Cumhuriyet Science Journal

e-ISSN: 2587-246X
ISSN: 2587-2680

Cumhuriyet Sci. J., 42(4) (2021) 934-941
http://dx.doi.org/10.17776/csj.974701

An econometric model for popularrity on media

***Orçun AYDIN*[1], *Erol TERZİ*[2*]**

[1] *Turkish Statistical Institute (TurkStat), Ankara/TURKEY*

[2] *Ondokuz Mayıs University, Faculty of Science, Department of Statistics, Samsun/TURKEY*

Abstract

This paper aims to determine and estimate an econometric model which can be used to forecast media popularity of a governmental organization. Number of media sources monitored was used as regressors while taking types of these sources into account. Some linear models were estimated besides some non-linear models. According to the results, number of national, local, regional newspapers and number of television channels monitored were not found important to estimate number of news caught through media monitoring. On the other hand, number of internet media sources was found important to estimate the dependent variable. Additionally, number of news caught on select subjects in previous year was also found important. In the end an autoregressive panel data model with some additional regressors such as number of monitored sources was suggested to forecast popularity of organization. Any data only accessible to TurkStat members was never used in this paper. TurkStat is not responsible for any inference made in this study.

Article info

History:
Received: 26.07.2021
Accepted: 21.11.2021

Keywords:
Panel data,
Media monitoring,
Regression,
TurkStat

1. TurkStat and Media Monitoring

Turkish Statistical Institution disseminates more than one hundred press releases each year. On most of the workdays announced press releases attracts wide attention from governmental and academic organizations also from public. In this paper dataset on news between the years 2012 and 2020 was studied. Any data only accessible to TurkStat members was never used in the study. TurkStat is not responsible for any inference made in this study.

Media monitoring is an interesting subject and risks can be minimized with monitoring [1]. To extract the useful information from this mass data, Press and Public Relations Department at TurkStat works on daily cycle agenda and gets the news form selected sources on every morning. Then a team leader attains the news to staff for tagging. The workload of each staff is determined by the consultant by taking their other daily tasks. Then each staff reads the news attained themselves and process them by entering data interested (tagging and classifying). Classification of Negative Information on Socially Significant Topics in Mass Media was also studied before [2]. But TurkStat needs to handle not only comments but also subjects to inference and decide on a detailed communication strategy. Government communication is framed within political communication itself and refers to the exercise that determines the management agenda of institutions, attitudes, and processes [3].

In this paper topmost important subjects were determined and then an econometric forecasting strategy was tried. At first, all the 369,938 news in 2020 are searched and then classified according to the list used by Press and Public Relations Department at TurkStat. However, all the dataset needed was formed and built up from scratch because it is not wanted to use any data only accessible to TurkStat members.

*Corresponding author. *e-mail address: eroltrz@omu.edu.tr*

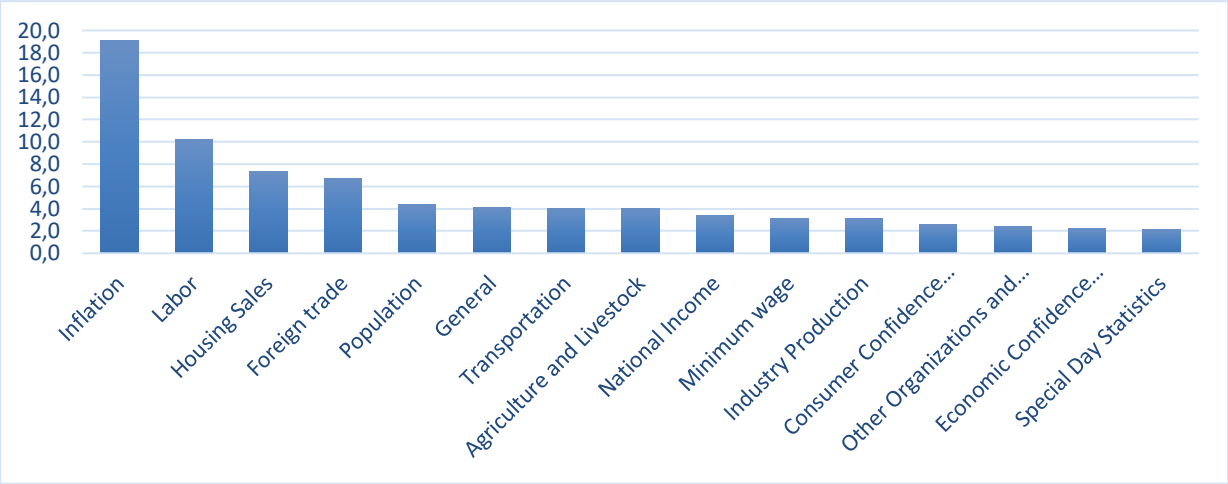

Figure 1. Top fifteen subjects on media in 2020

Then the order of the table was changed in decreasing order according to the amount of news for each subject. It has been seen that the top 12 subjects accounts for more than 70% off all the news in 2020. So, these subjects were chosen as cross-sections to eliminate heterogeneity bias [4].

In the above figure, minimum wage, foreign trade, inflation, labor force, house sales, general (corporation), gross domestic product (GDP), population, industrial production, agriculture, consumer confidence index (CCI), and transportation are the top twelve subjects in 2020.

2. Data and Method

Any data only accessible to TurkStat members was never used in the study. Only the open access data and meta-data were used to analyze the popularity of TurkStat on media. A third-party media monitoring tool was used as a service to search and access texts and media. By using this service news with given uniform resource locaters (URL) were searched with selected keywords above-mentioned. Then, they were classified using the list of subjects. In this process it is seen that some news belongs to more than one class. In such situations, their class is determined by looking the density and meaning, subjectively.

Table 1. Number of news according to selected subjects by year

	Years									
Subjects	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
Minimum Legal Wage	270	222	856	2,105	3,924	3,160	6,086	7,235	11,589	35,447
Foreign Trade	2,455	3,024	11,809	14,235	20,060	23,749	26,300	23,510	24,968	150,110
Inflation	2,696	3,116	16,228	23,116	29,711	36,990	48,547	57,937	70,637	288,978
Labour Force	2,249	3,312	11,620	14,270	16,988	24,543	22,491	27,799	37,846	161,118
House Sales	113	699	7,177	11,513	16,565	19,210	21,454	23,762	27,164	127,657
General	404	947	5,495	5,609	12,371	7,613	7,611	5,262	15,179	60,491
GDP	747	1,202	3,491	6,261	11,145	15,469	10,726	9,647	12,532	71,220
Population	1,152	2,023	6,339	5,998	12,136	12,635	14,721	13,339	16,130	84,473
Industry Production	766	1,233	3,782	5,065	1,770	8,326	8,509	8,764	11,477	49,692
Agriculture	1,386	2,136	9,266	12,454	14,834	18,386	15,075	15,223	14,788	103,548
CCI	251	451	1,971	4,000	4,265	4,975	4,887	5,634	9,611	36,045
Transportation	841	1,687	5,484	7,924	12,863	13,163	16,263	12,509	14,811	85,545
Total	13,330	20,052	83,518	112,550	156,632	188,219	202,670	210,621	266,732	1,254,324

In the above table, amount of news by selected subjects and by years can be seen for the time period between years 2012 and 2020. Top twelve subjects are the rows of table. This selection makes analyzing easier as the reduced table accounts for more than 72% of all data. For the year 2020, there are 266,732 news caught on these subjects. Without selection, there are 369,938 news caught on all the subjects including the unclassified ones.

In the below table, number of media sources for media monitoring can be seen. About 1,165 sources were monitored for 2012 while it has increased up to 4,172 in 2020. In this paper, the number of media sources used for monitoring was used as dependent variable and effect of this variable on the popularity of TurkStat on media was investigated. Amount of news was used as a proxy for this popularity.

Table 2. Number of media sources according to media types

Years	National Newspapers	Local N.	Regional N	Magazines	Tv Channels	Internet Media	Total
2012	45	250		800	70		1.165
2013	48	668		1.059	160		1.935
2014	46	503	210	1.489	81	394	2.723
2015	52	378	228	1.137	98	1.098	2.991
2016	51	536	212	1.189	74	1.501	3.563
2017	51	536	212	1.189	75	1.605	3.668
2018	51	536	212	1.189	75	1.605	3.668
2019	57	379	212	1.189	107	2.228	4.172
2020	57	379	212	1.189	107	2.228	4.172

According to the table above, internet as a media source type accounts for most of the sources. Cost of internet press is so less compared to all other traditional ways. Magazines takes the second place however amount of news on this type of source is not so high as expected. Because the time period from 2012 to 2020 is not enough for a time series analysis, a panel data approach is more convenient to apply. So, selected subjects are used as cross sections.

Many researchers illustrated panel data analyses deeply [4-10]. According to them, a general panel data regression model can be represented as below.

$$Y_{it} = \mu + \mu_i + \beta_1 X_{1,it} + \beta_2 X_{2,it} + \cdots + \varepsilon_{it} \quad (1)$$

Here, the variable Y_{it} is the dependent variable and regressed on $X_{1,it}$, $X_{2,it}$ and so on. The heterogeneity between cross sections is represented by the cross-sectional intercept term μ_i. Here, μ is the general intercept for the panel data regression model. ε_{it} represents errors and β_i's are the slope parameters to be estimated.

Covariance analysis for this panel data model was explained to identify the source of sample variance [4]. As mentioned, this method allows the true relation for each individual to depend on the class to which the individual belongs. As explained in panel data regression analysis effects are assumed to be fixed or random [4]. Fixed effects models can be estimated with dummy variable approach using ordinary least squares method and this estimator is called as covariance estimator.

In this manner, it also includes some advantages of usual analysis of variance [4]. As explained, panel data gives the researcher large number of data points. Suppose researcher has ten individuals and gathered data about them for ten years' time period. So, ten years is not enough for a time series analysis neither for a cross sectional analysis. But in the end with this panel data approach, researchers have one hundred data to use. Additionally, panel data gives the opportunity to investigate some questions impossible to get answers with time-series or cross-sectional data sets. Making dynamic inference is not possible with a cross-sectional data set. Multiple information for an individual decreases the negative effect of measurement errors. Panel data gives the model a chance to learn individual's behavior from other ones' behaviors.

On the other hand, panel data sets have some disadvantages also [4]. Firstly, analyze is more complicated compared to simple cross-sectional or time-series analysis. Unobservable regressors correlated with other regressors in the model can lead to biased estimators as explained.

3. Results and Discussion

At first, data was visualized, descriptive statistics was found for the 2020 data. By this way, the range of the data to be used was discussed to make things simple while keeping study representative. At the figure below, all the news by months can be seen for the year 2020.

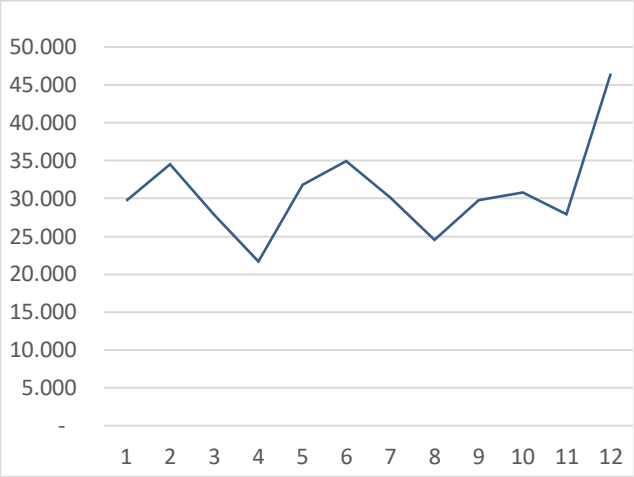

Figure 2. Amount of news by months in 2020

The number of news is not sable among months and follows a fluctuating pattern. Additionally, second, sixth and twelfth months experienced the top three amounts of news.

At the figure below, amount of news by months and by media type can be seen. In total, 369,938 news were caught to our monitoring system. Data for internet media type as 2254 thousand news does worth mentioning for the internet media type. This is the case for all the years from 2012 to 2020. This domination comes from the fact that internet press publishing is much easier compared to traditional arenas as for low costs and also for the effort needed to publish. In the end, it is so wide to publish news with embedding in web pages.

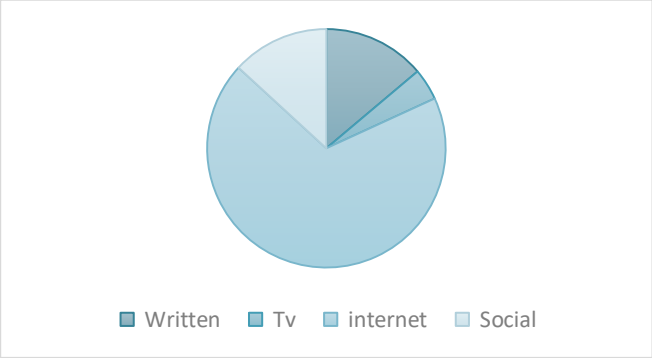

Figure 3. Number of news by nedia types in 2020

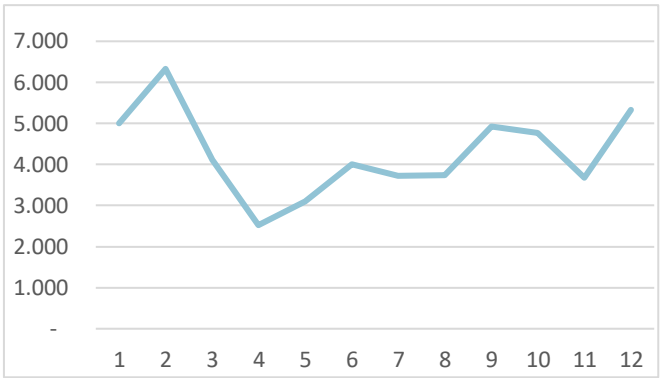

Figure 4. Number of news by months on written sources in 2020

In the figure above, written news caught by our monitoring system shows a different fluctuating pattern from the general line illustrated before. Observation performs a peak in second month with more than six thousand news. However, this trend falls below three thousand news on fourth month. An upward trend can be seen towards to the end of the year 2020. According to these points it can be said that every type of media source performs a different pattern peculiar to themselves.

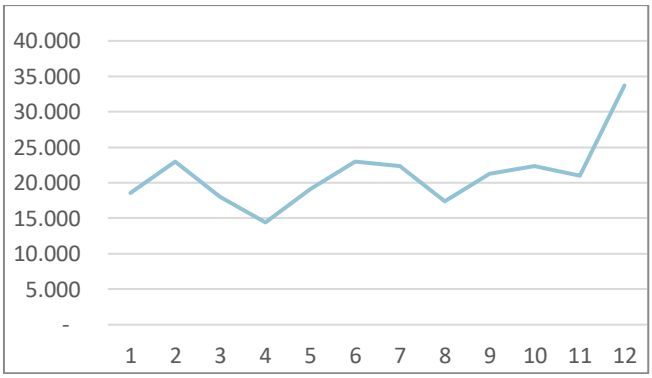

Figure 5. Number of news by months on internet sources in 2020

Internet media type is less volatile between 15,000 and 25,000 news. However, amount of news on internet media sources performs more than ten thousand news on each month. A similar peak and deep can be seen on second and fourth months as other media types.

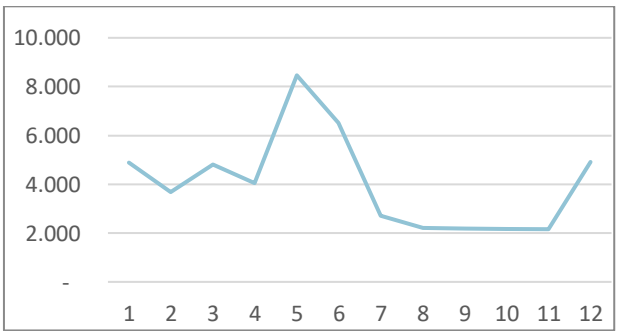

Figure 6. Number of news by months on social media sources in 2020

Social media sources perform a different pattern compared to other three types. Number of news fluctuates between two thousand and nine thousand. On the fifth month a strong peak can be seen with more than five thousand news. On the seventh month line falls below three thousand news and goes through an approximately stable level until eleventh month. Towards the end of the year a strong trend shows up again.

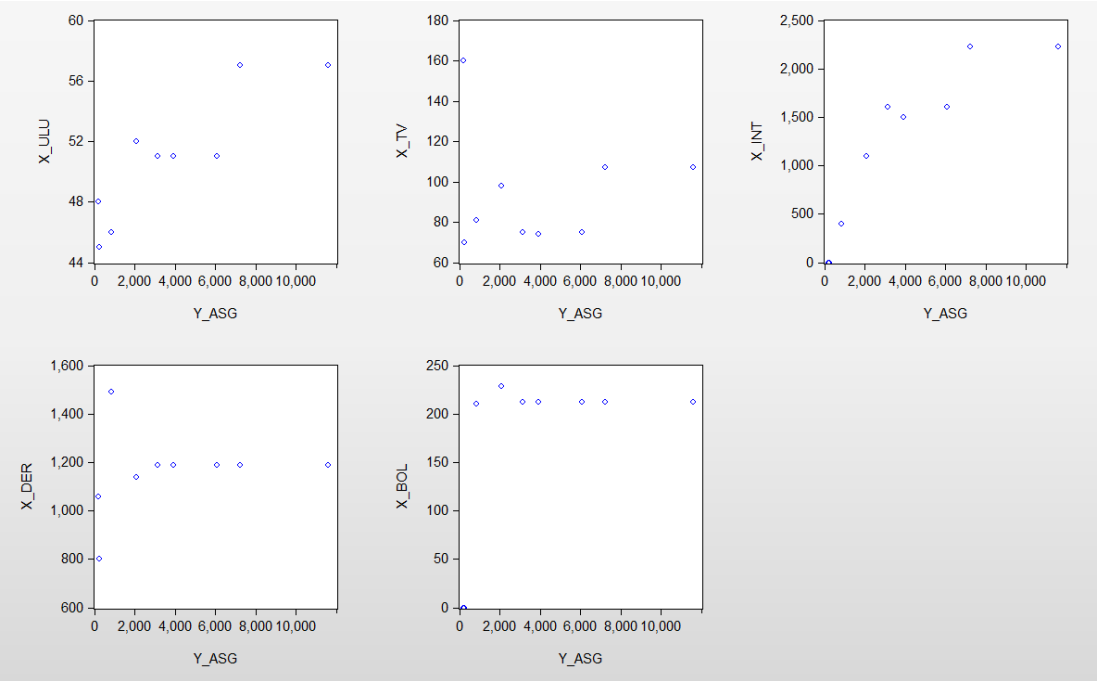

Figure 7. Scatterplot between minimum wage and types of media sources

Scatter plot above shows a nearly linear correlation between national newspapers. A similar linear structure can be seen between minimum wage and Tv's but with a less correlation noticed visually. Correlation between minimum wage and internet media type is like linear but in some scatterplots square like correlation can also be seen. However, magazine and regional media sources visually has not a strong correlation between minimum wage news.

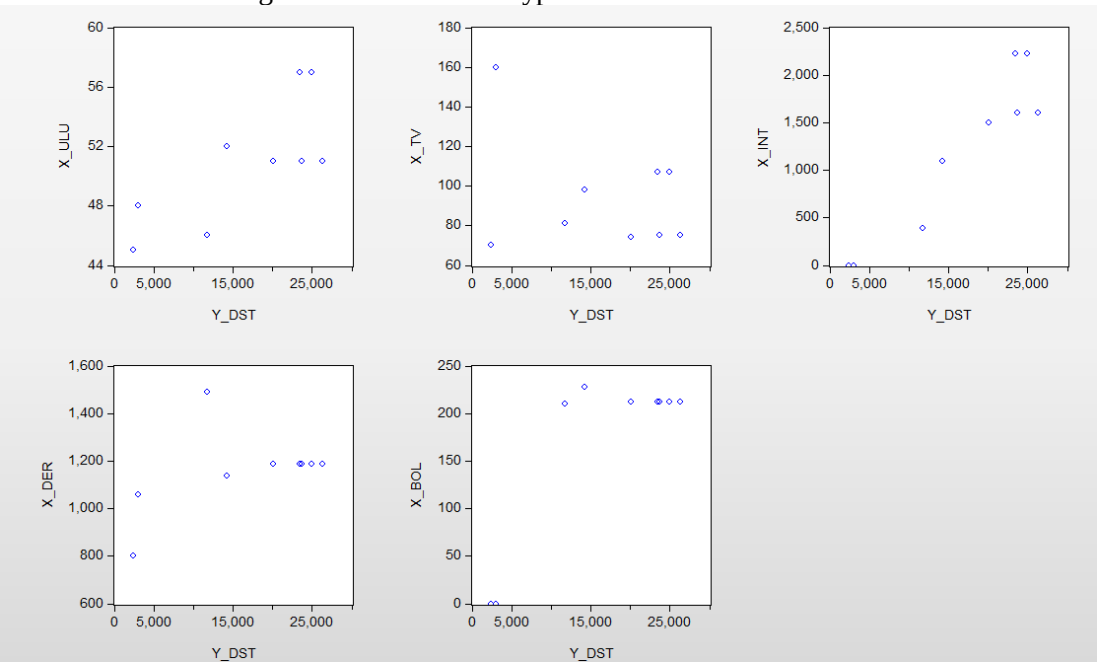

Figure 8. Scatterplot between foreign trade and types of media sources

Scatter plot above shows a nearly linear and positive correlation between national newspapers and foreign trade. A similar linear structure can be seen between foreign trade and internet media sources. Correlation between foreign trade and magazine media type is a square root like correlation. No correlation can be mentioned visually between foreign trade and regional media sources.

As mentioned below while checking scatterplots observed relations can be modelled as linear or square root like relations. In this sense, first model was set as following.

$$Y_{it} = \mu + \beta_1 X\_int_{it} + \beta_2 X\_Bol_{it} + \beta_3 X\_Der_{it} + \beta_4 X\_tv_{it} + \beta_5 X\_ulu_{it} + \varepsilon_{it} \quad (2)$$

Pooled estimation with Model (2) is found to be significant (P=0.0000) but most of the coefficients found to be insignificant. Only internet media type is significant at 0.10 level. Determination coefficient is so low (%33) and Durbin-Watson (0.159) is not acceptable. That is why Model (3) was estimated below.

$$Y_{it} = \mu + \mu_i + \beta_1 X\_int_{it} + \beta_2 X\_int_{it}^{1/2} + \varepsilon_{it} \quad (3)$$

Model (3) is found to be significant (P=0.0000). Determination coefficient is 0.75 and shows a better fit with this fixed effect approach compared to pooled regression in Model (2). β_1 is estimated as 6,89 (P=0.0345) and β_2 is estimated as 72.15 (P=0.6313). Internet media type was found significant in Model (2) as placed linearly but is found insignificant here in square root form. However, Durbin-Watson is not acceptable (0.0476) and should be seen as signs of autocorrelation.

$$Y_{it} = \mu + \mu_i + \beta_1 X\_int_{it} + \varepsilon_{it} \quad (4)$$

Model (4) was found to be significant (P=0.0000). Determination coefficient is again 0.75 and shows a good fit with this fixed effect approach. β_1 was estimated as 8,389 (P=0.0000). Internet media type was found insignificant in Model (3), so left aside in this Model (4). However, Durbin-Watson is not acceptable (0.0497) and signs some autocorrelation again as Model (3).

$$Y_{it} = \mu + \mu_i + \beta_1 X\_int_{it} + \beta_{2i} Y_{it-1} + \varepsilon_{it} \quad (5)$$

Model (5) was found to be significant (P=0.0000). Determination coefficient is better (0.96) with this cross-sectional fixed-effect dummy variable approach. Adjusted-R^2 is found as 0.948 and shows a good determination. Both intercept and slope coefficients are heterogenous in this model and this makes sense. Because each subject has its own trend in media popularity. Estimated values for β_1 and β_2 coefficients are given below. Durbin-Watson is acceptable (2.55) and signs that autocorrelation is removed with autoregressive cross-relational term in this Model (5).

Table 3. Coefficient estimates for Model (5)

Coefficient	Estimate	St. error	t-stat	Prob.
μ	2729.588	609.3685	4.479371	0.0000
β_1	3.673852	0.915684	4.012139	0.0002
β_{2i} (minimum wage)	0.317134	0.471417	0.672725	0.5035
β_{2i} (foreign trade)	0.506766	0.135512	3.739635	0.0004
β_{2i} (inflaton)	0.962326	0.062963	15.28409	0.0000
β_{2i} (labor)	0.750553	0.133379	5.627236	0.0000
β_{2i} (house sales)	0.642272	0.138175	4.648246	0.0000
β_{2i} (corporational)	-0.178859	0.306650	-0.583266	0.5617
β_{2i} (GDP)	0.286263	0.228706	1.251661	0.2152
β_{2i} (industrial production)	0.023742	0.358184	0.066285	0.9474
β_{2i} (agricukture)	0.269409	0.197289	1.365552	0.1768
β_{2i} (CCI)	-0.222603	0.598527	-0.371919	0.7112
β_{2i} (transportation)	0.295716	0.219205	1.349038	0.1820

Most of the coefficient estimates are significant for Model (5). With the assumption of this relationship structure keeps existing for the future period, this model can be used for forecasting. Cross-sectional coefficient estimates are given below.

Table 4. Cross-sectional coefficient estimates for Model(5)

Coefficient	Value
μ_1 (minimum wage)	-4173.183
μ_1 (foreign trade)	2905.119
μ_1 (inflaton)	1896.306
μ_1 (labor)	668.8228
μ_1 (house sales)	250.4876
μ_1 (corporational)	899.3940
μ_1 (GDP)	-915.4359
μ_1 (industrial production)	-1622.200
μ_1 (agricukture)	2156.622
μ_1 (CCI)	-2414.749
μ_1 (transportation)	348.8170

4. Conclusion

In this paper, news published on media about statistics disseminated by Turkish Statistical Institute (TurkStat) were analyzed. More than one million records over years from 2012 to 2020 were studied. All the records were studied by using URL as primary key on database. Texts for all the news were read and then news were classified according to related subjects. This was accomplished by recording interested variables for each while reading news. By this way after reading and before classifying more than six million data gathered. That is why open-source R software was used for this huge dataset. Reading media articles to discuss the popularity is a conventional way but still works [11].

Even countries and international organizations use content analysis in media articles to get a acknowledge on image and popularity [12]. In this study, some keywords adaptive with TurkStat's press release calendar were used for content analysis. By these keywords, news was matched with press releases or other subjects. After this preparation data was aggregated yearly and then modelled by using panel data regression methods. In this manner the effect of number of media sources monitored on popularity of organization was investigated. In the end, an econometric model was suggested for forecasting purposes.

Number of news published on media was used as a proxy for popularity on media. Only number of internet type media sources was found to be important on estimating number of news published. An autoregressive panel data model was used. A percentage increment in number of monitored media sources was found to increase media popularity by 3.67 percent. Previous amount of news on a certain subject was also found to be important on estimating amount of news in current year on selected subject. A percentage increment in number of news published on a certain subject was found heterogenous between subjects.

As in some studies, future popularity on the internet can be predicted by current available data [13]. Differently by the econometric model suggested in this study, number of news could be forecast because future number of sources to be monitored is an accessible information before that year starts. Previous year's number of news is also a variable and value is known beforehand. There is no other study like this for a governmental organization. That is why national statistical offices (NSO) can use this approach to forecast popularity of NSO and set a communication strategy relying on these statistics. Because there is not enough data for a time series approach, panel data method used in this paper can be used to succeed in estimating and forecasting.

Criticism on his study can be made on, no classifying is adopted for comments in texts. In case of a positive or a negative comment in meaning, "whether this effect the popularity or not?" is a nonrespondent question here. Only the absolute numbers by selected subjects are handled in this study. For future researchers inspired by this study it can be suggested to focus more on attitudes of the press staff. Because editors add interpretations while designing and editing texts of news. These interpretations are thought to have important effect on spreading acceleration of these news. This may indirectly determine the popularity of an organization.

Conflicts of interest

The author state that did not have conflict of interests

References

[1] Craufurd Smith R., Monitoring media pluralism in the digital era: application of the Media Pluralism Monitor 2020 in the European Union, Albania and Turkey in the years 2018-2019. Country report: United Kingdom , (2020).

[2] Mukhamediev R. I., Yakunin K., Mussabayev R., Buldybayev T., Kuchin Y., Murzakhmetov S., Yelis M. (2020). Classification of Negative Information on Socially Significant Topics in Mass Media, *Symmetry*, 12(12) (2020) 1945.

[3] Sánchez-Núñez P., Yanez E. R., Cabrera F. E., Peláez-Repiso A., Government Communication Management in Digital Ecosystems: A Real Case of Country Brand Analysis, In 2020 *Seventh International Conference on eDemocracy & eGovernment (ICEDEG)* (pp. 264-268) (2020, April) IEEE.

[4] Hsiao C., Panel data analysis, (2003).

[5] Arellano M., Panel data econometrics. Oxford university press, (2003).

[6] Baltagi B., Econometric analysis of panel data. John Wiley & Sons, (2008).

[7] Frees E. W., Longitudinal and panel data: analysis and applications in the social sciences. Cambridge University Press, (2004).

[8] Hsiao C., Why panel data?. *The Singapore Economic Review*, 50(02) (2005) 143-154.

[9] Hsiao C. Analysis of panel data (No. 54). Cambridge university press, (2014)..

[10] Wooldridge J. M., Econometric analysis of cross section and panel data. MIT press, (2010).

[11] Regusci E., A Content Analysis of News Coverage about Plant-Based Milk, master thesis, Faculty of Texas Tech University, (2020).

[12] Chouliaraki L., Georgiou M., Zaborowski R., Oomen W. A., The European 'migration crisis' and the media: a cross-European press content analysis, (2017).

[13] Trattner C., Moesslang D., Elsweiler D.,. On the predictability of the popularity of online recipes. *EPJ Data Science*, 7(1) (2018) 1-39.