Cluster analysis of agricultural household production of self-employed

A V Plotnikov
Perm National Research Polytechnic University, 29 Komsomolsky Av., Perm, 614000, Russia
E-mail: plotnikov-av@mail.ru

Abstract. The paper analyzes product data collected from the VK social network (N=26295). The study aims to analyze the prices of ordered goods from the VK social network communities of self-employed citizens and then find any exciting observations from the data. The research methods are correlation analysis and K-means cluster analysis. We are interested in finding correlated prices for self-employed goods to use these results in future studies to determine the balance of the market for household goods. As a result of the study, we identified correlations between the following products: 'rabbit carcass' and 'salted cucumbers'; 'Belper Knolle' and 'chicken roll'; 'breast fillet' and 'chicken eggs'; 'hot smoked' and 'homemade cutlets'; 'pork tenderloin' and 'pork ribs'.

1. Introduction
In comparison with hired employment, self-employment is a more independent and strong-willed choice of profession, which strengthens the experience of meaningfulness at work and the perception of labor autonomy [1]. Sustainable self-employment careers have higher gross labor income and higher job and life satisfaction than all other self-employment career models [2]. The choice in favor of self-employment does not depend on the person's education [3]. Self-employment allows you to engage in activities at retirement age [4].

There are several self-employed categories: freelancers, innovators, traditional small business owners (including farmers), dependent self-employed, and hybrid self-employed [5]. In our work, we will look at the features of self-employed home farms based on social network data. The result is a unique study from the perspective of the interaction of the self-employed in the VK social network. In the study, we analyze the "digital footprints" of the self-employed and their clients. Agricultural producers have the right to switch to a tax on professional income. You can become "self-employed" (article 4, 15 of the Federal law of 27.11.2018 N 422-FZ "on experimenting with establishing a special tax regime "tax on professional income").

A self-employed person who independently runs a subsidiary farm and sells their products can apply for self-employed status. Products must be manufactured or processed with your own hands without involving employees. The annual income should not exceed 2.4 million Russian rubles. Following article 5 of Federal law No. 422-FZ of 27.11.2018, citizens of the Eurasian Economic Union can register in Russia as self-employed. In addition to Russia, the EEU includes Belarus, Armenia, Kazakhstan, and Kyrgyzstan. Pushing from unemployment to self-employment shifts social risks away from the state [6].
Self-employed people can use Bank loans based on Bank statements and a transition certificate to the tax regime. According to a study by Junyi Xiang et al., access to credit can benefit a significant portion of the self-employed. Effectively targeting the minority of self-employed with higher growth potential is essential, especially in low-income settings [7].

2. Scientific problem
The social network VK has ceased to be just a social network for people to communicate. Today it is a full-fledged platform for doing business, which opens up selling goods to the self-employed. Thus, it is possible to analyze competitors, determine their price for the product, and determine the differences in self-employed and trading companies' prices.

We are interested in finding correlated prices for self-employed goods to use these results in future studies to determine the balance of the market for household farm products.

3. Purpose of the study
Analyze the collected products from the communities of self-employed citizens of the social network VK, and then find any interesting observations from the data.

4. Research questions
What products of farm households in the VK social network are linked by prices depending on the time of year, when, for example, the rise in the price of some goods causes other goods' growth?

5. Research methods
We used the following Python packages to work with tabular data and visualizations in graphical form: pandas, re, numpy, matplotlib, sklearn, seaborn.

Total unique farm products in the sample of 26295 (N=26295) (dataset [8]), the number of unique sellers = 935. We use the K-means method to cluster products based on the principle of maximum similarity of names. Then we build a correlation matrix for clusters over time and define groups of correlating clusters to explain their relationship in the future. Let's look at the distribution of prices for goods, except for emissions related to real estate and agricultural machinery.

6. Research result

Figure 1. Distribution of product prices.

The primary quantity and product names have a price of up to 1000 rubles (figure 1). Let's pay attention to the appearance of the goods on dates (figure 2).
Figure 2. Distribution of the appearance of products on the social network showcase.

In 2020 (figure 2), we see a surge in new products. Perhaps this fact is related to the introduction of a special tax regime, "tax on professional income" (or tax for the "self-employed"), and the popularization of this tax regime. We see minor spikes in January 2017, 2018, and 2019. Perhaps freelancers or self-employed people start a new life cycle from the New year and increase their efficiency from the beginning of the year. Next, let's pay attention to the appearance of the goods by the time of day (figure 3).

Figure 3. The appearance of the goods, depending on the time of day.

We can assume that the self-employed, despite their free schedule, adhere to the standard work schedule from morning to evening (figure 3). We did not analyze which days of the week or on weekends, the self-employed made the display of goods.

After general data analysis, we will apply k-means clustering on the vector representations of words of products based on the principle of similarity of names. Let's create an array of averaged vectors for each product name. Then we apply SVD decomposition to the resulting feature matrix (dimension: number of products per size of the averaged vector) to reduce the dimension. Next, we will perform clustering using the K-Means method.

We will remove products with a price of more than 10,000 from the dataset since they are mainly real estate, farm equipment, etc. In this work, we are primarily interested in food and related products.

First, we tokenize the "Name" field and define a class for calculating the "average" vector of words in the text field.

We use a pre-trained model from the site https://rusvectores.org/ru/models/"geowac_tokens_none_fasttextskipgram_300_5_2020". To train this model, we used a sample of Russian-language documents from the CommonCrawl dump, balanced by geography, compiled by Jonathan Dunn and Ben Adams; the corpus Size is 2.1 billion words.
To select the optimal number of clusters, we use quality criteria such as inertia and silhouette. Consider the range of the number of clusters from 2 to 9 (then up to 50).

Inertia (figure 4) is the average sum of the squares of the distance to the centroid.

**Figure 4.** Inertia: the Elbow method showing the optimal 3 clusters.

The average silhouette (figure 5) for all points from \( \mathbf{X} \) is the clustering quality criterion. It shows how the average distance to objects in your cluster differs from the average distance to objects in other clusters. Takes the values \([-1, 1]\):

- -1 is poor (scattered) clustering;
- 0 is clusters intersect and overlap each other;
- 1 is "dense" clearly marked clusters.

The larger the silhouette, the more the clusters are highlighted, and they are compact, tightly grouped point clouds.

Since the feature space turned out to be quite large, we use SVD decomposition to simplify and speed up the model without significant information loss.

Let's look at the proportion of variance explained by the new features. A good value is the proportion of the explained variance of \( \geq 80\% \)

\[ V = 0.85 \]

Let's plot the growth of the share of the explained variance (figure 6).

**Figure 6.** Increase in the proportion of explained variance.
Using SVD, we effectively reduced the dimension of the feature space from 300 to 100. Then, using the silhouette sampling method (figure 7), we determined the optimal number of clusters = 45.

For n_clusters = 2 The average silhouette_score is : 0.09439639225814947
For n_clusters = 3 The average silhouette_score is : 0.09144050346052632
For n_clusters = 4 The average silhouette_score is : 0.08852162838439609
For n_clusters = 5 The average silhouette_score is : 0.08234756316699983
For n_clusters = 6 The average silhouette_score is : 0.08410155322493175
For n_clusters = 14 The average silhouette_score is : 0.08302760066547653
For n_clusters = 23 The average silhouette_score is : 0.09853700766496705
For n_clusters = 45 The average silhouette_score is : 0.11298699317443985

**Figure 7.** Silhouette: the Elbow method showing the optimal 45 clusters.

Let's perform a more detailed analysis of the silhouette in graphical form to select the number of clusters. Let's run the function on the feature matrix and the considered range of the number of 45 clusters (figure 8). We will write the cluster numbers for different n_clusters in a separate data frame.

**Figure 8.** Silhouette analysis for K-Means clustering on sample data with n_clusters = 45.

0-cluster is mixed into many clusters, and the remaining clusters are distributed separately from each other. Even though there are no clear borders, we manually checked the content of these clusters, and it suits us. As an example, let's illustrate two random clusters of the 45 clusters (figure 9).
Next, consider clusters by price distribution (figure 10).

The median (50 percentiles) is shown as a yellow line, and the 25 and 75 percentiles are the borders of the box. Whiskers display the entire spread of points except for outliers, i.e., the minimum and maximum values that fall within the interval \((Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR)\), where \(IQR = Q_3 - Q_1\) is the interquartile range. Points on the graph indicate outliers — those that do not fit into the range of values specified by the graph's whiskers. Let's display the result graphically.

Recall that in Python programming language calculus starts with 0, the graph shows 45 clusters: '... hand-made' – 0; 'Homemade cheese' – 1; 'Rabbit carcass' – 2; 'Chicken, gherkin' – 3; 'Homemade dumplings' – 4; 'Cow's milk' – 5; 'Adyghe cheese' – 6; 'Butter' – 7; 'Kefir' – 8; 'Cream, honey' – 9; 'Pork, beef' – 10; 'Belper knolle' – 11; 'Pork tenderloin' – 12; '... Hot smoked' – 13; 'Carcass weight' – 14; 'Goat milk' – 15; 'Chicken egg' – 16; 'Sponge cake' – 17; 'Salted cucumbers' – 18; 'Curd' – 19; 'White cabbage' – 20; 'Bones, pork' – 21; '... Cold smoked' – 22; 'Breast fillet' – 23; 'Strawberry seedlings' – 24; 'Cloth mask' – 25; 'Adyghe cheese' – 26; 'Onion' – 27; 'Chicken roll' – 28; 'Cottage cheese' – 29; 'Cream' – 30;
'Compound feed' – 31; 'Walnut' – 32; 'Potato dumplings' – 33; 'Chicken eggs' – 34; 'Glass, beer' – 35; 'Sour cream' – 36; 'Homemade chicken' – 37; 'Shu Puer tea' – 38; 'blooming Sally' – 39; 'Homemade cutlets' – 40; 'Pork ribs' – 41; 'extra virgin' – 42; 'soup set' – 43; 'autumn apple tree' – 44.

Data can be useful for self-employed people for pricing and for selecting a niche with demand. Now let's build a correlation matrix and determine the clusters that correlate with each other. The value of the correlation coefficient reflects the strength of the connection (figure 11).

![Figure 11. Correlation matrix for the clusters.](image)

When constructing the matrix, the correlation close to zero is indicated in black, using the color scheme (figure 11). The Chaddock scale is used to estimate the strength of the correlation coefficients. We select only those correlating clusters with a correlation coefficient $r \geq 0.7$, which corresponds to
the threshold value of a strong relationship between two products. Then we will output a text list from the graph only correlated clusters of products:

- {'Rabbit carcass' – 2; 'Salted cucumbers' – 18}
- {'BelperKnolle' – 11; 'Chicken roll' – 28}
- {'Breast fillet' – 23; 'Chicken eggs' – 34}
- {'... hot smoked' – 13; 'Homemade cutlets' – 40}
- {'Pork tenderloin' – 12; 'Pork ribs' – 41}

Now let's see how the price of goods for clusters changes over time. We will take for illustration only two pairs of comparisons that we consider to be the most correlative (figure 12).

![Figure 12. Price changing when products appear, depending on the date.](image)

7. Conclusion

We found high correlations between the following products: 'rabbit carcass' and 'salted cucumbers'; 'Belper knolle' and 'chicken roll'; 'breast fillet' and 'chicken eggs'; '... hot smoked' and 'homemade cutlets'; 'pork tenderloin' and 'pork ribs'. It may reflect the demand by season for goods or resources from which this product will make. We also noticed that there are a lot of end products on display. Thus, self-employed farmers, by selling the final product to consumers without intermediaries, increase their income. Scientists and specialists can use correlated prices for self-employed goods to determine the balance of the market for household goods.

Acknowledgement

The study was supported by a grant from the Russian Science Foundation (project No. 20-78-00100).

References

[1] Stephan U, Tavares S M, Carvalho H, Ramalho J J, Santos, S C and van Veldhoven M 2020 Self-employment and eudaimonic well-being: Energized by meaning, enabled by societal legitimacy *J of Business Venturing* 35(6) 106047

[2] Koch M, Park S and Zahra S A 2019 Career patterns in self-employment and career success *J of Business Venturing* 105998 https://doi.org/10.1016/j.jbusvent.2019.105998

[3] Cowling M, Millán J M and Yue W 2019 Two decades of European self-employment: Is the answer to who becomes self-employed different over time and countries? *J of Business Venturing Insights* 12(C) https://doi.org/10.1016/j.jbvi.2019.e00138

[4] Nolan A and Barrett A 2019 The role of self-employment in Ireland’s older workforce *The J of
the Economics of Ageing 14 100201

[5] Skrzek-Lubasińska M and Szaban J M 2019 Nomenclature and harmonised criteria for the self-employment categorisation. An approach pursuant to a systematic review of the literature European Management J 37(3) 376-86

[6] Danson M, Galloway L and Sherif M 2020 From unemployment to self-employment: Can enterprise policy intensify the risks of poverty? Critical Perspectives on Accounting 102164 https://doi.org/10.1016/j.cpa.2020.102164

[7] Gindling T H and Newhouse D 2012 Self-employment in the developing world The World Bank

[8] Plotnikov A 2020 Household goods Mendeley Data V1 https://doi.org/10.17632/nb42nh6bjm.1