are three possible ways to decide. Method 1 picks the largest product values for different comps. Method 2 chooses the smallest product values for different comps. Method 3 chooses the largest product values, but with values greater than 0.5 first. Table 8 compares the RMSEs.

5. Discussion and conclusion

To tackle a seasonal time series for housing demand, this study uses fsQCA (1) to model the time series of an AR(k) and (2) employs the model to forecast the AR
time series. First, this study demonstrates how Google Trends data can represent a proxy source of big data, offering a proper standard for industry. Secondly, fSQCA generates multiple relationships that lead to the same outcome. In other words, multiple causes can lead to the same result. Thirdly, when facing a small dataset, fSQCA is still able to successfully model and forecast a higher-order time series that outperforms conventional statistical methods.

There are a few limitations of this study. As discussed in the text, the equation to calculate consistency involves a minimum function, which absorbs some amount of information. Hence, during out-of-sample forecasting, there is no way to discover the absorbed information. In addition, multiple relationships also cause bias or problems in out-of-sample forecasting. A consistent approach is needed to determine which relationship is more suitable in a given situation.

Following these limitations, we suggest that future studies focus on how to choose the proper relationship from multiple relationships. Such investigations can be done by applying domain-specific knowledge, such as through heuristics or with domain-independent knowledge. Another interesting topic to work on is how to retain the absorbed information in a minimum function.

Acknowledgement

The authors acknowledge and are grateful for the financial support of the Ministry of Science and Technology, Taiwan, ROC under grant MOST 105-2410-H-035 -029 -MY2.

References

Addison-Smyth, D., & McQuinn, K. (2010). Quantifying revenue windfalls from the Irish housing market. *The Economic and Social Review, 41*(2), 201–223.
Ahmad, S. (2015). Housing demand and housing policy in urban Bangladesh. *Urban Studies, 52*(4), 738–755.
Badun, M., & Franić, J. (2015). The importance of government incentives for housing savings in Croatia. *Economic Research-Ekonomska Istraživanja, 28*(1), 331–353.
Denk, T., & Lehtinen, S. (2014). Contextual analyses with QCA-methods. *Quality & Quantity, 48*(6), 3475–3487.
Diaconita, V. (2015). Processing unstructured documents and social media using Big Data techniques. *Economic Research-Ekonomska Istraživanja, 28*(1), 981–993.
Grimes, A., & Aitken, A. (2010). Housing supply, land costs and price adjustment. *Real Estate Economics, 38*(2), 325–353.
Huarng, K.-H. (2015). Configural theory for ICT development. *Journal of Business Research, 68*(4), 748–756.
Huarng, K.-H. (2016). Identifying regime switches using causal recipes. *Journal of Business Research, 69*(4), 1498–1502.
Huarng, K.-H., & Yu, T. H.-K. (2015). Healthcare expenditure with causal recipes. *Journal of Business Research, 68*(7), 1570–1573.
Huarng, K.-H., & Yu, T. H.-K. (2017). Using qualitative approach to forecasting regime switches. *Quality & Quantity, 51*(5), 2035–2048.
Ragin, C. C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond.* Chicago, IL: University of Chicago Press.
Rey-Martí, A., Ribeiro-Soriano, D., & Palacios-Marqués, D. (2016). Entrepreneurial attributes of human capital and contingency factors in the culinary tourism. International Entrepreneurship and Management Journal, 12(1), 67–85.

Roig-Tierno, N., Gonzalez-Cruz, T. F., & Llopis-Martinez, J. (2017). An overview of qualitative comparative analysis: A bibliometric analysis. Journal of Innovation & Knowledge, 2(1), 15–23.

Trüeb, B. (2013). Integrating qualitative and quantitative data: index creation using fuzzy-set QCA. Quality & Quantity, 47(6), 3537–3558.

Tsai, H. C., & Tsai, I. C. (2018). Market depth in the UK housing market. Economic Research-Ekonomska Istraživanja, 31(1), 406–427.

Vandenbussche, J., Vogel, U., & Detragiache, E. (2015). Macroprudential policies and housing prices: A new database and empirical evidence for Central, Eastern, and Southeastern Europe. Journal of Money, Credit and Banking, 47(S1), 343–377.

Vis, B., Woldendorp, J., & Keman, H. (2013). Examining variation in economic performance using fuzzy-sets. Quality & Quantity, 47(4), 1971–1989.

Woodside, A. G. (2013). Moving beyond multiple regression analysis to algorithms: Calling for adoption of a paradigm shift from symmetric to asymmetric thinking in data analysis and crafting theory. Journal of Business Research, 66(4), 463–472.

Woodside, A. G., Nagy, G., & Megehee, C. M. (2018). Applying complexity theory: A primer for identifying and modeling firm anomalies. Journal of Innovation & Knowledge, 3(1), 9–25.

Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker (Eds.), Economic analysis of the digital economy (pp. 89–118). University of Chicago Press.