**Appendix**

**Content analysis of documents from *China Judgment Online* using LDA model**

# *Methodology and estimation procedure of LDA algorithm*

Since Blei et al. (2003), the Latent Dirichlet Allocation model is well developed in natural language process (NLP). We use this algorithm to estimate the causes of each administrative litigation case. This algorithm discovers highly substantive and coherent topics, and measurement errors in estimated topic proportions are likely to be small. There are two advantages to using LDA algorithm. First, this algorithm clusters words strictly according to their co-occurrence patterns, thus avoiding the arbitrariness and errors in hand coding. Second, using an unsupervised machine learning algorithm tremendously reduces the cost of parsing thousands of lengthy and dry documents, making these documents more operable in quantitative analysis. As concluded by (Blei 2012), an LDA algorithm has been widely used in multiple disciplines, including political science, psychology, population genetics, computer vision, etc.

Before conducting LDA analysis, we need to carry out word segmentation on the judgment documents. We apply the jieba segmentation module to the corpus. The jieba module, which has been proved to be much faster than other packages, combines an existing dictionary of Chinese words and the ability to learn new words from the text. The segmented text is naturally tokenized, and each piece becomes a token in subsequent computation. After word segmentation, we need to remove all punctuation, numbers and a standard list of Chinese stop words. This stop words list, developed and maintained by the Harbin Institute of Technology, includes words such as some, what, he, she, as well as thousands of meaningless words.

After retrieving the tokens, we can then run the LDA algorithm. The data generating process is as follows:

A document contains words. Each word is a member of the vocabulary of words. There are topics, . has to be specified by the modeler. Each topic is a distribution over the words. These are characterized by a matrix , where = .

*Step 1:*The term distribution is drawn for each topic from a Dirichlet distribution with parameter

*Step 2:* The topic proportions of document are drawn from a Dirichlet distribution with parameter

so that , where is the proportion of topic .

*Step 3:*For each of the words ,

(a) Choose a topic .

(b) Choose a word from , a multinomial probability conditioned on the topic .

# *Results of the LDA algorithm*

One needs to specify the number of topics for the model. In the following analysis, we use results from both 10 topics and 20 topics. Here, we present the results for a 10-topics model.

Appendix Table 1: **Results of the LDA Model**

|  |  |  |
| --- | --- | --- |
| **Num.** | **Feature Words (in Chinese)** | **Feature Words (in English)** |
| 1 | '0.007\*"村" + 0.006\*"标准" + 0.006\*"合法" + 0.006\*"判决" + 0.005\*"程序" + 0.005\*"通知" + 0.005\*"意见"' | '0.007\*"Village" + 0.006\*"Standard" + 0.006\*"legitimate" + 0.006\*"judgment" + 0.005\*"program" + 0.005\*"notify" + 0.005\*"opinion"' |
| 2 | '0.007\*"合法" + 0.005\*"标准" + 0.005\*"答辩" + 0.005\*"委员会" + 0.005\*"法律" + 0.005\*"符合" + 0.004\*"撤销"' | '0.007\*"legitimate" + 0.005\*"Standard" + 0.005\*"reply" + 0.005\*"Committee" + 0.005\*"law" + 0.005\*"conform" + 0.004\*"Revoke"' |
| 3 | '0.005\*"标准" + 0.005\*"补偿款" + 0.004\*"程序" + 0.004\*"宅基地" + 0.004\*"协议书" + 0.004\*"登记" + 0.004\*"同意"' | '0.005\*"Standard" + 0.005\*"compensation" + 0.004\*"program" + 0.004\*"Homestead" + 0.004\*"Agreement" + 0.004\*"Register" + 0.004\*"agree"' |
| 4 | '0.006\*"内容" + 0.006\*"法律" + 0.006\*"合法" + 0.005\*"程序" + 0.005\*"Revoke" + 0.005\*"村" + 0.005\*"大闸"' | '0.006\*"content" + 0.006\*"law" + 0.006\*"legitimate" + 0.005\*"program" + 0.005\*"Revoke" + 0.005\*"Village" + 0.005\*"Big gate"' |
| 5 | '0.007\*"标准" + 0.006\*"意见" + 0.005\*"合法" + 0.005\*"程序" + 0.005\*"真实性" + 0.004\*"通知" + 0.004\*"答辩"' | '0.007\*"Standard" + 0.006\*"opinion" + 0.005\*"legitimate" + 0.005\*"program" + 0.005\*"Authenticity" + 0.004\*"notify" + 0.004\*"reply"' |
| 6 | '0.008\*"标准" + 0.006\*"一审" + 0.006\*"内容" + 0.006\*"通知" + 0.005\*"办事处" + 0.005\*"村委会" + 0.005\*"判决"' | '0.008\*"Standard" + 0.006\*"first instance" + 0.006\*"content" + 0.006\*"notify" + 0.005\*"office" + 0.005\*"Village Committee " + 0.005\*"judgment"' |
| 7 | '0.005\*"真实性" + 0.005\*"村" + 0.005\*"内容" + 0.005\*"合法" + 0.004\*"程序" + 0.004\*"开发区" + 0.004\*"涉及"' | '0.005\*"Authenticity" + 0.005\*"Village" + 0.005\*"content" + 0.005\*"legitimate" + 0.004\*"program" + 0.004\*"Development zone" + 0.004\*"Involved"' |
| 8 | '0.005\*"补偿款" + 0.005\*"承包" + 0.005\*"面积" + 0.005\*"标准" + 0.005\*"合法" + 0.004\*"查明" + 0.004\*"一审"' | '0.005\*"compensation" + 0.005\*"contract" + 0.005\*"area" + 0.005\*"Standard" + 0.005\*"legitimate" + 0.004\*"Find out" + 0.004\*"first instance"' |
| 9 | '0.006\*"程序" + 0.005\*"一审" + 0.005\*"合法" + 0.005\*"意见" + 0.005\*"通知" + 0.005\*"承包" + 0.004\*"内容"' | '0.006\*"program" + 0.005\*"first instance" + 0.005\*"legitimate" + 0.005\*"opinion" + 0.005\*"notify" + 0.005\*"contract" + 0.004\*"content"' |
| 10 | '0.006\*"答辩" + 0.006\*"城市" + 0.005\*"合法" + 0.004\*"判决" + 0.004\*"承包" + 0.004\*"撤销" + 0.004\*"意见"' | '0.006\*"reply" + 0.006\*"city" + 0.005\*"legitimate" + 0.004\*"judgment" + 0.004\*"contract" + 0.004\*"Revoke" + 0.004\*"opinion"' |

# *Reclassification for clustering topics*

After retrieving the results of the LDA model, we can identify the closest topic of each document according to the feature words. Based on the theoretical background, we reclassify the above ten groups into four topics: inadequate compensation, illegal land expropriation, illegal procedure, and others. The reclassification results are shown in Figure1.

Figure 1: **Reclassification of LDA Results**

|  |
| --- |
|  |

To make it easier to understand, we supplement three examples corresponding to three clear topics as follows. Since the whole documents are very long, we only intercepted the core part of them.

*Inadequate compensation*

原告诉称，亭湖区政府作出的《关于公布〈纺织厂地块棚户区改造项目二期房屋征收补偿方案（征求意见稿）〉并征求意见的公告》（以下简称《征收补偿方案公告》）违反法律规定。请求法院依法确认亭湖区政府作出的《征收补偿方案公告》违法，判令亭湖区政府中止征收拆迁并承担诉讼费用。事实与理由：1、征收补偿方案没有国务院及省政府征地批文红头文件。亭湖区政府没有江苏省征地批文红头文件，违反法律规定。2、征收补偿价远低于市场价。仁恒置业盐城城投集团、仁恒滨河世纪高档小区楼盘均价25000元左右，现在拆迁办给其定的拆迁补偿单价为8000元＋2600元，共10600元，远低于市场价。

The original complaint alleged that the Tinghu District Government’s “Announcement on the announcement and solicitation of the ‘Announcement of the Second Phase of Housing Expropriation and Compensation Plan (draft for comment)’” (hereinafter referred to as the “Announcement of Expropriation and Compensation Plan”) made by the Tinghu District Government violated the law. The court requested the court to confirm that the “Announcement of Expropriation and Compensation Plan” made by the Tinghu District Government was illegal and ordered the Tinghu District Government to suspend the expropriation and demolition and bear litigation costs. Facts and reasons: 1. The expropriation compensation plan does not have a red-headed document for land expropriation approval by the State Council and the provincial government. The government of Tinghu District did not have a red-headed document of approval for land acquisition in Jiangsu province, which violated the law. 2. The requisition compensation price is far lower than the market price. The average price of real estate in Yancheng Urban Investment Group and Yanlord Binhe Century High-end Community is about 25,000 yuan. Now the demolition compensation unit set by the Demolition Office is 8,000 yuan + 2,600 yuan, a total of 10,600 yuan, which is far lower than the market price.

*Illegal land expropriation*

被告在征地过程中存在未批先征，少批多征、非法征用基本农田的违法行为。省政府批准征收筒车、丁家、南岸三个村的土地面积共计652.0155亩，而被告实际却征收了二千余亩耕地。在省政府批准征收的土地中有80％以上属于基本农田。根据《土地管理法》第四十五条的规定，征收基本农田应由国务院批准。被告未经国务院批准，征收基本农田的行为明显违反了国家法律的规定；被告虽然有省政府的部分征地批复，但明显存在少批多征，属于越权、滥用行政权力的行为，严重损害了原告的合法权益，特诉请法院确认被告犍为县人民政府、犍为县国土资源局、犍为县清溪镇人民政府征收原告承包耕地和本集体经济组织集体土地、宅基地的行为违法。

In the process of land requisition, the defendant committed illegal acts such as unapproved prior requisition, less requisition for more requisition, and illegal requisition of basic farmland. The provincial government approved the expropriation of a total of 652.0155 *mu* of land in the three villages of Tongche, Dingjia, and Nan’an, but the defendant actually expropriated more than 2,000 *mu* of cultivated land. More than 80 per cent of the land approved by the provincial government is basic farmland. According to Article 45 of the Land Administration Law, the expropriation of basic farmland shall be approved by the State Council. The defendant’s expropriation of basic farmland without the approval of the State Council clearly violated the provisions of national laws; although the defendant had partial land expropriation approval from the provincial government, there were obviously fewer approvals and excessive expropriation, which was *ultra vires* and abuse of administrative power, which seriously harmed the plaintiff. The court specifically appealed to the court to confirm that the defendants Qianwei County People’s Government, Qianwei County Land and Resources Bureau, Qianwei County and Qingxi Town People’s Government’s expropriation of the plaintiff’s contracted farmland and collective land and homesteads of the collective economic organization violated the law.

*Illegal procedure*

原告家庭承包的土地未经合法征收程序转为国有土地的情况下，荥经县城乡规划建设和住房保障局等单位擅自在原告家庭承包的土地上施工建设，违反了《土地管理法》关于建设用地应当使用国有土地等相关规定。根据《土地管理法》的规定，原告依法向被告提出申请，请求被告查处违法用地行为。2016年9月29日，原告收到被告作出的答复。

When the land contracted by the plaintiff’s family was converted to state-owned land without legal expropriation procedures, Yingjing County Urban and Rural Planning and Construction and Housing Security Bureau and other units arbitrarily constructed construction on the land contracted by the plaintiff’s family, violating the Land Management Law regarding construction. The land should use state-owned land and other relevant regulations. According to the provisions of the Land Administration Law, the plaintiff filed an application with the defendant in accordance with the law, requesting the defendant to investigate and deal with illegal land use. On 29 September 2016, the plaintiff received a reply from the defendant.

**References:**

Blei, David M. 2012. “Probabilistic topic models.” *Communications of the ACM* 55(4), 77–84.

Blei, David M, Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research* 3, 993–1022.