Analysis of E-taxi Market and Policies in Different Cities

Fan Jinlin

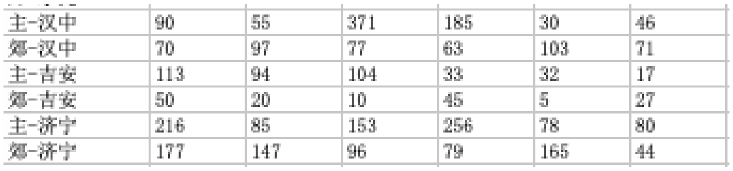
2019-Spring: AAS Final Report

1 Introduction

E-taxies are more and more important in people's daily choices of transportations. Not only for the convenience but also for the cheaper price and much better customer experience. Recently, lots of events come in front of public attention. Some of them are good, while some of them are bad. Taking the safety of passengers into consideration, government and relevant companies publish some items and lists to control the quality of the E-taxies and drivers.

All the analysis in this project will focus on some senario under this background based on true data collected by students in Law School of RUC. These datasets contain information about the requirments to taxis in several cities, news about certain events happened in several cities and competition between different taxi companies.

To get a deeper insight about the balance or the relationship between government's supervision strength and lobbying power of market and commercial cooperations, I will try to construct a frame for attaining quantified strength of supervision and monopolization of taxi companies, which can be used for further clustering of cities in order to find out similar patterns about cities in one same cluster. Some of the results will make sense, and others may be surprising, but useful for leading us to a brand new branch in the logical chain.



主-汉中	90	55	371	185	30	46
郊-汉中	70	97	77	63	103	71
主-吉安	113	94	104	33	32	17
郊-吉安	50	20	10	45	5	27
主-济宁	216	85	153	256	78	80
郊-济宁	177	147	96	79	165	44

Figure 1: Some samples in dataset1

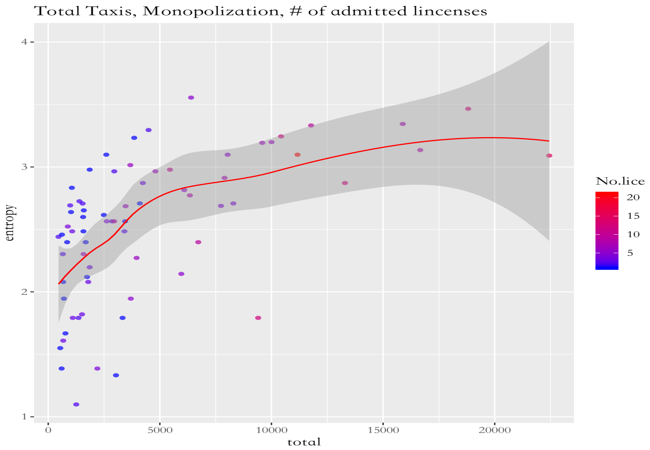


Figure 2: This figure shows that larger amount of taxies in a city means smaller monopolization power(larger entropy), and more companies can reduce the monopolization also.

2 EDA

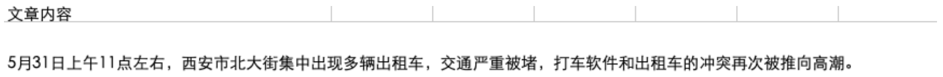
In this part, I will show what my datasets look like, and breifly express some results after exploratory data analysis. Basic tools for data processing also will be referred if necessary.

One excel table includes the total number of taxis for different companies, and the companies are divided into two groups, downtown or not. Some samples are in Figure 1.

Motivated by the entropy, which is a core concept in information theory, and defined as

$$l = -\sum_{i} p_i log p_i.$$

Entropy is used to measure unstability in a dynamic system, and in statistical cases, if n observations come from a uniform distribution, then the entropy



文章内容
5月31日上午11点左右，西安市北大街集中出现多辆出租车，交通严重被堵，打车软件和出租车的冲突再次被推向高潮。

Figure 3: One news in dataset

of these n obs attain the maximum. On contrary, if the entropy is smaller, the latent distribution is further from uniform, which means the system has kind of information about certainty, such a thing can also be explained as monopalization power in an unfair market.

In Figure 2, there are three features, every point stands for a city. And the color means the number of permitted taxi companies in this city. The fitted curve

Then I can also simply show datasets containing news in Figure 3. Except the content of news, the datasets also contain cities appearing in this news. Text-mining methods for these texts will be discussed in next section.

I use a map to express basic numeric features about different cities in Figure 4. This graph capture some advantages of using map and it contain many other information. Usually, such a map-type figure, sometimes, may express many other interesting patterns exsiting in geological aspects. But in my case, because of lack of data for other cities, there isn't any significant pattern we can get only from this graph.

3 text-mining and LDA

4 Clustering

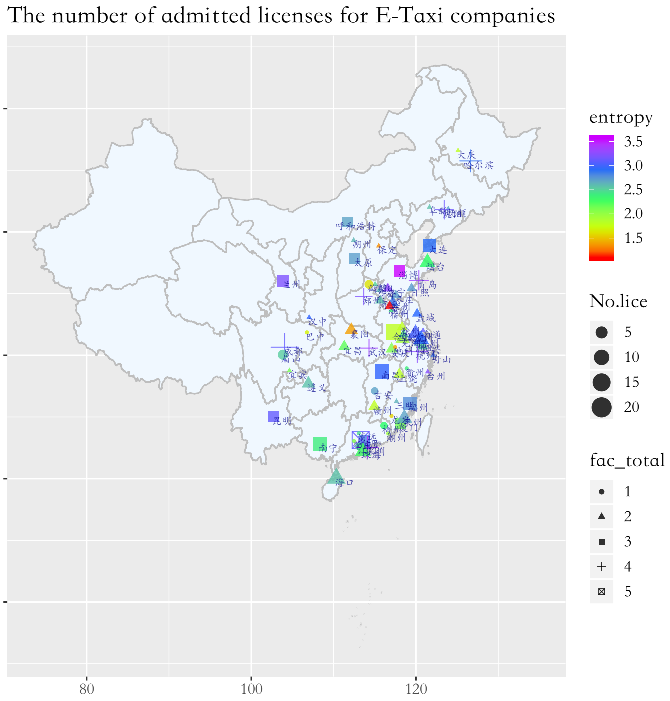


Figure 4: This graph contain more geological information. Although not significant, one interesting thing is that, severe monopolaztion cases usually are not big cities. Most of them are second-level or lower-level cities.