 Posted prediction in social media base on Markov chain model: twitter dataset with covid-19 trends

W Suryaningrat\textsuperscript{1}, D Munandar\textsuperscript{1,2}, A Maryati\textsuperscript{1,3}, A S Abdullah\textsuperscript{1,4}, and B N Ruchjana\textsuperscript{1,*}

\textsuperscript{1}Department of Mathematics, Universitas Padjadjaran, Jl. Raya Bandung-Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia
\textsuperscript{2}Research Center for Informatics, Indonesian Institute of Sciences, Jl. Cisitu No.21/154D Komplek LIPI Gedung 20 Lantai 3, Bandung 40135, Indonesia
\textsuperscript{3}Senior High School 1, Cileunyi, Jl. Pendidikan No.6, Bandung 40625, Indonesia
\textsuperscript{4}Department of Computer Science, Universitas Padjadjaran, Jl. Raya Bandung-Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia

*budi.nurani@unpad.ac.id

Abstract. The influence of social media is very attractive in disseminating information; even social media analysis is one of the focuses in the field of research in terms of data mining. In its development not only the field of social science that exists but many studies of social media that can be solved stochastically to calculate the trend of the emergence of a discussion on social media. In this paper, we investigated calculations and predictions using Markov Chains on the emergence of discussions on Twitter media related to coronavirus disease tweets or better known as covid-19. The tweet data obtained is a random sample of the tweet posts that are crawled at the specified time. The tweet data is crawled at three different observations each day for thirteen days continuously. The results of data crawling are calculated to determine the transition from one observation to the next observation. The stages of the process are; crawling tweet data with keywords coronavirus and covid-19; data cleaning process; data processing; Markov Chain modeling; n-step distribution and long-term prediction; interpretation of results. The computational results used are opportunity distribution conditions for the number of tweets. As a transition between two states, namely low (0) and high (1) relative to mean or median. The results of the opportunity distribution obtained in the next 145-time steps (0.28767, 0.71233) and (0.47368, 0.52632) in the probability distribution of the number of tweets are respectively the mean and median values. The results of the modeling show that the conversation on Twitter for 145-time steps in the next prediction is estimated to remain high along with the outbreak of coronavirus or covid-19 before this epidemic subsides.

1. Introduction
Various social media platforms have been widely used as a means of exchanging information both by individuals and organizations throughout the world. The number of social media users has continued to increase rapidly, especially in the past decade. Facebook, Twitter, Youtube, LinkedIn and Pinterest saw significant increases over the past year. By number of user, Facebook is the most popular social media with 1.2 billion active users. While Twitter took second place with around 300 million users. [1].

Each social media platform has certain characteristics and can be classified based on its function [2]. Twitter is an online social networking and microblogging service that provides facilities for users to send and read text-based messages. Twitter is experiencing rapid growth and is quickly gaining popularity all over the world. Twitter service is used by some users to support various aspects, for example as a means of protest, political campaigns, knowledge spreading, and as playing an important role in a social movement [3–5].
Twitter makes many parameters of social learning that can be learned. Twitter content, such as tweets and retrieval aggregated by the system following the frequency of specific words created can provide an overview of the phenomena of everyday life that occur [6; 7]. The social phenomena in question can be in the form of opinions on the Indonesian presidential election [8], the dynamics of consumer online-shop usage [9–11] changes in the level of popularity of a brand [12], and also talks about health topics especially since WHO declared the coronavirus novel (covid-19) as a global pandemic [13].

Twitter as a social media is very widely used as an opinion, discussion, and research material. Several researchers have previously modeled the exchange of information on twitter. The Non-Homogeneous Poisson Process (NHPP) hierarchy model is used to predict and calculate the final retweet based on the original tweet using the hybrid method as conducted by Mathew et al [14] but without discretizing and carrying out the hierarchy of the retweet [15]. But it is different from the use of Twitter in the field of tourism, by utilizing the Geo attribute of Twitter data to detect tourist movement patterns utilizing DBSCAN-based and Markov clustering [16]. On the other hand by utilizing Twitter or hashtag trending topics that often appear in dynamic programming for modeling infection rates and recovery in infectious diseases using the Bayesian Markov Chain Monte Carlo (MCMC) method [17].

2. Method
2.1. Dataset Description
The dataset used is sourced from Twitter social media by crawling based on keywords. Some keywords used are Coronavirus, covid-19, and Covid19. By using the Application Programming Interface (API) which is open-source, it is a library of Twitter. Crawling results for 13 days with these 3 keywords at 3 different times, and 39 indexes of observation. The three combined tweet keywoards statistically can be seen in Table 1. We get the number of tweets for 13 days with keywords that define 921,462 tweets, with a mean of 70,882 tweets per day, and a standard deviation of 11,528.

| Days to | 1  | 2  | 3  | 4  | 5  | ... | 13 |
|--------|----|----|----|----|----|-----|----|
| Number of tweet keywords | 73,745 | 60,993 | 84,627 | 59,265 | 78,089 | ... | 80,830 |
| Mean of keywords | 24,582 | 20,331 | 28,209 | 19,755 | 26,030 | ... | 26,943 |
| Standard deviation | 13,527 | 12,387 | 6,639 | 11,807 | 10,055 | ... | 457 |

2.2. Markov Chain Model
A Markov chain is a stochastic model characterize a sequence of possible events in which the probability of each state or event depends only on the previous state. Markov chain in general can be specified based on state-space and time-index. Markov chains were named after their inventor, A. A. Markov, a Russian Mathematician who worked in the early 1900’s [20]. Simply put, a Markov chain uses a matrix and a vector to model and predict the behavior of a system that moves from one state to another state in a way that depends only on the current state.

The term “Markov chain” is used for a process with a discrete set of parameter or times, also known as, a discrete-time Markov chain (DTMC), while some authors use the term “Markov process” to refer to a continuous-time Markov chain (CTMC). However, many applications of Markov chains used discrete-time with finite or countably infinite state-spaces, that have further statistical analysis.
2.2.1. Discrete-Time Markov Chain

**Definition 1.** Let \( \{X(n), n = 1, 2, \ldots \} \) stochastic process with discrete time-index and state-space \( i = 0, 1, 2, \ldots \) satisfy

\[
P\left\{X(n + 1) = j \mid X(0) = i_0, X(1) = i_1, \ldots, X(n - 1) = i_{n-1}, X(n) = i\right\} = PX(n + 1) = j \mid X(n) = i = p_{ij}
\]

\forall i_0, \ldots, i_{n-1}, i, j, \text{ and } n, \text{ then this stochastic process named Discrete-Time Markov Chain and } p_{ij} \text{ named transition probability.}

The probability of transition from state \( i \) to state \( j \) \((p_{ij})\) of Equation 1 depends only on the present time. If the transition probability is independent of time \( n \), then it is called the stationary transition probability, and the Markov chain is called Homogeneous Markov Chain.

**Definition 2.** One step transition probability matrix of \( \{X(n), n = 0, 1, 2, \ldots \} \) defined as

\[
P = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{01} & p_{02} & \cdots \\ p_{20} & p_{01} & p_{02} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}
\]

(2)

where \( p_{ij} \geq 0 \) and \( \sum_{j=0}^{\infty} p_{ij} = 1 \) \((i, j = 0, 1, 2, \ldots)\)

2.2.2. Chapman-Kolmogorov

The probability that the process in state \( i \) will be in state \( j \) after \( n \) transitions stated by \( p_{ij}^n \). Where it can be calculated by adding up all the probability of state \( k \), at time \( r \) \((0 \leq r \leq n)\) and the transition of state \( j \) from state \( k \) in the remaining time \( n - r \) as follows

\[
p_{ij}^n = P\left\{X(n) = j \mid X(0) = i\right\} = \sum_{k=0}^{\infty} P\left\{X(n) = j \mid X(n-r) = k, X(0) = i\right\}P\left\{X(n-r) = k \mid X(0) = i\right\} = \sum_{r=1}^{\infty} p_{ik}^r p_{kj}^{n-r}.
\]

(3)

The Equation 3 is called the Chapman-Kolmogorov equation. In matrix form it can be written by

\[
p^{(n)} = P^{(r)} . P^{(n-r)}.
\]

(4)

2.2.3. Probability Distribution of opportunities

The combined probability can be calculated from initial probability (initial distribution \( \pi_0 \)) and transition probability \((P \text{ matrix})\). Let, \( p_{ij}(n) = P\{X(n) = j\}, \text{ and } \pi_{(n)} = [\pi_0(n), \pi_1(n), \ldots] \) as n-step distribution then it holds

\[
\sum_{j=0}^{\infty} \pi_j(n) = 1 \text{ and } \pi(n) = \pi(0) P^n.
\]

(5)
3. Result and Discussion

3.1. Data Preparation

3.1.1. Review of the Twitter dataset

The dataset crawled through Twitter social media is crawled at 3 different times with 20 minutes of crawling duration. We assume that the time when Twitter users tweet the most is at three specific time per day. The chosen time is; the first at 11.00 a.m.-11.20 a.m., the second at 04.00 p.m.-04.20 p.m., the third at 09:00 p.m.-09.20 p.m. The retrieval time is conditioned by using a scheduling application program that is set up on a computer that acts as a server. Crawling data runs for 24 hours non-stop by following the time settings above. Choosing time is not without reason. After observing, these times crawling show increase the volume of tweets compared to other times. By selecting 3 different times, this is considered sufficient to see the conditions and calculate the probability of these 3 keywords.

3.1.2. Cleaning the dataset

This process is carried out after raw data stored from 10-22 March 2020 is obtained and stored in a repository contained of a text file. Every data is stored in a file, so every day there are 3 data tweet files. The attributes taken are id, create_at, text. Especially for text attribute to be cleaned to facilitate the process of counting keywords. The cleaning begins by removing the ASCII character that represents unwanted symbols for each line of tweets, then all words are converted to low case for uniformity of words to make it easier to calculate, then the tweet attribute that is the text is placed in one column as the initialization of the calculation.

3.1.3. Counting keywords every time crawling

The calculation of each keyword is obtained on a file containing tweets that have been cleaned. It will count how many coronavirus, covid-19, covid19 keywords appear that were discussed by netizens for a specified duration of time. Some examples of tweets obtained are shown in the Table 2.

| Time Posted | Tweets |
|-------------|--------|
| 2020-03-22 21:05:02 | Lockdowns not enough to defeat #coronavirus |
| 2020-03-22 21:05:39 | The only person that could end Coronavirus is her . . . |
| 2020-03-22 21:05:55 | Symptoms of COVID-19 may appear anywhere from 2-14 days after exposure |
| 2020-03-22 21:06:04 | NEW: India reports 3 new coronavirus deaths |
| 2020-03-22 21:06:38 | Update Jumlah Kasus Covid-19 di Indonesia dan Jawa Timur per hari ini . . . |
| 2020-03-22 21:07:17 | Thailand plans $3 billion liquidity support fund amid coronavirus outbreak |
| 2020-03-22 21:08:01 | Many people are affected by the CoronaVirus. Avoid public places, wash your hands regularly. . . |
| 2020-03-22 21:08:30 | . . . this must be the headlines. We can defeat #coronavirus |
| 2020-03-22 21:09:39 | Receta la OCDE rescate tipo Plan Marshall por pandemia de coronavirus |
| 2020-03-22 21:10:18 | *#PrayIndonesia* Korban Wabah Covid19 Setiap Hari Terus Bertambah |
| 2020-03-22 21:10:39 | . . . Update situasi terkini perkembangan #COVID19 di Indonesia (22/3) #tenangdanwaspada #LawanCovid19 |
| . . . | . . . |
By using the substring and sum-product functions we get the number of keywords that are defined, and so on until we get the number per time up to the day as in Table 1. The transition from these numbers will be used as a state for the Markov Chain process which will be explained in the next section.

3.2. Data Processing

In data processing, the numbers of the dataset that has been collected through crawling will be calculated on the trend of tweets with the keywords specified above. $N_T$ is a time series of tweets. While the arithmetic mean (average) value of the number of tweets is $N_{\text{bar}}$. Then the median value of the number of tweets is $N_{\text{med}}$. Mean and Median Numbers of coronavirus, covid-19 Tweets keywords are $N_{\text{bar}}=23,627$, $N_{\text{med}}=27,282$ respectively.

The values between the average number and median of tweets have different values, but showing the value of $N_{\text{med}}>N_{\text{bar}}$ indicates a positive frequency distribution curve. Figure 1 observation for the tendency of high number of tweets is at the observation hours 9.00 - 9.20 pm Western Indonesia Time, the number of tweets obtained at that time exceeds above mean line and median line. These observations can be used as a reference that the tendency of netizen’s at that time to discuss covid-19 on social media through twitter. Outlier is shown in Figure 1. is a very different data point from the other data [18]. The existence of outliers can affect the performance of the models created [19]. The graph detected that there are outlier data at the time index $t = 18$ and $t = 25$ which are relatively distorted from the mean of all data. This condition is then solved by removing outlier data and replacing them with a median of all data or determine the state space based on the median. As for the outlier data that will be solved are $N_{18}=2,129$, $N_{25}=1,131$.

3.3. Classification and State Transition

The changing dynamics of many covid-19 tweets can be viewed as the transition between two states high (1) and low (0) relative to the number of $N_{\text{bar}}$ and $N_{\text{med}}$ previously defined. The corresponding discrete state space can be classified as follows:

$$X_t = \begin{cases} 
\text{State 0} : N_t < N_{\text{bar}} \\
\text{State 1} : N_t \geq N_{\text{bar}} 
\end{cases}$$

$$Y_t = \begin{cases} 
\text{State 0} : N_t < N_{\text{med}} \\
\text{State 1} : N_t \geq N_{\text{med}} 
\end{cases}$$

$X_t$: State space for the number of tweets with average threshold value $N_{\text{bar}}$

$Y_t$: State space for the number of tweets with the median threshold value $N_{\text{med}}$

The parameter space in this case is 39 time index $t \in T$, where $T = \{1, 2, 3, 4, ..., 39\}$. 

![Figure 1. Number of covid-19 tweets to mean and median.](image-url)
Table 3. Data classification number of covid-19 tweets in state space

| State space | Elements |
|-------------|----------|
| $X_t$       | 011001011001110110110011111111111111111 |
| $Y_t$       | 01100101100111011011001111111111111111111 |

Table 4. The number of transition conditions from state $i$ to state $j$ for state space $X_t$

| State $i$ | State $j$ | 0 | 1 | Sum |
|-----------|-----------|---|---|-----|
| 0         | 0         | 4 | 8 | 12  |
| 1         | 0         | 7 | 19| 26  |
| Sum       |           | 11| 27| 38  |

Table 5. The number of transition conditions from state $i$ to state $j$ for state space $Y_t$

| State $i$ | State $j$ | 0 | 1 | Sum |
|-----------|-----------|---|---|-----|
| 0         | 0         | 9 | 9 | 18  |
| 1         | 0         | 9 | 11| 20  |
| Sum       |           | 18| 20| 38  |

From the process to produce Table 3, then the $X_t$ state is 1 if the number of $N_t$ tweets at one time is greater than the mean of all observations. while $X_t$ state is 0 if the number of $N_t$ tweets at one time of observation is smaller than the mean of all observations. likewise for $Y_t$ state is 1 if the number of tweets of $N_t$ at one time of observation is greater than the total median value of the observations, and $Y_t$ is 0 if it is smaller than the median value. the transition grouping to calculate the values 0 to 0, 0 to 1, 1 to 0 and 1 to 1 can be seen in Table 4 & Table 5.

3.4. Transition Opportunity Matrix

The process of classifying the number of tweets in Table 4 & Table 5, then a transition matrix can be made that represents the transition state of a low tweet (0) to a high tweet state (1), or a transition to itself that is low to low or high to high. Let’s say $P$ matrix of transition opportunities from the many covid-19 tweets. $P_1$ and $P_2$ show the transition probability matrix for state spaces $X_t$ and $Y_t$ respectively.

$$P_1 = \begin{bmatrix} \frac{4}{12} & \frac{8}{12} \\ \frac{7}{26} & \frac{19}{26} \end{bmatrix}, \quad P_2 = \begin{bmatrix} \frac{9}{18} & \frac{9}{18} \\ \frac{11}{20} & \frac{11}{20} \end{bmatrix}.$$

But to show graphically the transition can be modeled visually by using a transition diagram shown in Figure 2.

Figure 2. Transition diagram probability of covid-19 tweets; to the left (2a) is a data transition diagram of the number of tweets for $X_t$; to the right (2b) is the data transition diagram number of tweets for $Y_t$.

After the transition matrix is obtained, then the distribution of n-step opportunities is determined as a prediction of the probability of tweets with the covid-19 discussion trend. For example $\pi(n)$ n-step
probability distribution of the number of tweets with state space \(X_t\) (state space with mean), then from Figure 2 there are many transitions from state \(i\) to state \(j\) for state space \(X_t\), obtained initial distribution known \(\pi(0) = (12/38 \ 26/38)\). By using n-step probability distribution from Equation 5 the prediction of 1-step distribution \(\pi(1)\) is obtained as follows:

\[
\pi(1) = \pi(0) P^1 = \begin{pmatrix} 12/38 & 26/38 \end{pmatrix} \begin{pmatrix} 4/12 & 8/12 \\ 7/26 & 19/26 \end{pmatrix} = \begin{pmatrix} 0.289 & 0.711 \end{pmatrix}.
\]

The 1-step condition predicted that the probability of tweets was 0.289 in the low state, while 0.711 in the high state. Then in the 2-step distribution prediction \(\pi(2)\) by using by using a 1-step calculation technique, obtained \(\pi(2) = (0.287 \ 0.709)\). On 2-step conditions predicted the probability of tweets is 0.287 in low states, while 0.709 in high states. And so on until the 145-step is obtained, namely the number of tweets on May 10, 2020, the probability of tweets is 0.28767 in the low state and 0.71233 in the high state. So it is predicted that on that date the probability of tweets discussing with coronavirus and covid-19 keywords is still high as the outbreak of the epidemic has not subsided.

As for \(\pi(n)\) step opportunity distribution of the number of tweets with the state space \(Y_t\) (state space with median), the initial distribution is known \(\pi(0) = (18/38 \ 20/38)\). Following the method in the state space \(X_t\), the n-step opportunity distribution obtained using Equation 5. The values of 1-step and 2-step opportunity distribution are \(\pi(1) = (0.474 \ 0.526)\) and \(\pi(2) = (0.473 \ 0.527)\), respectively. Until the 145-step is obtained, namely the number of tweets in the median state space on May 10, 2020, the probability of a tweet is 0.474 in the low state and 0.526 in the high state.

Table 6. Computational Results Prediction n-step Distribution.

| \(n\)-step | Time               | Mean Space State \(X_t\) | Median Space State \(Y_t\) |
|------------|--------------------|--------------------------|---------------------------|
| Date       | Hour (a.m./p.m.)   | Low          | High         | Low | High         |
| 1          | 11.00 a.m.-11.20 a.m. | 0.28947       | 0.71053      | 0.47368 | 0.52632 |
| 2          | 23/3/2020 04.00 p.m.-04.20 p.m.  | 0.28779       | 0.71221      | 0.47368 | 0.52632 |
| 3          | 09.00 p.m.-09.20 p.m.     | 0.28768       | 0.71232      | 0.47368 | 0.52632 |
| ...        | ...                | ...           | ...          | ...   | ...         |
| 145        | 11.00 a.m.-11.20 a.m. | 0.28767 0.71233 | 0.47368 0.52632 |
| 146        | 10/5/2020 04.00 p.m.-04.20 p.m.  | 0.28767 0.71233 | 0.47368 0.52632 |
| 147        | 09.00 p.m.-09.20 p.m.     | 0.28767 0.71233 | 0.47368 0.52632 |

From Table 6 Predictions for the 10\(^{th}\) May 2020 or the next 49 days can be calculated the same as counting predictions for the 49 x 3 = 147, precisely starting at the 145\(^{th}\) steps. In the same way the probability distribution of the number of tweets is obtained with a relative average average and median covid-19 in the future 145 times (10 May 2020). Many tweets about covid-19 from two models in the next 7 steps and 147 are predicted to be in state-1 which is a lot of high tweets. Obtained opportunity distribution in the next 145-steps (10 May 2020) for many tweets in \(X_t\) and \(Y_t\) state space are respectively (0.28767, 0.71233) & (0.47368, 0.52632). The selection of different state spaces results in different probability distribution values for the predicted 7-step distribution for the same data.
Table 7. Difference in distribution of $i$ steps and $i-1$ steps.

| Step-$i$ | 2     | 3     | 4     | ...  | 145 |
|---------|-------|-------|-------|------|-----|
| $\pi(i) - \pi(i-1)$ | 0.001168691 | 0.000108135 | 6.93175E-06 | ...  | 4.0764E-174 |

The convergence of the n-step distribution can be calculated as the difference between the distribution for the 0 state from steps $i$ and $i+1$. In the number of tweet dataset in step 4 the distribution of relative conditions does not change for the first 6 decimal places as shown in Table 6.

4. Conclusion
The number of tweets about covid-19 on May 10, 2020 is estimated to be at a high level. The opportunity distribution of the many tweets about covid-19 is relatively unchanged in the first 7 times. Suggestions use more data and longer observation time, to get a better Markov Chain model. For policy makers to consider Twitter as one of the covid-19 information centers.

Acknowledgments
The authors thank Rector Universitas Padjadjaran who provided financial assistance for the dissemination of research through Library and Online Data Research Grant 2020 with contract number: 1735/UN6.3.1/LT/2020. Also thanks to the reviewers for the valuable review of this paper.

References
[1] Golbeck J 2015 Introduction to Social Media Investigation: A Hands-on Approach (Amsterdam: Syngress)
[2] Kietzmann J H , Hermkens K, McCarthy I P, and Silvestre B S 2011 Social media? Get serious! Understanding the functional building blocks of social media Business Horizon 54 242
[3] Tremayne M 2014 Anatomy of Protest in the Digital Era: A Network Analysis of Twitter and Occupy Wall Street Social Movement Studies 13 115
[4] Wang X, Yu Y, and Lin L 2020 Tweeting the United Nations Climate Change Conference in Paris (COP21): An analysis of a social network and factors determining the network influence Online Social Networks and Media 15 100059
[5] Mohammadi E, Thelwall M, Kwasny M, and Holmes K L 2018 Academic information on Twitter: A user survey Plose One 13 e0197265
[6] Golder S A and Macy M W 2011 Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures Science 333 1880
[7] Alalwan A A, Rana N P, Dwivedi Y K, and Algharabat R 2017 Social Media in Marketing: A Review and Analysis of the Existing Literature Telematics and Informatics 34 1183
[8] Miranda E, Aryuni M, Hariyanto R, and Surya E S 2019 Sentiment Analysis using Sentiwordnet and Machine Learning Approach (Indonesia general election opinion from the twitter content) ICIMTech 1 2019 19-20 August (Bali: Indonesia) 65
[9] Marseto F, Handayani P W, and Pinem A A 2019 Push, Pull, and Mooring Evaluation of User Switching Intention from Social Commerce to E-Commerce 1 2019 19-20 August (Bali: Indonesia) 576
[10] Pennacchiotti M and Gurumurthy S 2011 Investigating topic models for social media user recommendation Proc. 20th Int. Conf. Companion World Wide Web. 2011 March 28 - April 1 (Hyderabad: India) 101
[11] Zollo L, Filieri R, Rialti R, and Yoon S 2020 Unpacking the relationship between social media marketing and brand equity: The mediating role of consumers’ benefits and experience *Journal of Business Research* **117** 258

[12] Rutter R, Roper S, and Lettice F 2016 Social media interaction, the university brand and recruitment performance *Journal of Business Research* **69** 3099

[13] Cucinotta D and Vanelli M 2020 WHO declares COVID-19 a pandemic *Acta Bio Med* **91** 158

[14] Mathews P, Mitchell L, Nguyen G, Bean N 2017 The Nature and Origin of Heavy Tails in Retweet Activity *Proc. of the 26th Int. Conf on World Wide Web Companion* April 2017 (Perth: Australia) 1495

[15] Lee C and Wilkinson D 2018 A Hierarchical model of nonhomogeneous Poisson processes for twitter retweets *Journal of the American Statistical Association* **115** 4

[16] Hu F, Li Z, Yang C, and Jiang Y 2018 A graph-based approach to detecting tourist movement patterns using social media data *Cartography and Geographic Information Science* **46** 380

[17] Skaza J and Blais B 2017 Modeling the Infectiousness of Twitter Hashtags *Physica A: Statistical Mechanics and its Applications* **465** 292

[18] Hawkins D M 1980 *Identification of outliers* (London: Chapman and Hall)

[19] Aggarwal C C 2017 *Outlier Analysis* (New York, USA: Springer)

[20] Gagniuc P A 2017 *Markov Chains: From Theory to Implementation and Experimentation* (Hoboken, NJ, USA: John Wiley)