

Received September 29, 2018, accepted October 21, 2018, date of publication October 23, 2018, date of current version November 30, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2877716

Incorporating LDA Based Text Mining Method to Explore New Energy Vehicles in China

SUSAN (SIXUE) JIA¹⁰ AND BANGGANG WU²

¹School of Business and Management, Shanghai International Studies University, Shanghai 200083, China ²Economics and Management School, Wuhan University, Wuhan 430072, China

Corresponding author: Banggang Wu (banggangwu@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 71702107, in part by Shanghai International Studies University under Grant 2015114050, in part by the National Science Fund for Outstanding Young Scholars under Grant 71325005, in part by the National Natural Science Foundation of China (Key Program) under Grant 71532006, and in part by the Ministry of Education, Humanities and Social Science Research Base Project under Grant 16JJD630006.

ABSTRACT Tracking the evolution of policy and development of new energy vehicles (NEVs) in China is of critical significance, because it helps generate rational prediction regarding future trends. To this end, this paper investigated the 5185 articles on NEV obtained from China National Knowledge Infrastructure by means of latent Dirichlet allocation (LDA)-based text mining. Word count was performed to highlight important keywords for different periods of years of publication. In addition, topics were identified from the abstracts of these articles using LDA. Findings suggest that attention on NEV in China has been growing and will continue to grow in the predictable future. Full electric vehicle, being the currently dominating form of NEV, will continue to play the leading role. Meanwhile, China's NEV industry requires further investment into charging, battery, personnel training, and patent portfolio.

INDEX TERMS Battery, China, full electric vehicle, LDA, new energy vehicle, text mining.

I. INTRODUCTION

New energy vehicles (NEVs) are the collective ensemble of alternative fuel (ethanol, biogas, biodiesel) vehicles, electric vehicles, hybrid electric vehicles, and fuel cell vehicles [1], [2]. NEVs not only have the advantage of fuel oil energy conservation and pollutant emission reduction [3], but also play the role of sustaining the automotive and transportation industry of a nation [4]. For China, developing and promoting NEVs are of even greater significance, considering the country's existing huge volume of production [5], the prediction of increasing future demand [6], as well as the strategic concern of energy safety [7]. Therefore, it is extremely meaningful to track the evolution of policy and development of NEVs in China so as to generate rational prediction regarding future trends.

One efficient type of way of analyzing NEV evolution is by discovering knowledge from existing data source [8]–[12]. For instance, Guo and Liu [8] employed social network analysis to construct China's NEV industry alliance network based on gathered internet information, the process of which is sometimes called "netnography" [11]. Wang *et al.* [9] revealed the impact of government subsidy on NEV market performance with the help of an automobile database containing model and production information. Both these studies [8], [9] dealt with relatively structured data, i.e., numbers and pre-determined company names.

To tackle less structured information such as natural language, researchers use either content analysis [10] or text mining [11], [12]. For example, Xie and Tian [10] manually categorized the NEV policy instruments from 37 policy texts. Sung and Park [12] uncovered the high impact words embedded in the 265 million pieces of online news and documents about renewable energy by applying term frequency-inverse document frequency (TF-IDF), an automatic word frequency and impact analyzing tool. Moreover, Zeng [11] leveraged latent Dirichlet allocation (LDA) algorithm, a topic identification method, as a foresight tool in the renewable energy industry. More advanced text clustering methods include vector space model [13], hybrid particle swarm optimization algorithm with genetic operators [14], and krill herd algorithm [15].

Based on the significance of NEV study for China, as well as the technological readiness of analyzing the vast amount of information composed of natural language, this study aims at uncovering the evolution of policy and development of NEVs in China by means of text mining. More specifically, this paper unfolds as follows. The "data and method" section describes the data source of 5185 articles obtained from China National Knowledge Infrastructure (CNKI) as well as the LDA algorithm for data reduction. The "results and discussion" section reveals the evolution of high frequency words and major topics with year, predicts the future trend, and discusses the limitation of LDA. The potential contribution of this study manifests as the retrospect and prediction of China's NEV industry, along with an optimized protocol of structuring unstructured data.

II. DATA AND METHODS

A. DATA SOURCE

All the data analyzed in this study were collected from China National Knowledge Infrastructure (CNKI), the largest knowledge database in China based on which a number of influential researches were conducted [16]–[18]. The search conditions were set as,

- Searching for articles published in academic and industrial journals,
- And that the article title should include "new energy vehicle".

B. DATA STRUCTURE

The data structure of the obtained articles unfolds as follows,

- Title, which is the identity of an article,
- Year of publication, which was used for trend analysis,
- · Keyword, which was used for word frequency analysis,
- Abstract (or summary for some industrial journal articles), which was used for topic identification.

Article information other than the above was not the focus of this study.

C. WORD COUNT

Word count was performed to highlight important words. For a certain period of years of publication, all the keywords of the articles published in that period were counted for times of appearance. The words were then listed in a table with more frequently mentioned words arranged on the top.

D. TOPIC IDENTIFICATION

Besides investigating individual words, this study also examines word clusters, or topics, by LDA. LDA is based on the assumption of probabilistic topic model [19], [20]. It assumes the word generation in a document as a two-stage process:

(I) Randomly choose a distribution of topics.

(II) For each word in the document,

(i) Randomly choose a topic from the distribution of topics in (I).

(ii) Randomly choose a word from the corresponding distribution of the vocabulary.

In real situations, neither the distribution of topics over documents nor the distribution of words over topics is known a priori; only the documents are observed, which in this paper were the abstracts of the obtained articles. Mathematically, the connection between hidden and observed variables is the

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \cdot \prod_{d=1}^{D} p(\theta_d)$$
$$\cdot \sum_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \quad (1)$$

 β_i distribution of word in topic *i*, altogether *K* topics

 θ_d proportions of topics in document *d*, altogether *D* documents

 z_d topic assignment in document d

 $z_{d,n}$ topic assignment for the n^{th} word in document d, altogether N words

 w_d observed words for document d

 $w_{d,n}$ the n^{th} word for document d

The identification of topics and words is thus a posteriori estimation (Equation 2) using Gibbs sampling [21].

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$
(2)

In this study, the estimation was realized using Python LDA 1.0.5 [22]. To implement LDA, one needs to prepare a word-sentence matrix M from all the obtained abstracts. Suppose the abstracts have been broken into a set of sentences {S₁, S₂,, S_i,}, with S_i designating the ith sentence of all sentences. Meanwhile, {W₁, W₂,, W_j,} is a list of words that have appeared in these sentences, with W_j representing the jth word of all words. Then in the M matrix {M₁₁, M₁₂,, M_{ij},}, M_{ij}=1 would mean that sentence S_i contains word W_j, whereas M_{ij}=0 would stand for that word W_j is not in sentence S_i. With this matrix, the Python LDA 1.0.5 module is able to perform the topic identification task.

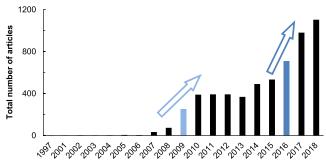


FIGURE 1. Total number of articles per year. The number for 2018 is an estimated value.

III. RESULTS AND DISCUSSION

A. INCREASING NUMBER OF ARTICLES

From CNKI, a total of 5185 articles were retrieved. The number of articles per year has been counted and summarized in Figure 1. From the figure, it can be seen that reports on NEV in China first appeared in 1997, but had not reached a significant volume until around 2007. The year 2009 witnessed the first boost of article number to approximately 400 per year. Recently in around 2016, the number

1997-2008		2009-2010		2011-2012 2013-20		2013-2014	4 2015-2016			2017-2018	
word	#	word	#	word	#	word	#	word	#	word	#
NEV	79	NEV	82	NEV	74	NEV	<mark>75</mark>	NEV	77	NEV	73
Hybrid electric vehicle	23	NEV industry	12	NEV industry	18	NEV industry	18	NEV industry	16	NEV industry	9
Fuel cell vehicle	18	Corporation	10	Full electric vehicle	17	Full electric vehicle	10	Electric vehicle	8	New energy	6
Corporation	14	Corporate management	10	Plan	9	Electric vehicle	8	Full electric vehicle	7	Electric vehicle	5
Corporate management	14	Industrialise	10	Long term plan	8	P.R.China	7	Electric cars	7	Power battery	4
Hybrid electric passenger car	10	P.R.China	9	Industrialise	8	Electric cars	6	P.R.China	6	Full electric vehicle	4
P.R.China	10	Hybrid electric vehicle	9	P.R.China	8	Energy	6	Corporate management	5	Automobile	3
Industrialise	10	EV	9	Electric vehicle	7	Popularisation	6	Corporation	5	Passenger car	3
Full electric vehicle	10	Electric locomotive	9	Plug-in hybrid	7	EV	5	Popularisation	5	New energy car	3
Automobile	10	Electric vehicle	9	Electric cars	7	New energy car	5	Charging facility	5	Charger	3
Hybrid power	8	Full electric vehicle	9	EV	6	Electric locomotive	5	Charger	5	Charging infrastructure	2
Fuel cell passenger car	7	Electric cars	8	Electric locomotive	6	Charging facility	4	Energy	4	Charging facility	2
Production admittance	7	New energy car	8	Hybrid electric vehicle	6	Corporation	4	New energy car	4	Corporation	2
Toyota	6	Energy	6	Energy	5	Corporate management	4	New energy	4	Corporate management	2
Energy	6	Hybrid electric car	6	Automobile	5	Automobile	4	Automobile	4	NEV technology	2
New energy car	6	Subsidy	6	NEV technology	5	Plug-in hybrid	4	Sales volume	4	Electric cars	2
Full electric car	5	Additional payment	6	Full electric car	4	BYD	3	Power battery	3	Popularisation	2
Electric vehicle	5	Automobile	5	Corporation	4	Subsidy	3	Charging infrastructure	3	Status quo	2
Engine	5	Hybrid power	5	Corporate management	4	Charger	3	Plug-in hybrid	3	Fuel vehicle	2
Continuous battery	5	Full electric car	4	Hybrid electric car	3	Full electric car	3	BYD	3	Full electric passenger car	1
Fuel cell	5	Key part	3	New energy car	3	New energy	3	Subsidy	2	Ministry of Ind & Info Tech	1
NEV industry	5	Consumer	3	Hybrid power	3	Industrialise	3	Model	2	Consumer	1
NEV technology	5	NEV alliance	3	Demonstration operation	3	Additional payment	3	Full electric passenger car	2	Energy	1
Electric cars	4	Plan	3	New energy	3	Hybrid power	3	Additional payment	2	Power storage battery	1
Hybrid electric car	4	New energy	3	Passenger car	3	Consumer	3	Consumer	2	Recycle	1

TABLE 1. Evolution of high frequency words (top 25) with year, converted to per every 100 articles. Words of same colors have similar meanings or are highly related. Arrows indicate the rising or falling trend.

of articles per year experienced a second booming to over 1000 per year. Overall, the number of articles is still increasing, indicating the persisting importance and attention of the NEV industry in China.

B. EVOLUTION OF HIGH FREQUENCY WORDS

The keywords of the articles published in each two-year period were counted for times of appearance, except that the years from 1997 to 2008 were treated as one period due to relatively limited number of articles. The high frequency words (top 25) of each period, converted to per every 100 articles, were demonstrated in Table 1. Table 1 shows that,

- Fuel cell vehicle was rarely mentioned in NEV articles after 2009, indicating its immaturity as an industrial product for automotive application in China even for now. Nevertheless, fundamental research regarding fuel cell vehicle is still being conducted in Chinese academia and industry.
- Hybrid electric vehicle, being an important form of NEV, became seldom mentioned after 2016,

	2013-2014								
Торіс	1. Global situation	2. Volume of production	3. Environmental protection	4. Demonstration operation	5. Subsidy				
	China	Automobile	Automobile	Automobile	Policy				
	Market	New energy	Development	Promotion	New energy				
	U.S.	Ten thousand	Energy	Operation	Automobile				
	Automobile	Growth	Environment	Public transport	Subsidy				
	Global	Volume	Economy	Private	Release				
	Tesla	Produce	Emission	Number	Government				
Word	World	Sell	Global	Government	Support				
word	Electric vehicle	Data	Pollution	Energy saving	Nation				
	Japan	Exceed	Social	Purchase	Local				
	New energy	Sum	World	Rental	Strength				
	Domestic	Goal	Pressure	Policy	Market				
	Development	Inventory	Industry	Demonstration	Promote				
	BYD	Double	Continual	Bus	Encourage				
	Germany	Half year	Petroleum	Management	Corporation				
Торіс	6. Popularisation	7. Market	8. Charging	9. Design					
	Automobile	Automobile	Charging	System					
	Popularisation	New energy	Construction	Battery					
	Notification	Market	Facility	New energy					
	Development	Consumer	Auxiliary	Design					
	Ministry of Finance	Technology	Improvement	Motor					
	State Department	Cost	Infrastructure	Drive					
Word	Ministry Sci Tech	Purchase	System	Technology	-				
wora	Industry	Product	Electric vehicle	Experiment	-				
	Opinion	Promote	Standard	Development	-				
	Release	Price	Foundation	Performance					
	Informationise	Demand	Supply	Vehicle	-				
	Energy saving	Battery	Accelerate	Platform	-				
	Ministry Ind Info	Range	Beijing	Control					
	National Dev Reform	Risk	Quick	Material]				

TABLE 2. Topics identified from the abstracts of the articles published during 2013–2014.

suggesting its role as an intermediate form of NEV is moving towards the finale in China.

- Full electric vehicle has always been frequently mentioned over the years, because it is regarded as an ideal form of NEV for China in the predictable future.
- Recent focus has been on the construction of charging facility and the development of power battery, both of which are highly related to full electric vehicle.

C. EVOLUTION OF MAJOR TOPICS

The abstracts of the obtained articles have been processed using LDA to identify major topics. Table 2 shows the identified topics from the abstracts of the articles published during 2013–2014 as an example. Altogether 9 meaningful topics were identified, with corresponding words listed below the topics. With the same approach, the topics of articles published during other periods were also identified. They were compared in Table 3 to reveal their evolution with time.

From Table 3, it can be seen that,

 Energy saving was once a motivation of NEV development. However, it is less mentioned in recent years. The reason for this is that before 2009 vehicles in China had relatively low fuel efficiency. At that time, the energy saving effect of NEV was quite significant. Later on, new generations of vehicles had much improvement on fuel economy, which gradually made energy saving a minor strength of NEV.

- Before 2014, global situation was a major topic in published articles. At that time, the NEV industry in China was still under incubation, with global firms such as Toyota as benchmarks. Things have changed since 2015 when global situation was less frequently mentioned. Instead, reports of growth of NEV popped up, suggesting the rise of the NEV industry of China.
- The preparation of NEV policy dates back to the 2000s when the appeal of NEV development was brought up to NPC (National People's Congress) and CPPCC (Chinese People's Political Consultative Conference), the two highest level administrative conferences in China. Shortly afterwards, NEV development was incorporated into the 863 Project, a high technology development plan leaded by the State Department. Meanwhile, the "production

Period	I. Background and motivation topics							
1997-2008	Energy saving	Global situation	Volume of production	Environmental protection				
2009-2010		Global situation	Volume of production	Environmental protection				
2011-2012		Global situation	Volume of production	Environmental protection				
2013-2014		Global situation	Volume of production	Environmental protection				
2015-2016			Volume of production	Environmental protection	Growth of NEV			
2017-2018			Volume of production	Environmental protection	Growth of NEV			
Period	II. Preparation and standardisation topics							

TABLE 3. Topics identified from the abstracts of the articles published during each period. The topics are grouped in four categories: I, II, III, and IV.

Period	II. Preparation and standardisation topics							
1997-2008	NPC and CPPCC	863 Project	Production admittance	Ministries				
2009-2010			Production admittance	Ministries	State department			
2011-2012				Ministries	State department			
2013-2014								
2015-2016								
2017-2018								

Period	III. Policy and popularisation topics						
1997-2008							
2009-2010	Demonstration operation	Subsidy					
2011-2012	Demonstration operation	Subsidy					
2013-2014	Demonstration operation	Subsidy	Popularisation				
2015-2016	Demonstration operation	Subsidy	Popularisation				
2017-2018		Subsidy (diminishing)	Popularisation	Credit system			

Period	IV. Market and technology topics						
1997-2008							
2009-2010	Market	Charging					
2011-2012	Market	Charging	Design				
2013-2014	Market	Charging	Design				
2015-2016	Market	Charging	Design	Personnel training	Patent		
2017-2018	Market	Charging	Design	Personnel training	Patent		

admittance" of NEV began to be authorized by the Ministry of Industry and Information Technology to automobile companies batches by batches. • From 2009 to 2016, government subsidy was an important motivation and support to automobile companies that invest heavily in the development and

manufacturing of NEVs. It is worth noting that since 2017, subsidy has been diminishing, with a credit system gradually taking the place. The credit system, also established by the Chinese government, requires that for every fuel oil car being produced, a corresponding number of NEVs have to be produced in the meantime. In other words, the economic incentives will soon transform fully into policy requirements.

• The newly increased focus of NEV has been on personnel training and patent in China. To provide better NEV products and services, Chinese researchers and industrial practitioners are accelerating the process of innovation in terms of highly efficient charging facility and service, NEV lease in the sharing economy context, and batteries of higher capacity and safer operation. Meanwhile, China has begun targeting overseas markets, which makes patent issue of increasing priority.

D. FUTURE TREND PREDICTION

Based on the above observation, the following predictions are made,

- Regarding the general trend, attention on NEV in China is expected to continue to grow in the predictable future because of the persisting motivation of environmental protection and energy safety concern.
- In terms of technology roadmap, full electric vehicle will continue to be the dominating form of NEV in China in the predictable future. Hybrid electric vehicle, being an intermediate form, will soon come to a finale in China. Fuel cell vehicle, being still under development on the fundamental level, is not expected to meet mass production in China in the near future.
- To further enhance the performance and export of full electric vehicle, investments are expected to be directed to improved charging strategy and facility, innovative battery material and system, NEV personnel training, as well as patent portfolio.

E. OVERCOMING THE LIMITATIONS OF LDA

The LDA based text mining method is not without limitations. On one hand, it only provides word clusters instead of explicit topic names. On the other hand, it only identifies major topics, meaning it can miss some important information which is not mathematically significant.

This study has successfully overcome the above limitations by taking two actions. Firstly, the authors carefully assigned names to each topic after examining the words in the clusters. By incorporating human knowledge into the text mining task, the topics made better sense and were more comparable. Secondly, the authors conducted word frequency analysis as an important complement to LDA. Word frequency analysis revealed the decreasing mentioning of fuel cell vehicle and hybrid electric vehicle, which was not captured by LDA.

VOLUME 6, 2018

IV. CONCLUSIONS

Tracking the evolution of policy and development of NEVs in China is of critical significance. For this purpose, this paper investigated the 5185 articles on NEV obtained from CNKI by means of text mining. Findings suggest that attention on NEV in China has been growing and will continue to grow in the predictable future. Full electric vehicle, being the currently dominating form of NEV, will continue to play the leading role. Meanwhile, China's NEV industry requires further investment into charging, battery, personnel training, and patent portfolio.

REFERENCES

- O. Minggao, "Chinese strategies and countermeasures for energy saving and vehicles with new types energy," *Automot. Eng.*, vol. 28, no. 4, pp. 317–322, 2006.
- [2] Y. Liu and A. Kokko, "Who does what in China's new energy vehicle industry?" *Energ Policy*, vol. 57, pp. 21–29, Jun. 2013.
- [3] Y. Li, Q. Zhang, B. Liu, B. McLellan, Y. Gao, and Y. Tang, "Substitution effect of new-energy vehicle credit program and corporate average fuel consumption regulation for green-car subsidy," *Energy*, vol. 152, pp. 223–236, Jun. 2018.
- [4] X. Zhang and X. Bai, "Incentive policies from 2006 to 2016 and new energy vehicle adoption in 2010–2020 in China," *Renew. Sustain. Energy Rev.*, vol. 70, pp. 24–43, Apr. 2017.
- [5] X. Wang, C. G. Li, J. Shang, C. Yang, B. Zhang, and X. Ke, "Strategic choices of China's new energy vehicle industry: An analysis based on ANP and SWOT," *Energies*, vol. 10, no. 4, p. 537, 2017.
- [6] J. Ren, "New energy vehicle in China for sustainable development: Analysis of success factors and strategic implications," *Transp. Res. D, Transp. Environ.*, vol. 59, pp. 268–288, Mar. 2018.
- [7] L. Zhang and Q. Qin, "China's new energy vehicle policies: Evolution, comparison and recommendation," *Transp. Res. A, Policy Pract.*, vol. 110, pp. 57–72, Apr. 2018.
- [8] X. Guo and Y. Liu, "The analysis of China new energy vehicle industry alliance status based on UCINET software," J. Comput., vol. 6, no. 9, pp. 1852–1856, 2011.
- [9] H. Wang, H. Shi, P. Chen, and M. Ouyang, "Analysis on the progress of energy saving and new energy vehicle industrialization in China based on a database," *J. Automot. Saf. Energy*, vol. 5, no. 3, pp. 294–297, 2014.
- [10] Q. Xie and Z. L. Tian, "Text analysis of China's new energy vehicle industry policy," Adv. Mater. Res., vols. 1073–1076, pp. 2499–2502, Dec. 2015.
- [11] M. A. Zeng, "Foresight by online communities—The case of renewable energies," *Technol. Forecasting Social Change*, vol. 129, pp. 27–42, Apr. 2018.
- [12] B. Sung and S.-D. Park, "Who drives the transition to a renewable-energy economy? Multi-actor perspective on social innovation," *Sustainability*, vol. 10, no. 2, p. e448, 2018.
- [13] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "Hybrid clustering analysis using improved krill herd algorithm," *Appl. Intell.*, vol. 48, no. 11, pp. 4047–4071, 2018.
- [14] L. M. Abualigah and A. T. Khader, "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering," *J. Supercomput.*, vol. 73, no. 11, pp. 4773–4795, 2017.
- [15] L. M. Abualigah and E. S. Hanandeh, "Applying genetic algorithms to information retrieval using vector space model," *Int. J. Comput. Sci. Eng. Appl.*, vol. 5, no. 1, pp. 19–28, 2015.

- [16] S. Sauerland, T. Jaschinski, and E. A. M. Neugebauer, "Laparoscopic versus open surgery for suspected appendicitis," *Cochrane Database Syst. Rev.*, vol. 10, 2010, Paper CD001546.
- [17] T. Zhang, L.-H. Shen, L. H. Hu, and B. He, "Statins for the prevention of contrast-induced nephropathy: A systematic review and meta-analysis," *Amer. J. Nephrol.*, vol. 33, no. 4, pp. 344–351, 2011.
- [18] Y. Tong, J. Liao, Y. Zhang, J. Zhou, H. Y. Zhang, and M. Mao, "LOC387715/HTRA1 gene polymorphisms and susceptibility to agerelated macular degeneration: A HuGE review and meta-analysis," *Mol. Vis.*, vol. 16, no. 213, pp. 1958–1981, 2010.
- [19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.
- [20] D. M. Blei, "Probabilistic topic models," Commun. ACM, vol. 55, no. 4, pp. 77–84, Apr. 2012.
- [21] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004.
- [22] LDA 1.0.5: Topic Modeling With Latent Dirichlet Allocation. Accessed: Aug. 21, 2018. [Online]. Available: https://pypi.python.org/ pypi/lda



SUSAN (SIXUE) JIA received the Ph.D. degree in management from Tsinghua University, Beijing, China, in 2015. She is currently an Assistant Professor with the School of Business and Management, Shanghai International Studies University, Shanghai, China.

BANGGANG WU received the Ph.D. degree in management from Tsinghua University, Beijing, China, in 2018. He is currently an Assistant Professor with the Economics and Management School, Wuhan University, Wuhan, China.

...