Communications for Statistical Applications and Methods
2020, Vol. 27, No. 4, 459–468

https://doi.org/10.29220/CSAM.2020.27.4.459
Print ISSN 2287-7843 / Online ISSN 2383-4757

Statistical analysis of the employment future for Korea

SangHyuk Lee[a], Sang-Gue Park[a], Chan Kyu Lee[b], Yaeji Lim[1,a]

[a]Department of Applied Statistics, Chung-Ang University, Korea;
[b]Department of Korean Language and Literature, Chung-Ang University, Korea

Abstract

We examine the rate of substitution of jobs by artificial intelligence using a score called the "weighted ability rate of substitution (WARS)." WARS is a indicator that represents each job's potential for substitution by automation and digitalization. Since the conventional WARS is sensitive to the particular responses from the employees, we consider a robust version of the indicator. In this paper, we propose the individualized WARS, which is a modification of the conventional WARS, and compute robust averages and confidence intervals for inference. In addition, we use the clustering method to statistically classify jobs according to the proposed individualized WARS. The proposed method is applied to Korean job data, and proposed WARS are computed for five future years. Also, we observe that 747 jobs are well-clustered according to the substitution levels.

Keywords: automation, employment, job clustering, job replacement, weighted ability rate of substitution

1. Introduction

Automation and digitalization have changed jobs and employment. Computers have substituted for a number of jobs and functions, including those of bookkeepers, cashiers, and telephone operators (Charles *et al.*, 2013). Employment has changed dramatically along with the development of artificial intelligence (AI). Even art and other creative fields, which have always been seen as distinctly human domains, have been affected by recent AI.

Numerous studies analyze this trend and model the effect of AI on employment. Autor and Dorn (2013) showed that as the costs of robots decline and technological capabilities expand, robots can be expected to gradually substitute for labor in a wide range of low-wage service occupations. Frey and Osborne (2013) estimated the susceptibility of employment to computerization using a Gaussian process classifier and concluded that about 47% of total US employment is in the high risk category. This result implies that the associated occupations may be automated in the next decade. More recently, Kim (2015) applied the model of Frey and Osborne (2017) to Korean job data and showed that 57% of jobs in Korea are in danger of being replaced by AI.

Arntz *et al.* (2016) pointed out that the model proposed by Frey and Osborne (2017) takes an occupation-based approach and may overestimate the results. Instead, they proposed a new method that estimates the probability of replacement by taking a task-based approach, and they concluded that, on average across the 21 OECD countries, only 9% of jobs are automatable. The share of automatable jobs in Korea is 6%, which is very small compared to the results of Kim (2015).

[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: yaeji.lim@gmail.com

Published 31 July 2020 / journal homepage: http://csam.or.kr

Park *et al.* (2016) compared these two approaches in the context of Korean employment and introduced the "weighted ability rate of substitution (WARS)" by taking a task-based approach. The WARS is a new indicator that represents potential of each job for substitution by automation and digitalization. However, since the WARS is computed from the average responses from the employees, it can not explain differences between individual employees in a given job. It is also highly sensitive to outliers. To handle these limitation, we propose a modified version of the WARS based on statistical methods. We compute an individual-based WARS to account for the variability of employees in a particular job. For statistical inference, we compute trimmed means and confidence intervals. In addition, we apply a clustering method to classify jobs based on their replacement rates.

The paper proceeds as follows. In Section 2, we briefly review the conventional WARS and propose a new version of the WARS and its clustering method. The analysis results and their interpretations are presented in Section 3. Lastly, Section 4 provides the conclusion and final remarks.

2. Methodology

2.1. Weighted ability rate of substitution

In this section, we first review the original WARS introduced by Park *et al.* (2016). The survey from the Korea Network for Occupations and Workers (KNOW) report provides the input variables for computing the WARS. This annual survey measures the abilities needed to perform tasks related to each job and the importance of these abilities to the job, and the questionnaire is designed using a five- or seven-point scale.

For $i = 1, \ldots, N$, $j = 1, \ldots, N_i$, $k = 1, \ldots, p$, let a_{ijk} be the level of the k^{th} ability related to a task for the i^{th} job based on the response of the j^{th} individual, and, for $i = 1, \ldots, N$, $j = 1, \ldots, N_i$, $k = 1, \ldots, p$, let m_{ijk} be the level of importance of the k^{th} ability related to a task for the i^{th} job based on the response of the j^{th} individual. Here, $N = 747$ is the number of jobs in the KNOW survey, N_i is the number of employees in the i^{th} job who responded to the KNOW survey, and $p = 44$ is the number of abilities in the survey, including "listening and understanding," "speaking," and "writing."

Consider two matrices, $\boldsymbol{A}$ and $\boldsymbol{M}$, which contain the average responses to the KNOW survey.

- $\boldsymbol{A}$ is an $N \times p$ matrix containing the average abilities required to perform tasks for each job;

$$\boldsymbol{A} = \begin{pmatrix} a_{1\cdot 1} & \cdots & a_{1\cdot p} \\ \vdots & \ddots & \vdots \\ a_{N\cdot 1} & \cdots & a_{N\cdot p} \end{pmatrix}, \tag{2.1}$$

  where $a_{i\cdot k} := (1/N_i) \sum_{j=1}^{N_i} a_{ijk}$, for $i = 1, \ldots, N$, $k = 1, \ldots, p$.

- $\boldsymbol{M}$ is an $N \times p$ matrix containing the average importance of the abilities for each job;

$$\boldsymbol{M} = \begin{pmatrix} m_{1\cdot 1} & \cdots & m_{1\cdot p} \\ \vdots & \ddots & \vdots \\ m_{N\cdot 1} & \cdots & m_{N\cdot p} \end{pmatrix}, \tag{2.2}$$

  where $m_{i\cdot k} := (1/N_i) \sum_{j=1}^{N_i} m_{ijk}$, for $i = 1, \ldots, N$, $k = 1, \ldots, p$.

Park *et al.* (2016) surveyed AI and robot experts to obtain replacement levels for each ability. If the k^{th} ability is likely to be replaced by AI in the future, then the experts are asked to give it a high score

in the questionnaire. These replacement levels were obtained for $p = 44$ abilities and $T = 5$ future time points: the years 2020, 2025, 2030, 2035, and 2045. Let e^t_{lk}, where $t = 1, \ldots, T$, $l = 1, \ldots, q$, and $k = 1, \ldots, p$, be the replacement level of the k^{th} ability in future year t according to the l^{th} expert. Then, consider the following matrix:

- $\boldsymbol{E}_t$ is a $q \times p$ matrix containing the replacement levels in future year t;

$$\boldsymbol{E}_t = \begin{pmatrix} e^t_{11} & \cdots & e^t_{1p} \\ \vdots & \ddots & \vdots \\ e^t_{q1} & \cdots & e^t_{qp} \end{pmatrix}, \quad \text{for } t = 2020, 2025, 2030, 2035, 2045, \tag{2.3}$$

  where $q = 21$ is the number of experts surveyed. As mentioned above, greater values of e^t_{lk} indicate a greater possibility that the k^{th} ability will be replaced by AI in year t.

Using $\boldsymbol{A}$, $\boldsymbol{M}$, and $\boldsymbol{E}_t$, the following algorithm is the procedure for generating the WARS for the i^{th} job in future year t.

1. Define the threshold vector $\boldsymbol{d}_t$ as

$$\boldsymbol{d}_t := \left(\tilde{e}^t_{\cdot 1}, \ldots, \tilde{e}^t_{\cdot p}\right),$$

   where $\tilde{e}^t_{\cdot k} = \bar{e}^t_{\cdot k} - \mathrm{sd}(e^t_{\cdot k})$ with $\bar{e}^t_{\cdot k} = (1/q) \sum_{l=1}^{q} e^t_{lk}$ and $\mathrm{sd}(e^t_{\cdot k}) = \sqrt{1/(q-1) \sum_{l=1}^{q} (e^t_{lk} - \bar{e}^t_{\cdot k})^2}$.

2. Compare the i^{th} row of $\boldsymbol{A}$ with $\boldsymbol{d}$. That is, generate the vector $\boldsymbol{b}_{i,t}$ of length p as

$$\boldsymbol{b}_{i,t} := \left(\mathbf{I}\left\{a_{i\cdot 1} < \tilde{e}^t_{\cdot 1}\right\}, \mathbf{I}\left\{a_{i\cdot 2} < \tilde{e}^t_{\cdot 2}\right\}, \ldots, \mathbf{I}\left\{a_{i\cdot p} < \tilde{e}^t_{\cdot p}\right\}\right),$$

   where $\mathbf{I}$ is an indicator function. If $\boldsymbol{b}_{i,t}$ contains a large number of ones, then the i^{th} job can be easily replaced by AI in future year t.

3. Now, we obtain the WARS of the i^{th} job at time t by multiplying $\boldsymbol{b}_{i,t}$ by the weight $\boldsymbol{M}$:

$$\mathrm{WARS}_{i,t} :=< \boldsymbol{b}_{i,t}, \bar{\boldsymbol{m}}_i >, \tag{2.4}$$

   where $\bar{\boldsymbol{m}}_i = \{1/(\sum_{k=1}^{p} m_{i\cdot k})\}(m_{i\cdot 1}, \ldots, m_{i\cdot p})$, for $i = 1, \ldots, N$.

2.2. Individualized WARS

Park *et al.* (2016) averaged the responses to the KNOW survey, as shown in (2.1) and (2.2), and this method obtains exactly one WARS for each job in year t. However, tasks may differ across individual employees in a given job, and thus, a job's replacement level may vary (Autor and Handel, 2013; Arntz *et al.*, 2016). We therefore propose an individualized WARS as an alternative to generating a WARS using averaged data.

The detailed description of the proposed algorithm is as follows.

1. Define the threshold vector $\boldsymbol{d}_t$ as

$$\boldsymbol{d}_t := (\tilde{e}^t_{\cdot 1}, \ldots, \tilde{e}^t_{\cdot p}),$$

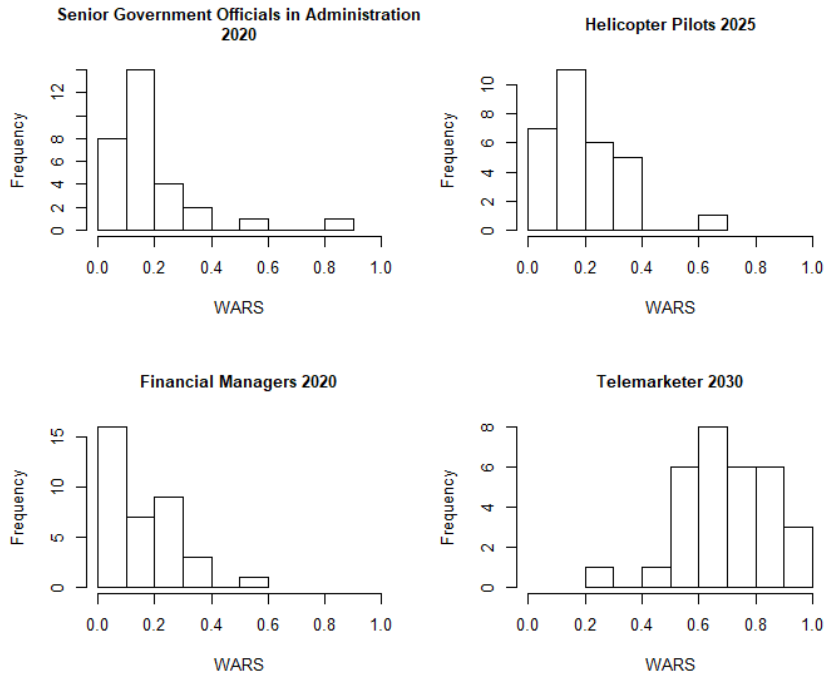

Figure 1: *Histograms of the individualized WARS for four randomly sampled jobs in the KNOW survey.*

where $\tilde{e}^t_{\cdot k} = \bar{e}^t_{\cdot k} - 0.5$. Note that $\boldsymbol{E}_t$ in (2.3) is not available; only $\bar{e}^t_{\cdot k}$ is provided to the public.

2. Generate the vector $\boldsymbol{b}_{i,j,t}$ of length p for $i = 1, \ldots, N$, $j = 1, \ldots N_i$, and $t = 1, \ldots, T$, as

$$\boldsymbol{b}_{i,j,t} := \left(\mathbf{I}\left\{a_{ija} < \tilde{e}^t_{\cdot 1}\right\}, \mathbf{I}\left\{a_{ij2} < \tilde{e}^t_{\cdot 2}\right\}, \ldots, \mathbf{I}\left\{a_{ijp} < \tilde{e}^t_{\cdot p}\right\}\right).$$

3. Now, the WARS of the j^{th} individual employed in the i^{th} job in year t is defined as

$$\text{WARS}_{i,j,t} := \left\langle \boldsymbol{b}_{i,j,t}, \bar{\boldsymbol{m}}_{ij} \right\rangle,$$

where $\bar{\boldsymbol{m}}_{ij} = \{1/(\sum_{k=1}^{p} m_{ijk})\}(m_{ij1}, \ldots, m_{ijp})$, for $i = 1, \ldots, N$, $j = 1, \ldots, N_i$, $t = 2020, \ldots, 2045$.

This measure differs from (2.4) because it uses j in $\text{WARS}_{i,j,t}$ to index individuals in the i^{th} job. Figure 1 shows histograms of the individualized WARS for four randomly selected jobs in the KNOW survey. These jobs are senior government officials in administration, helicopter pilots, financial managers, and telemarketers. We observe that the individualized WARS has a skewed distribution, implying that a simple average can provide misleading results.

We now compute summary statistics for inference, as follows.

- Robust WARS

We consider trimmed means instead of simple means since the distribution of the individualized

WARS is skewed (Figure 1). Then, the WARS for the i^{th} job in future year t is defined as

$$\overline{\text{WARS}}_{i,t} = \sum_{j=1+k}^{N_i-k} \text{WARS}_{i,j,t}, \tag{2.5}$$

where $k = \lfloor N_i \times a \rfloor$, with $a = 0.05$. Here, $\lfloor x \rfloor$ returns the greatest integer less than or equal to x.

Because we exclude extreme values of the WARS in computing its average value, the proposed WARS is more robust to outliers than those of Park *et al.* (2016). We call this proposed robust WARS the *RWARS*.

- Bootstrapping-based confidence interval

  For inference, we compute the confidence interval (CI) for the RWARS using bootstrapping. Efron (1979) first proposed the bootstrapping method, which is based on a resampling procedure, to estimate standard errors and biases. However, CIs can also be estimated using the bootstrapping method. Bootstrapping does not depend on the distribution assumption and provides more accurate results than conventional methods (Efron and Tibshirani, 1994; Carpenter and Bithell, 2000; Mudelsee and Alkio, 2007).

  Here, we use the bootstrap CI proposed by Efron (1987), that is, the nonparametric BC$_a$ method, which is a bias-corrected and accelerated version of the CI.

  Let $\hat{\theta}$ be the estimate of θ based on the observed data, and let $\hat{\theta}^*(b)$ be the estimate of θ using the bootstrap sample $b = 1, \ldots, B$. Then, define the cumulative density function (cdf) of $\hat{\theta}^*(b)$ as

$$\hat{H}(c) = \frac{1}{B} \# \left\{ \hat{\theta}^*(b) < c,\ b = 1, \ldots, B \right\}.$$

  Then, the $(1-\alpha) \times 100\%$ BC$_a$ interval is given by

$$\left(\hat{H}^{-1}\left(\tilde{z}_\alpha\right),\ \hat{H}^{-1}\left(\tilde{z}_{1-\alpha}\right) \right), \tag{2.6}$$

  with

$$\tilde{z}_\alpha = \Phi\left(z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)} \right), \tag{2.7}$$

  where $\Phi(\cdot)$ is the standard normal cdf and $z_\alpha = \Phi^{-1}(\alpha)$.

  This interval depends on two parameters, a and z_0, which are an acceleration parameter and a bias-correction factor, respectively. If both a and z_0 equal zero, (2.6) is simply obtained from the percentiles of the bootstrap replications. We estimate z_0 as

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\# \left\{ \hat{\theta}^*(b) < \hat{\theta},\ \ b = 1, \ldots, B \right\}}{B} \right),$$

  and we estimate a as

$$\hat{a} = \frac{\sum_{i=1}^{n} \left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)} \right)^3}{6 \left\{ \sum_{i=1}^{n} \left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)} \right)^2 \right\}^{\frac{3}{2}}},$$

  where $\hat{\theta}_{(i)}$ is the estimate of θ excluding the i^{th} observation and $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$.

  For more details on the nonparametric BC$_a$ method, see Efron and Tibshirani (1994).

2.3. Job clustering based on the WARS

Using the RWARS in (2.5), we cluster jobs based on the replacement potential. Park *et al.* (2016) also classified jobs using the WARS, but they simply divided the jobs into two groups using a threshold, 0.7, that was derived by Frey and Osborne (2017). They grouped jobs with a WARS above 0.7 into the "high risk group" and other jobs into the "low risk group." In addition, they calculated the geometric mean of the relative rates of change derived using the years 2016, 2020, and 2025. Using this geometric mean as a threshold, they defined two groups: the "high-changeable group" and the "low-changeable group."

Here, we statistically cluster the jobs using the density-based spatial clustering of applications with noise (DBSCAN) algorithm. The DBSCAN algorithm is a density-based clustering method that uses a non-parametric algorithm (Ester *et al.*, 1996). It groups points that are closely packed together and marks outliers points that lie alone in low-density regions. The DBSCAN algorithm has several advantages. First, unlike other clustering methods, it does not require the number of clusters in the data to be specified a priori. Furthermore, it can find arbitrarily shaped clusters, and is robust to outliers.

To apply the DBSCAN algorithm, two hyper-parameters need to be determined:

- epsilon (`eps`): Two points are considered neighbors if the distance between them is below the threshold epsilon.
- min.samples (`minPts`): Min.samples is the minimum number of neighbors a given point that must be classified as a core point.

In this study, we follow heuristic guidelines for these parameters. `minPts` is set equal to $2 \times d$, where d is the dimension of the dataset (Schubert *et al.*, 2017), and `eps` is determined based on the k-dist plot. The average distance from every point to its k nearest neighbors is computed, and these k-distances are plotted in an ascending order. We find an elbow point and define the distance of this point as `eps` (Ester *et al.*, 1996).

More details of the DBSCAN algorithm are provided by Tran *et al.* (2013).

3. Results

3.1. Data

In this study, we primarily use the 2015 KNOW survey, and also use the 2017 KNOW survey for newly introduced jobs, such as 3D printing technicians and peer-to-peer lending specialist. The dataset and survey are provided by the "Korea Employment Information Service" (`https://www.keis.or.kr`). We consider 747 jobs in the analysis.

3.2. Modified WARS

The individualized WARS values are computed for 747 jobs, and the RWARS and its CI for four randomly selected jobs are presented in Figure 2. We observe that the RWARS for food delivery is relatively high, whereas the RWARS for specialist surgeons is less than 0.6 even in 2045. However, the CI of the RWARS for specialist surgeons has widened over time. We also observe the difference of slope of RWARS. The RWARS for restaurant managers are steeper than that of food delivers.

Table 1 presents the list of jobs with large and small RWARS values in 2025. Similar results are obtained for other future years, but they are omitted from this study. The RWARS is larger for shopping hosts and farmers than for jobs that require more advanced abilities (such as comprehension or analytic thinking) such as credit analysts and accountants.

Figure 2: *RWARS and bootstrapping CI for four randomly selected jobs in the KNOW survey.*

Table 1: List of jobs with large and small RWARS values in 2025

Job	2025 RWARS (95% C.I)
Shopping Hosts	0.857 (0.798, 0.923)
Domestic Chores Helpers	0.824 (0.766, 0.893)
Crop Farmers	0.787 (0.750, 0.833)
Cement and Mineral Products Production Machine Operators	0.784 (0.737, 0.840)
⋮	⋮
Foreign Exchange Dealers	0.120 (0.082, 0.156)
Credit Analysts	0.109 (0.063, 0.150)
Accountants	0.103 (0.060, 0.137)
Financial Investment Analysts	0.102 (0.068, 0.127)

The box-plot in Figure 3 is used to compare the proposed RWARS with that of Park *et al.* (2016). We observe that the variance of the RWARS is lower than the original WARS, and the median of the RWARS is lower than the original WARS in 2030, 2035, and 2045, indicating that the proposed method is more conservative for estimates in the far future.

3.3. Clustering results

We apply the DBSCAN clustering method to the RWARS with `minPts` = 6 and `eps` = 0.3. We only use the RWARS in 2020, 2025, and 2030 for the clustering analysis because the reliability of the RWARS declines with time.

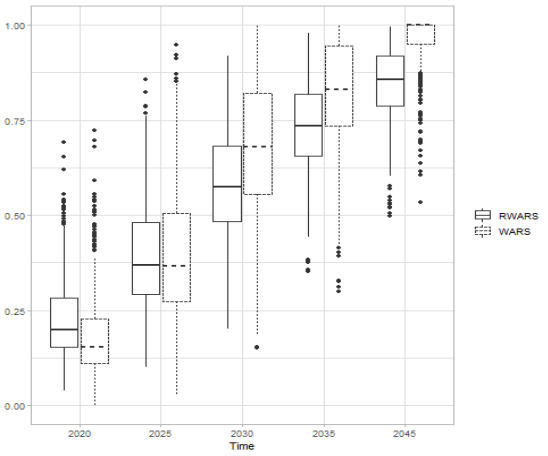

Figure 3: *Box-plot of the original WARS and the proposed RWARS.*

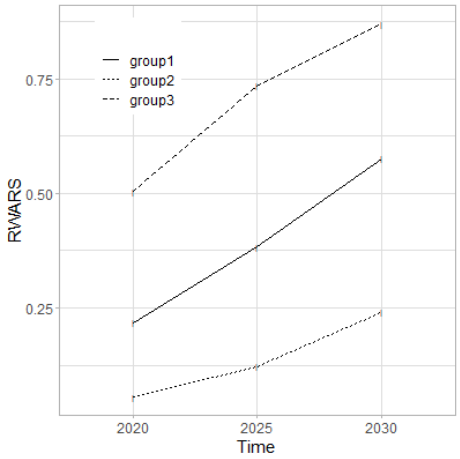

Figure 4: *Average RWARS with three cluster groups.*

The DBSCAN results show three clustering groups. Figure 4 plots the average RWARS in each group and Table 2 presents the list of jobs. Note that 12 jobs are considered as outliers in the DBSCAN algorithm and excluded from the results. The 12 outlier jobs are Product Planners Poet, Novelist, Players, Back Dancers, Discs Production Planners, Shopping Hosts, Domestic Chore Helpers, Bath Attendants, General Machinery Assemblers, Metal Processing Related Operators, and Telecommunication Network Operation Engineers. Especially, 'Shopping Hosts' and 'Domestic Chores Helpers' have the highest RWARS in 2025, while the 'Metal Processing Related Operators' has the highest RWARS in 2030.

We observe that group 3 has the largest RWARS, which implies that the jobs in group 3 are at a high risk of replacement. The jobs in group 2, which includes jobs like accountants and trading brokers, have the smallest RWARS. In addition, the RWARS for group 2 has a gentle slope, indicating that the jobs in group 2 experience less change over time.

Table 2: Clustering results obtained from the DBSCAN algorithm. The numbers in parentheses indicate the number of jobs in each group.

Group 1 (705)	Group 2 (9)	Group 3 (21)
Central Government Legislators	Accountants	Displayers
Local Government Legislators	Futures Trading Brokers	Door-to-Door Deliverers
Senior Corporate Officials	Foreign Exchange Dealers	Guards
General and Human Resources Managers	Public Prosecutors	Cleaners
⋮	⋮	⋮

4. Conclusion and discussion

In this study, we investigate the conventional WARS introduced by Park *et al.* (2016), and we improve it statistically. We propose an individualized WARS and compute its trimmed mean and CI for inference. The proposed method is applied to the KNOW survey, and the results show that this method classifies jobs well according to their replacement risk.

This study uses limited input data from the KNOW survey, meaning that the results can be improved with richer information related to jobs and employment. More information, such as salary data, can also be used to better interpret the results.

Acknowledgements

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No.2017S1A6A3A01078538).

References

Arntz M, Gregory T, and Zierahn U (2016). The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis, OECD Social, *Employment and Migration Working Papers*, No.189, OECD Publishing, Paris.

Autor DH and Dorn D (2013). The growth of low-skill service jobs and the polarization of the US labor market, *American Economic Review*, **103**, 1553–1597.

Autor DH and Handel MJ (2013). Putting tasks to the test: Human capital, job tasks, and wages, *Journal of labor Economics*, **31(S1)**, S59–S96.

Carpenter J and Bithell J (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, *Statistics in Medicine*, **19**, 1141–1164.

Charles KK, Hurst E, and Notowidigdo MJ (2013). *Manufacturing decline, housing booms, and non-employment* (Technical report), NBER Working Paper No. 18949. National Bureau of Economic Research.

Efron B (1979). Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, **7**, 1—6.

Efron B (1987). Better bootstrap confidence intervals, *Journal of the American statistical Association*, **82**, 171–185.

Efron B and Tibshirani RJ (1994). *An Introduction to the Bootstrap*, CRC Press, Florida.

Ester M, Kriegel HP, Sander J, and Xu X (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, (Vol. 96, No. 34, pp. 226-231).

Frey CB and Osborne MA (2013). *The Future of Employment: How Susceptible are Jobs to Computerization?*, University of Oxford.

Frey CB and Osborne MA (2017). The future of employment: How susceptible are jobs to computerisation?, *Technological Forecasting and Social Change*, **114**, 254–280.

Kim S (2015). Labor Market Changes and Response to Technological Progress, Korea Labor Institute.

Mudelsee M and Alkio M (2007). Quantifying effects in two-sample environmental experiments using bootstrap confidence intervals, *Environmental Modelling & Software*, **22**, 84–96.

Park G, Kang K, Kim D, Park S, Lee L, Hwang Y, Jun H, and Son Y (2016). A study on the future job, Korea Employment Information Service.

Schubert E, Sander J, Ester M, Kriegel HP, and Xu X (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN, *ACM Transactions on Database Systems (TODS)*, **42**, 1–21.

Tran TN, Drab K, and Daszykowski M (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters, *Chemometrics and Intelligent Laboratory Systems*, **120**, 92–96.

Received March 31, 2020; Revised May 13, 2020; Accepted May 14, 2020