Summer Invitational Data Open 2020

From Industry Clustering to Regional Difference: Small Business Growth in 2010 Winter Olympics

Team #23
Jiachuan Bi, jb4360@columbia.edu
Jiajun Song, js814@duke.edu
Peimou Sun, ps3136@columbia.edu
Wenyi Xu, wx2226@columbia.edu

Jul 19, 2020

Contents

1 Problem Statement

In British Columbia, small businesses continue to play a vital role, since the vast majority of businesses in the province have fewer than 50 employees. The small business sector, which generated 34% of overall provincial GDP[1], is a key instrument of job creation and economic growth. In light of this, our report aims to achieve a better understanding of how the Vancouver 2010 Winter Olympics has influenced small businesses in different regions of British Columbia.

2 Executive Summary

To explore the regional difference of the 2010 Winter Olympics' impact on small business growth, we investigated the industry composition in the 8 development regions (DR) of the British Columbia area. We found that different industry compositions result in different regional growth rates. The main procedure and results of our research and analysis are:

1. Structural break test of different regions: We visualized the growth of small business amount in region DR01 - DR08 and tested if there exist structural breaks in time series (QLR test). We concluded that DR01, DR04, and DR06 are significantly sensitive to the 2010 Winter Olympics, with p-value 0.141, 0.095, and 0.094.

2. Spectral clustering of industries: We utilized the Pearson correlation matrix of industries to identify the high & low sensitive industries to the 2010 Winter Olympics. The structural break p-values of cluster 1 (sensitive) and cluster 2 (insensitive) are 0.078 and 0.390. We further verified that the regions with higher growth rates have higher proportions of cluster 1 industries, which validated our argument.

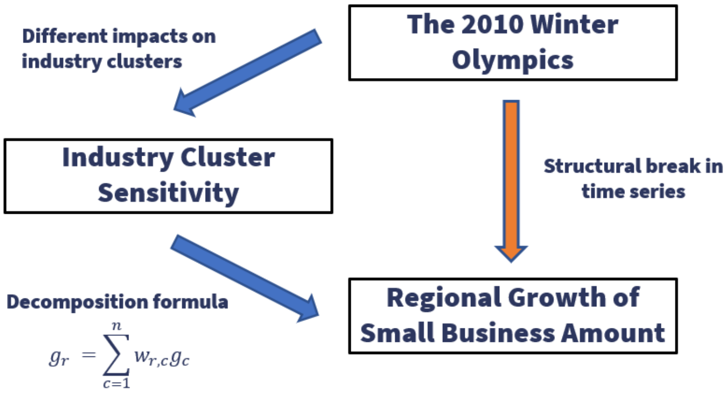

The 2010 Winter Olympics
Different impacts on industry clusters
Industry Cluster Sensitivity
Structural break in time series
Decomposition formula
$g_r = \sum_{c=1}^{n} w_{r,c} g_c$
Regional Growth of Small Business Amount

Figure 1: Summary of results

These two aspects are connected by the regional growth decomposition formula in Figure 1, which attributes the small business growth to different industry clusters and accounts for the regional difference.

[1] https://www2.gov.bc.ca/assets/gov/employment-business-and-economic-development/business-management/small-business/sb_profile.pdf

3 First Glance: Regional Structural Break of Small Business Growth

3.1 Data Processing and Statistical Description

All the economic growth data, including the growth of small business amount that we focused on, are significantly affected by global macro-economics. To isolate the impact of 2010 Vancouver Winter Olympics, it's necessary to rule out those external sources of economic variation (e.g., the financial crisis in 2008). Therefore, we did a regression analysis on growth rate g_t as below, using Canadian GDP growth rate data from the World Bank[2].

$$g_t = \alpha + \beta g_t^{\text{Canada}} + \epsilon_t \tag{1}$$

After the regression, we took $g_t^{\text{adjust}} = g_t - \hat{\beta} g_t^{\text{Canada}}$ as the adjusted growth rate. We omitted the superscript "adjust" hereafter for convenience purposes. However, it's worth noting that we conducted this adjustment for all kinds of economic growth data.

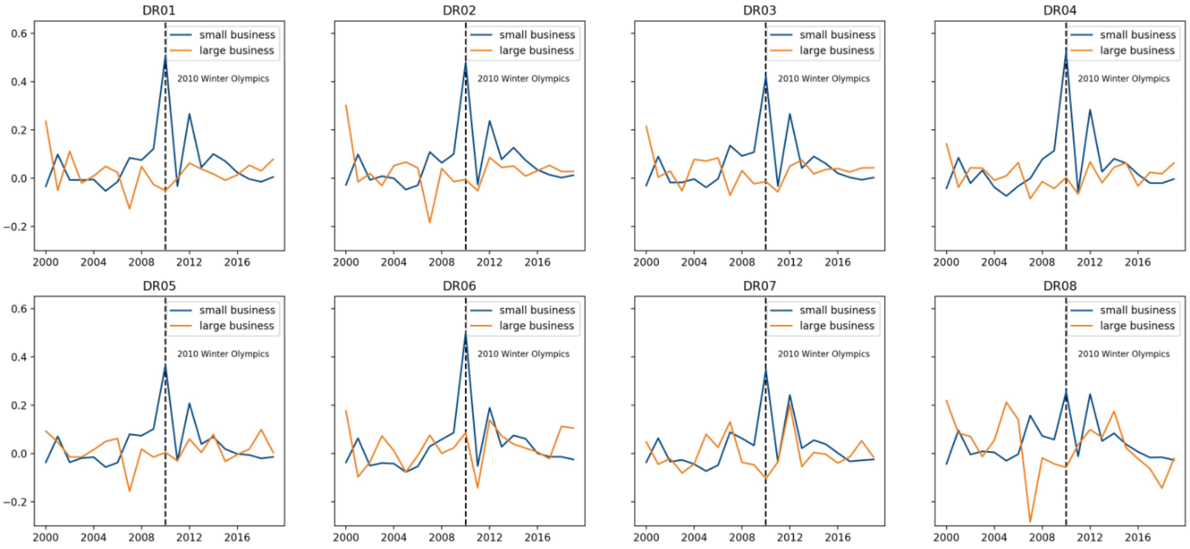


Figure 2: Growth of small and large business amount in different regions

Note: The growth rate is calculated by removing the global macro-economics impact. DR01 - DR04 and DR06 experienced a significant small business booming in 2010, while others were not affected so much by the 2010 Winter Olympics.

Figure 2 shows the growth rate of the small and large business amount in different regions, from 2000 to 2019. We noticed that the amount of small businesses is not as stable as large businesses. In most RDs, small business growth rate went up rapidly in 2010, followed by a drop-down in 2011 and a rebounce in 2012. These oscillations form a more volatile time series structure, which can be captured by a structural break test introduced later. Another observation is, the impact on small businesses is nonuniform among regions. DR05, DR07, and DR08 seem not to be affected so much.

[2] https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=CA

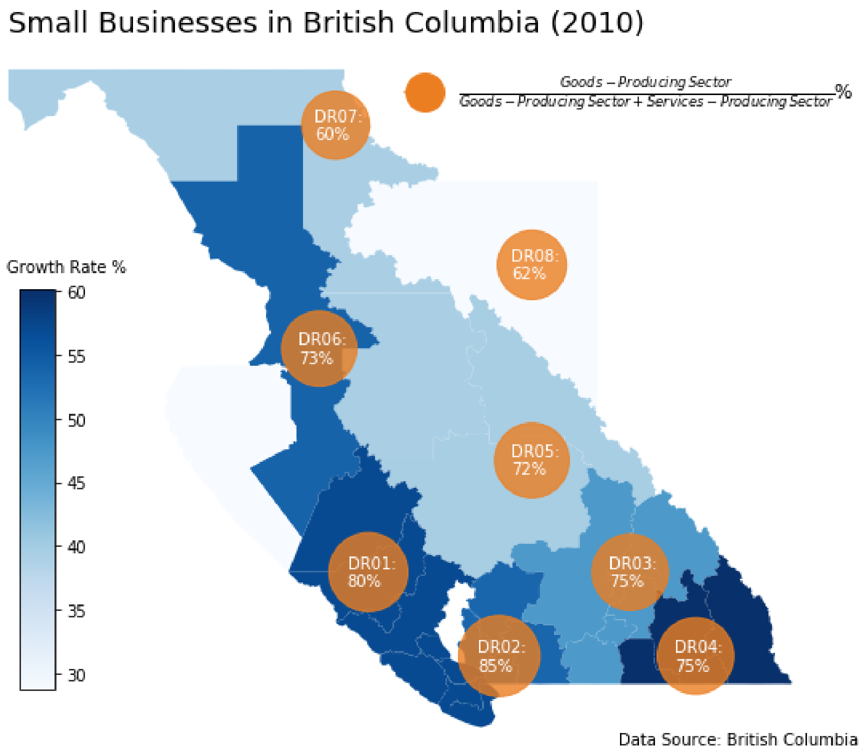

Small Businesses in British Columbia (2010)
Goods – Producing Sector / Goods – Producing Sector + Services – Producing Sector %
Growth Rate %
60
55
50
45
40
35
30
DR07: 60%
DR08: 62%
DR06: 73%
DR05: 72%
DR01: 80%
DR03: 75%
DR02: 85%
DR04: 75%
Data Source: British Columbia

Figure 3: 2010 Regional growth of small business and industry compositionss of British Columbia

Note: This figure displays the growth of small business amount in 2010. Deeper color indicates higher growth. We observed that the growth is usually higher in the region with higher proportion of business belonging to goods-producing sector.

Figure 3 displays the 8 RDs of British Columbia (DR01-DR08), including Vancouver Island/Coast (DR01), the host city of the 2010 Winter Olympics. The growth rate of small business amount in 2010 is positively correlated with the proportion of the goods-producing sector, which motivates us to study the relationship between industry composition and regional growth rate difference.

All these interesting phenomenons can be well-explained by the following statistical test and industry clustering. We quantified the change of time series structure in Section 3.3 and attributed the regional difference to industry distribution in Section 4.3.

3.2 Statistical Model

3.2.1 Auto-Regression (AR) Model

Intuitively, we assumed the growth rate of small business amount has an auto-regression structure, which can be expressed as

$$g_t = \alpha_0 + \alpha_1 g_{t-1} + \epsilon_t \tag{2}$$

where we only consider the AR(1) model, and the noise ϵ_t are independent identically distributed (i.i.d) and normally distributed. This linear structure is convenient for us to conduct statistical test of structural break and interpret the meaning of parameters α_0 and α_1.

3.2.2 Quandt Likelihood Ratio (QLR) Test

As mentioned above, we utilized Quandt likelihood ratio (QLR) test to verify whether there is a structural break between $\tau_0 = 2000$ and $\tau_1 = 2019$. In short, the QLR test is a point-wise Chow test of the linear model. For each given time $\tau \in [\tau_0, \tau_1]$, we split the test period into two intervals $[\tau_0, \tau]$ and $(\tau, \tau_1]$ and assumed

$$\begin{aligned} g_t &= \alpha_0 + \alpha_1 g_{t-1} + \epsilon_t, \quad t \in [\tau_0, \tau] \\ g_t &= \beta_0 + \beta_1 g_{t-1} + \epsilon_t, \quad t \in (\tau, \tau_1] \end{aligned} \tag{3}$$

Then we tested for the null hypothesis $H_0 : \alpha_0 = \beta_0, \alpha_1 = \beta_1$ using F-statistics. After going through all the possible break point τ, we picked the point with the highest F-value (lowest p-value) as the most likely structural breakpoint.

3.3 Empirical Analysis and Data Visualization

The result of QLR test is shown in Figure 4. We noticed that

1. DR01, DR04, and DR06 experienced a significant structural break in 2010 with p-values of around 0.010. Other regions, including the whole British Columbia, didn't significantly change in terms of time series structure, with a p-value larger than 0.200.

2. If structural break did happen in these insignificant regions, some of the breakpoints should have occurred in 2006, 4 years before the 2010 Winter Olympics.

These regional differences of structural break motivated us to further think of the underlying reasons. We conjectured that the 2010 Winter Olympics increased demands of some specific industries, and these industries located in various regions, which contributed to different small business growth. We justified our arguments in the following sections.

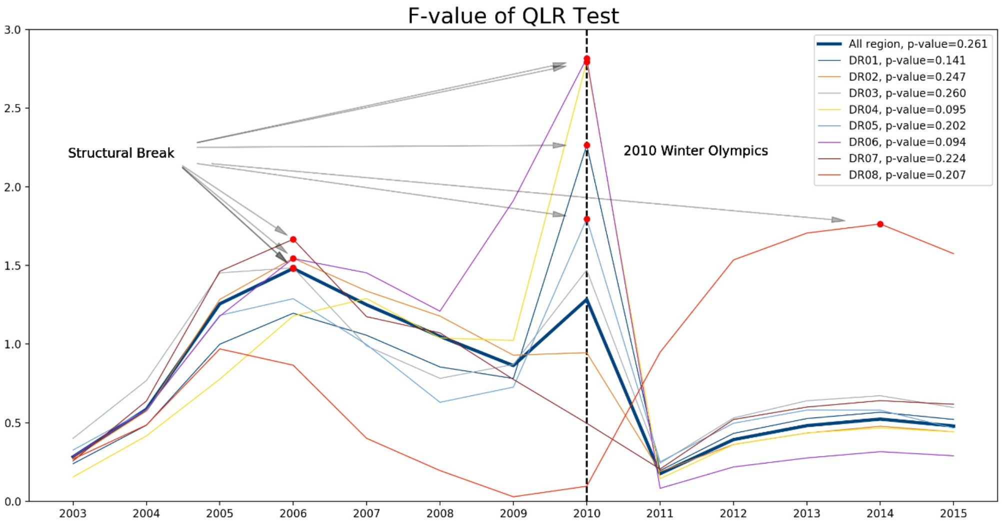


Figure 4: F-value of QLR test of different regions

Note: The QLR test is applied to the growth rate of small business amount, after ruling out the global macro-economics impact by regression. Different structural breakpoints are displayed.

4 Deeper Understanding: Winter Olympics Industry Clustering

4.1 Data Processing and Statistical Description

4.1.1 The Adjustment of Industry Labels

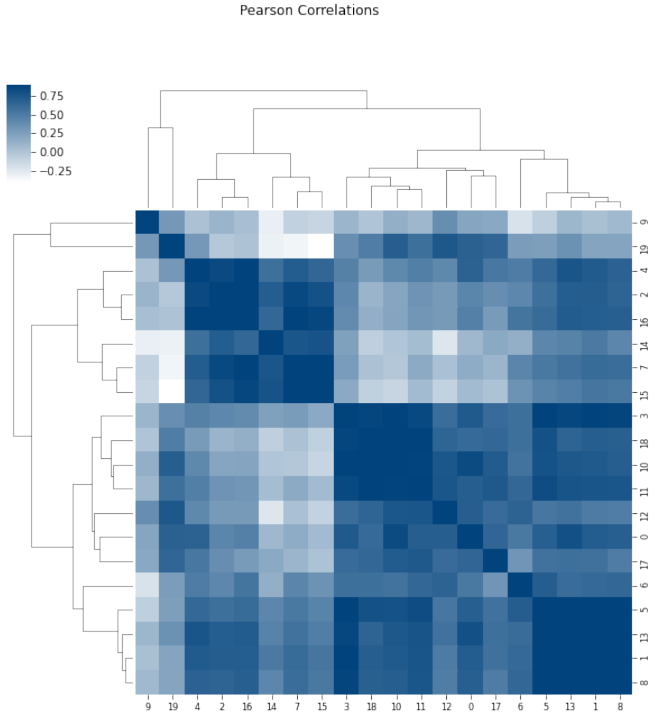
We noticed that labels of the same industries changed in the past 20 years. For instance, the industry called “Accom. & Food Services” was renamed to “Accommodation & Food Services” after 2002. We manually modified the names of industries to ensure the continuity of the data. We also ruled out the “Unclassified” industry, which only occurred after 2015.

After washing and aggregating the data, we got 20 industries in total. Their name and correlation matrix of small business growth rates are shown in Figure 5. This correlation matrix is vital for us to identify which industries are sensitive to the 2010 Winter Olympics.

4.1.2 Motivation from Real Estate & Rental & Leasing industry

After a glimpse at the industry growth datasets, we observed that the Real Estate & Rental & Leasing industry had the highest growth rate of small business amount, among 20 industries in 2010. Therefore, we focused on the monthly room revenues data to explore if the rental and leasing demand of rooms in the 2010 Winter Olympics leads to the higher growth rate.

Figure 6 indicates the Winter Olympics functions as a structural break point here, which relates to the increasing demand for small businesses. This motivated us to look at more industries, exploring which industries are sensitive to the 2010 Winter Olympics.



0	Accommodation & Food Services
1	Admin. & Support, Waste Mgmt. & Remedtn.
2	Agriculture, Forestry, Fishing & Hunting
3	Arts, Entertainment & Recreation
4	Construction
5	Educational Services
6	Finance & Insurance
7	Health Care & Social Assistance
8	Information & Cultural Industries
9	Management of Companies & Enterprises
10	Manufacturing
11	Mining & Oil & Gas Extraction
12	Other Services (excl. Public Administration)
13	Professional, Scientific & Technical Services
14	Public Administration
15	Real Estate & Rental & Leasing
16	Retail Trade
17	Transportation & Warehousing (48-49)
18	Utilities
19	Wholesale Trade

Figure 5: The Pearson correlation between industries

Note: The figure on the left shows the Pearson correlation of small business growth between industries. The table on the right lists the name of all the industries.

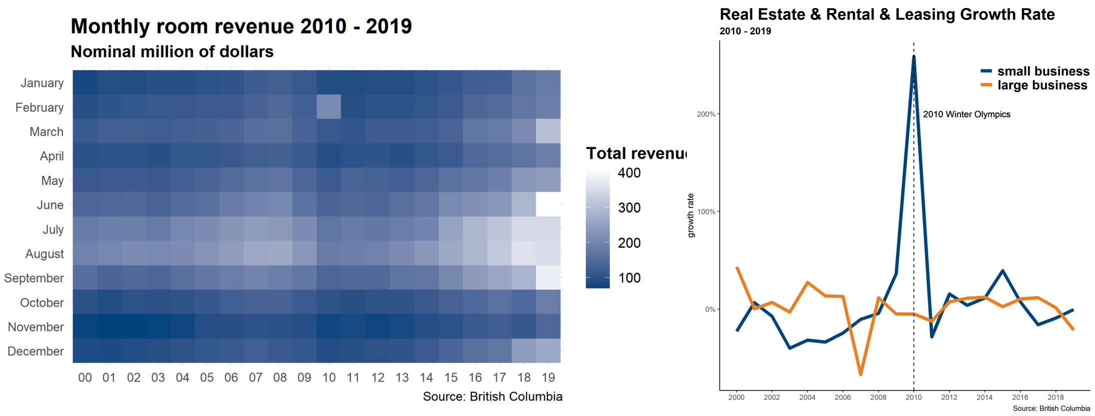


Figure 6: Monthly room revenues and business growth in Real Estate & Rental & Leasing

Note: The figure on the left shows the monthly room revenues, with a significant increase in February 2010. The figure on the right shows the growth of small and large business amount in Real Estate & Rental & Leasing, after adjustment by regression.

4.2 Spectral Clustering: A Similarity Analysis between Industries

Our structural break detection model (AR1) implies the assumption that the Winter Olympic games only affect the expectation of the growth rate of business size, while the correlation structure keeps constant. From this perspective, it is reasonable to treat the covariance matrix as a metric of the similarity of industries. We noticed that this is a perfect scenario to conduct spectral clustering since we can use Gaussian kernel to generate the affinity matrix.

4.3 Empirical Analysis and Data Visualization

4.3.1 Winter Olympics Industry Clustering

As we discussed above, we concentrated on the correlation between industries. The clustering result is shown in Table 1.

We were glad to observe that most of the clustering results are predictable based on economics sense. For example, as we mentioned above, we can expect the blooming of construction and real estate industries during the Olympic games since the city built new stadiums and infrastructures. However, we also noticed some surprising results. For example, we can hardly believe the Agriculture, Forestry, Fishing & Hunting is also sensitive to the Olympic games. We conjectured that this may be related to the supply chains between industries but we can not dig deeper into this topic without the support of relevant data.

Cluster 1 (high sensitivity)	**Cluster 2 (low sensitivity)**
Admin. & Support, Waste Mgmt. & Remedtn.	Accommodation & Food Services
Agriculture, Forestry, Fishing & Hunting	Management of Companies & Enterprises
Educational Services, Finance & Insurance	Mining & Oil & Gas Extraction
Health Care & Social Assistance	Other Services (excl. Public Administration)
Professional, Scientific & Technical Services	Transportation & Warehousing
Public Administration	Arts, Entertainment & Recreation
Retail Trade, Construction	Utilities, Wholesale Trade
Information & Cultural Industries	Manufacturing
Real Estate & Rental & Leasing	

Table 1: Spectral Clustering based on the sensitivity to 2010 Winter Olympics

We can also conclude that our clustering algorithm successfully captures the properties of industries, especially their sensitivity to the Winter Olympic games, based on the business size growth rate and QLR statistics over two clusters. The QLR test statistics and the corresponding p-values are shown in Figure 7.

4.3.2 Decomposition of Regional Growth

Our clustering algorithm shows that the regions with high sensitive industries are more likely to bloom under the influence of the Olympics. From this perspective, we argued that the cities with industries in cluster 1 are more suitable to bid for the Winter Olympic Games. To quantify this argument, we decomposed the growth of a given region into contributions of

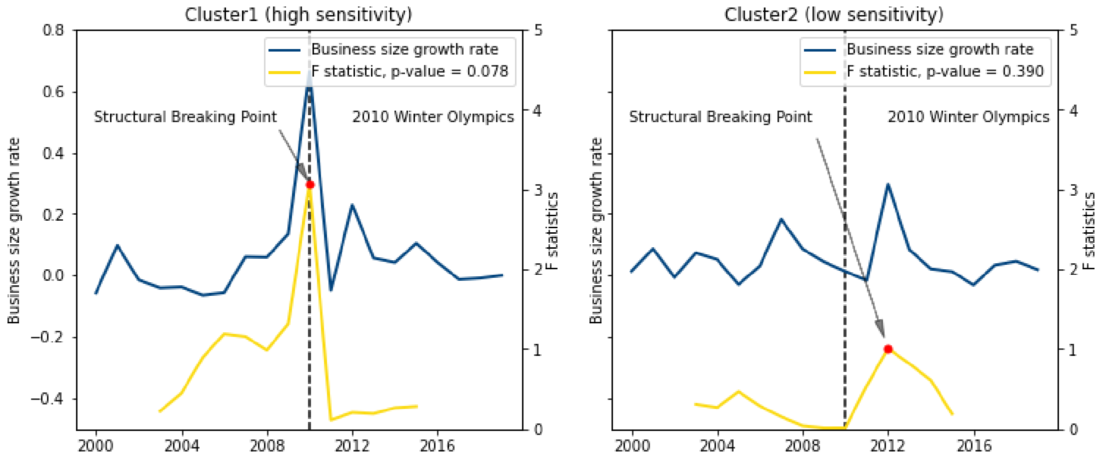


Figure 7: Business size growth rates and QLR test over clusters

Note: The p-value in the legends denotes the our confidence about the detection of structural break points. The smaller p-value of cluster 1 provides us with a significant evidence that growth rate jumped in the 2010 Winter Olympic games.

different industries.

$$g_{\text{region}} = \frac{\mathrm{d}y_{\text{region}}}{y_{\text{region}}} = \sum_{\text{industry}} \frac{y_{\text{industry}}}{y_{\text{region}}} \times \frac{\mathrm{d}y_{\text{industry}}}{y_{\text{industry}}} = \sum_{\text{industry}} w_{\text{r,i}} g_{\text{industry}} \tag{4}$$

where $w_{\text{r,i}}$ is the proportion of industry i in region r and g_{industry} is the small business growth rate of industry i. From the perspective of clustering, we can rewrite this formula as

$$g_{\text{region}} = \frac{\mathrm{d}y_{\text{region}}}{y_{\text{region}}} = \sum_{\text{cluster}} \frac{y_{\text{cluster}}}{y_{\text{region}}} \times \frac{\mathrm{d}y_{\text{cluster}}}{y_{\text{cluster}}} = \sum_{\text{cluster}} w_{\text{r,c}} g_{\text{cluster}} \approx w_{\text{r,c}_1} g_{\text{cluster}_1} \tag{5}$$

Since cluster 2 is not sensitive to the 2010 Winter Olympics, we omitted its contribution in (5). This formula identifies two critical points of small business growth:

1. The industry structure, which can be represented by the proportion $w_{\text{r,c}_1}$ of cluster 1 industries.

2. The impact on some specific industries (cluster 1 industries in our report), which can be represented by the cluster 1 growth rate g_{cluster_1}.

The proportion of cluster 1 in different regions and small business growth rate in 2010 are shown in Figure 8. The high correlation between these two quantities is convincing evidence of our decomposition formula (5). Furthermore, Table 2 indicates our clusters are better than the official business sector partition because our clusters focus on the sensitivity of the Winter Olympics, which provides a better understanding of the relationship between regions, industries, and the Olympics.

	Cluster 1 proportion	Goods-producing sector proportion
Correlation with g_{2010}	0.8463	0.7727

Table 2: Correlation with growth rate of small business amount in 2010

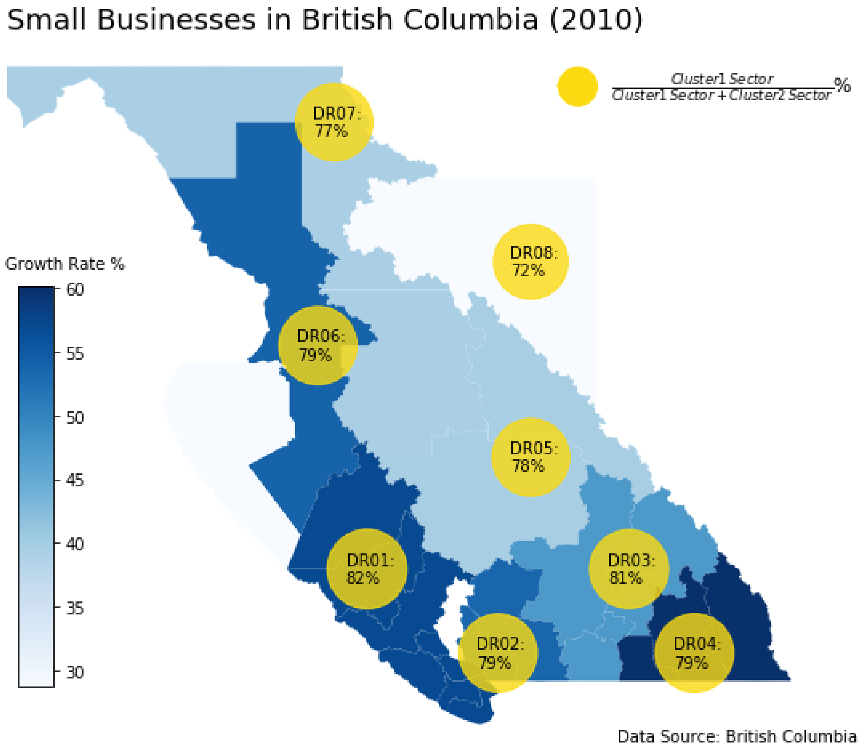

Small Businesses in British Columbia (2010)
$\frac{Cluster1\ Sector}{Cluster1\ Sector + Cluster2\ Sector}\%$
Growth Rate %
60
55
50
45
40
35
30
DR07: 77%
DR08: 72%
DR06: 79%
DR05: 78%
DR01: 82%
DR03: 81%
DR02: 79%
DR04: 79%
Data Source: British Columbia

Figure 8: 2010 Regional growth of small business and cluster 1 proportion

Note: This figure displays the growth of small business amount in 2010. The deeper color indicates higher growth. Cluster 1 proportions in different regions are shown in yellow circles, which has a higher Pearson correlation with the growth rate in 2010 than the goods-producing sector proportion in Figure 3.