

Deriving Public Sector Workforce Insights: A Case Study using Australian Public Sector Employment Profiles

Shameek Ghosh^{1,*}, Yi Zheng^{1,*}, Thorsten Lammers¹, Yinying Chen²,
Carolyn Fitzmaurice², Scott Johnston², Jinyan Li¹

¹University of Technology Sydney (UTS), Australia

²NSW Public Service Commission, Australia

{Shameek.Ghosh, Yi.Zheng-8}@student.uts.edu,
{Thorsten.Lammers, Jinyan.Li}@uts.edu.au, {Lucy.Chen,
Carolyn.Fitzmaurice, Scott.Johnston}@psc.nsw.gov.au

Abstract. Effective approaches for measurement of human capital in public sector and government agencies is essential for robust workforce planning against changing economic conditions. To this purpose, adopting innovative hypotheses driven workforce data analysis can help discover hidden patterns and trends about the workforce. These trends are useful for decision making and support the development of policies to reach desired employment outcomes. In this study, the data challenges and approaches to a real life workforce analytics scenario are described. Statistical results from numerous workforce data experiments are combined to derive three hypotheses that are useful to public sector organisations for human resources management and decision making.

Keywords: Workforce Analytics, Public Sector, Data Mining

1 Introduction

Human Resource Management in public administration plays a critical role in targeting and executing long term employment and workforce retention objectives. From a government perspective, it turns out to be a function correlating the employees performance to align with the government's strategic goals and priorities like sustainable budgeting for example [1, 2]. Additionally, with the increased use of digital technologies in the workforce, traditional jobs are at risk of being displaced and replaced by transformed job requirements and descriptions creating new types of jobs [6]. Thus the modernisation of existing workforce requires a robust workforce planning process that integrates future workforce needs so that capabilities are developed earlier. Typically, workforce planning processes involve the identification of strengths, weaknesses, opportunities and

*contributed equally

threats to develop appropriate interventions to mitigate future risks [3–5]. Previous applications of advanced data analysis techniques on workforce data have provided actionable insights on the effectiveness, efficiency, and impact of the workforce initiatives in the private sector related to retention and talent management [7–9, 16]. Data mining techniques like clustering and sequence mining have been used for resource management, staffing allocations and predicting employee attritions [13–15, 17]. In contrast, public sector organisations tend to have massive amounts of fragmented, disparate and unstructured datasets posing initial data processing challenges. Recent advances in large-scale data fusion [10], efficient service provisioning, and disaster mitigation [11] have been explored for public sector organisations. In the context of large-scale public sector workforce data, data-driven analysis drives a larger objective of conceptualizing and development of policy interventions that are relevant to the economic problems associated with maintaining an efficient and diverse national or state level public sector workforce.

The important contributions of this study are described as follows.

- A real life data exploration study for a subset of the Australian public sector workforce is reported.
- Existing data mining approaches are combined with summary statistics to derive actionable insights that can be used for workforce planning.
- Three data driven hypotheses are derived for future workforce data investigations.

2 Methods

2.1 Problem Description

Exploratory data mining questions were posed to derive workforce in-sights in two broad categories namely (1) Workforce Diversity, and (2) Ageing Workforce. In workforce diversity, the data analysis required finding emergent trends among employees for specific subpopulations viz. ethnic employees, employees with a disability and Aboriginal people. For an ageing workforce, data analysis was required to derive workforce implications due to population ageing.

2.2 Approach for addressing Workforce Diversity

For deriving relevant insights about workforce diversity, a number of subpopulations were extracted from the original datasets. Table 1 lists the variable names and valid values which were used to select three types of subpopulation datasets based on (1) Ethnicity, (2) Disability, and (3) Aboriginality.

Table 2 shows the various subpopulation combinations that were used to compare the population results for Aboriginal, Ethnic and mainstream groups of employees. Analogously, for comparisons with the Disability subpopulation, we use the combinations groups listed in Table 3.

Table 1. Variable names used in diversity analysis and their valid values

Variable Name (ID)	Valid Values
Aboriginal Person or Torres Strait Islander (2a)	1 = Aboriginal, 2 = Torres Strait Islander 3 = Aboriginal and Torres Strait Islander 4 = Not Aboriginal or Torres Strait Islander -7777/-9999 = Missing/Withdrawn
Person with a Disability (2b)	1 = Disability requiring adjustment at work. 2 = Disability not requiring adjustment at work. 3 = Person with a disability 4 = No disability -7777/-9999 = Missing/Withdrawn
Ethnicity (2c)	1 = Person from an Ethnic Minority Group 2 = Person not from an Ethnic Minority Group -7777/-9999 = Missing/Withdrawn

To understand the temporal trends of a certain subpopulation, two different methods are utilized. Taking the leadership as an example, in the first method, a temporal trend was derived using the percentage of leadership in a subpopulation - $Perc_L$, as given in equation 1. Leaders in the workforce data were defined as employees with an annual base salary greater than a specific salary band threshold (band 6 for senior executives encoded as GSE6, corresponding to \geq \$130929 for the year 2015).

Table 2. Selecting subpopulations for Aboriginal and Ethnic Minority

Name of the Subpopulation	Subpopulation Selection Criteria
Aboriginal	2a = 1 or 2 or 3
Ethnic	2c = 1 and 2a = 4
Mainstream ₁	2a = 4 and 2c = 2

Table 3. Selecting subpopulations for Disability

Name of the Subpopulation	Subpopulation Selection Criteria
Disability	2b = 1 or 2 or 3
Mainstream ₂	2b = 4

$$Perc_L^S = \frac{(EmployeeCountofsub - population \geq Band6)}{(TotalNumberofEmployeesinsub - population)} \quad (1)$$

For the second method, the percentage of employees (for each subpopulation) within the total leadership was calculated, as shown in equation 2.

$$Perc_L^T = \frac{(EmployeeCountofsub - population \geq Band6)}{(TotalNumberofEmployees \geq GSE6band)} \quad (2)$$

To acquire the distribution of the Aboriginal, ethnic, disability and main-stream populations in different occupations, the ANZSCO (Australian and New Zealand Standard Classification of Occupations) coding standard was employed [12]. ANZSCO is a classification system developed to collect, and analyse occupation statistics across government agencies in Australia. The x (x=1, 2, 3, 4, 5, 6) digits of an ANZSCO (Australian and New Zealand Standard Classification of Occupations) code follow a hierarchical tree structure, such that any digit at a given position is a child node of the prefix ANZSCO sequence. This means each position (or a digit) in the ANZSCO code corresponds to the suboccupation within the prefix code.

2.3 Approach for addressing Workforce Ageing Issues

For data analysis purposes, the calculation of age was carried out using the date-of-birth (DOB) variable, as shown in equation 3.

$$Age = CensusDate_{Year} - DOB \quad (3)$$

Within workforce ageing, for the identification of critical roles, specific occupations of interest were verified. Specifically, the occupations used for the ageing analyses are given in Table 4.

Table 4. Mapping of ANZSCO codes to selected Occupations

Occupation	ANZSCO
Accountants, Auditors and Company Secretaries	221
Information and Organisation Professionals	224
Engineering Professionals	233
Natural and Physical Science Professionals	234
Education Professionals	24
ICT Professionals	26
Office Managers and Program Administrators	51
General Clerical Workers	53
Numerical Clerks	55
Clerical and Office Support Workers	56

Additionally, to determine correlation scores with age, each of the numeric variables in the workforce dataset, were initially processed to remove missing values. Later, normalization of the data was carried out, with mean as 0 and

variance as 1. Following this, the Pearson product-moment correlation coefficient (PCC) was computed using the corresponding workforce numeric variable and age. Typically, the PCC is a measure of the linear correlation between two variables X and Y, giving a value between +1 and -1. Here, +1 is maximum positive correlation, 0 is no correlation, and -1 is maximum negative correlation. The PCC can be calculated as given in equation 4.

$$PCC = \frac{\sum(X_i - Mean(X))(Y_i - Mean(Y))}{\sqrt{\sum(X_i - Mean(X))^2} \sqrt{\sum(Y_i - Mean(Y))^2}} \quad (4)$$

Here, X and Y are considered to be the two variables for which PCC is being calculated.

3 Results

3.1 Dataset Description

For performing data exploration experiments, we obtained public sector workforce datasets from 1999 to 2015. The total number of variables in these datasets ranged between 40 to 83, for a given year. The number of variables for year varied due to the addition of new variables in later years. Summary statistics related to the Australian workforce data are reported next.

3.2 On Workforce Diversity

The ethnic sub-population demonstrated a three times higher likelihood of moving into a leadership position (defined by a salary band of 6 and above), in comparison to the aboriginal population (Figure 1).

The majority of the aboriginal population is working in the Community and Personal Service Profession (ANZSCO 4). Three of the top five ranked occupations belonged to ANZSCO code 4. However, there are only two aboriginal leaders of a total of 3493 aboriginal employees in ANZSCO code 4. In contrast, for every 956 mainstream employees in ANZSCO code 4, there is one mainstream leader. (Figure 2 and Figure 3).

In the ANZSCO code 4 occupations, which have been the third most popular occupation for the ethnic population and have almost 14% of the ethnic population, there is only one ethnic leader over GSE6 out of 5870 ethnic employees (Figure 3). Each subpopulation in the workforce exhibits an ageing trend (average age of 41.4 in 1999, compared to the average age of 45.2 in 2015). However, the disability population is ageing twice as fast as the mainstream group (7 years older on average for the disability population from 1999 to 2015, while only 2.7 years older for the mainstream population from 1999 to 2015, see Figure 4 and Figure 5).

The disability population in the lower salary bands of 2, 3, 4 and 5 are the main part leading to the population decrease of the disability group (Figure 6).

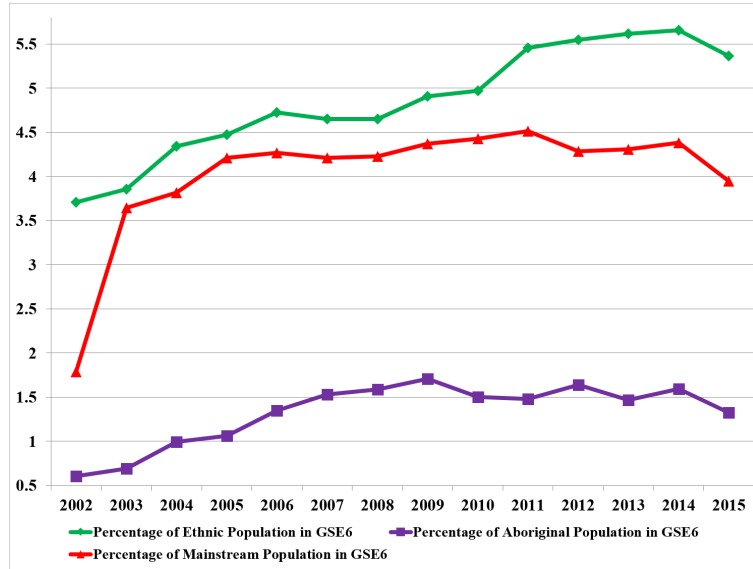


Fig. 1. Percentage of employees in each subpopulation with a GSE6 salary

Typically, in the leadership population, age has the highest negative correlation with unpaid sick leave (Table 5). Generally, in the leadership population leave variables tend to negative correlated with age.

3.3 On Workforce Ageing

Accountants had a median age of 43 years in 2015, which is two years younger than the median age of accountants in 2008. The median age for Clerical and Office Support Workers was 57 years in 2015, three years older than the median age of the workers of this profession in 2008.

In the leadership population, age has the highest negative correlation with unpaid sick leave (Table 5). Generally, in the leadership population leave variables tend to negative correlated with age. Evidence from PCC demonstrates that within the leadership cohort, age has a negative correlation to the unpaid and paid sick leave taken, when compared to the general population. This means that as age increases, less sick leaves had been taken in the senior leadership cohort, in 2015 (Table 5).

4 Hypothesis and Discussion

As part of the workforce study, the results lead us to determine specific insights about the public sector workforce which could be framed as hypotheses that may be further investigated. These are described further.

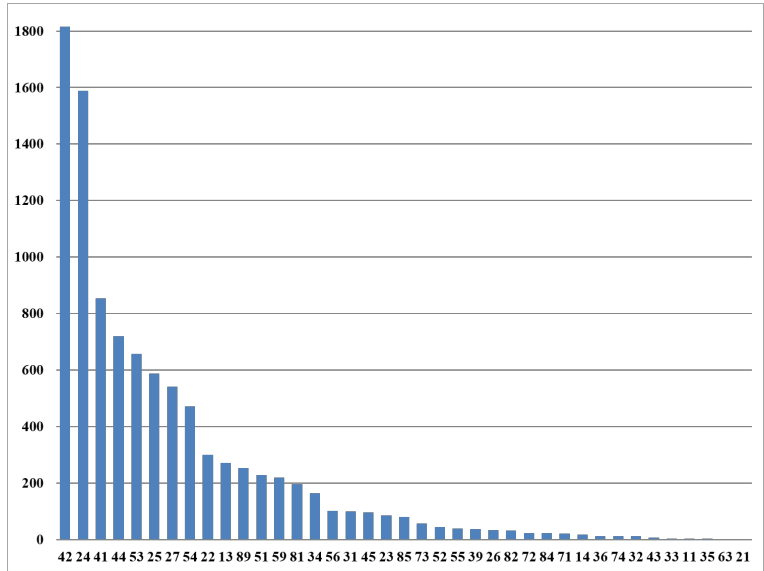


Fig. 2. Percentage of employees in each subpopulation with a GSE6 salary

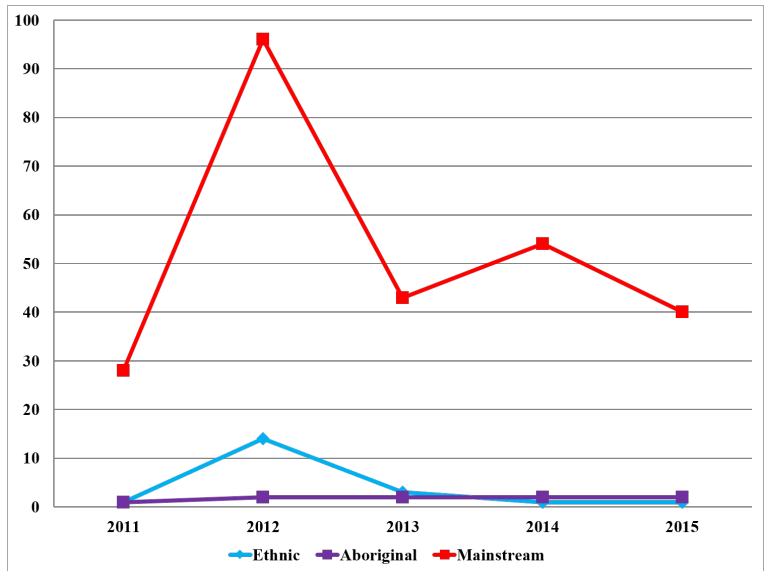


Fig. 3. Percentage of employees in each subpopulation with a GSE6 salary

Hypothesis 1: Given that bands 6 and 7 workers are older and ageing faster than the middle-band workers in the disability population, there might be a succession lag and gap in the leadership by the executives with a disability.

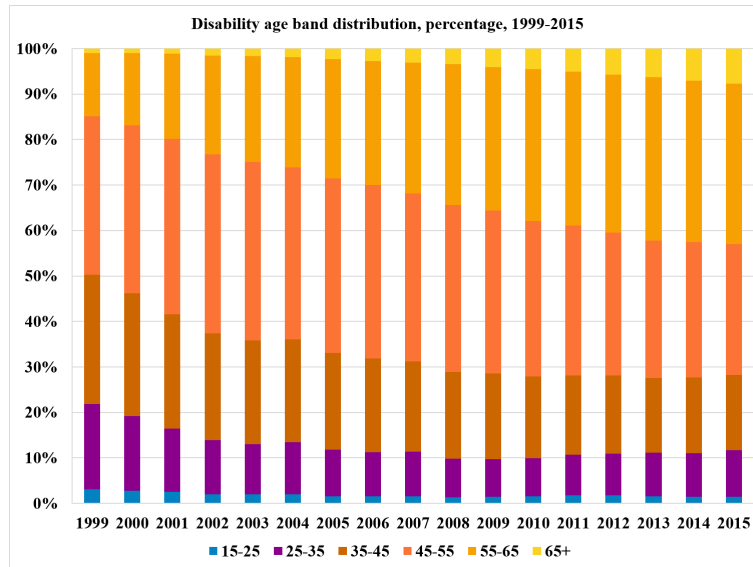


Fig. 4. Percentage of employees in each subpopulation with a GSE6 salary

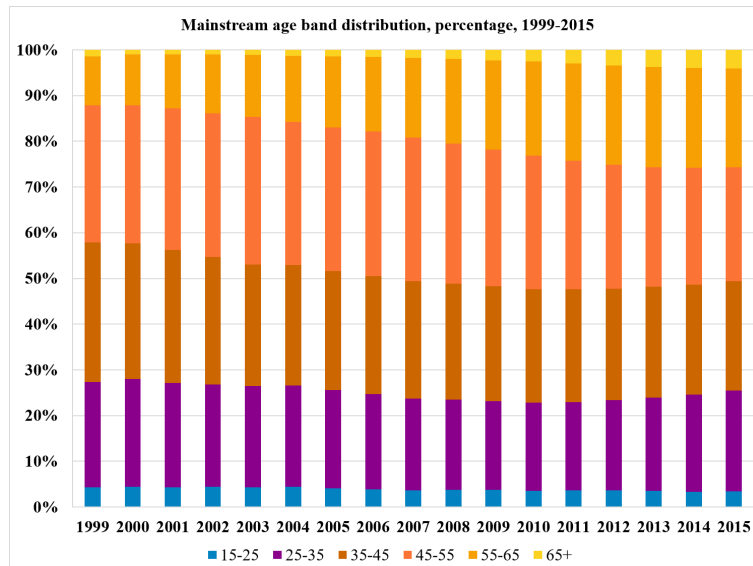


Fig. 5. Percentage of employees in each subpopulation with a GSE6 salary

Our results showed that the headcount of the workforce with a disability has been gradually decreasing from the year 1999 to 2015, which is mainly attributed to the decreasing subpopulation in salary bands 2, 3, 4 and 5. However, the head-

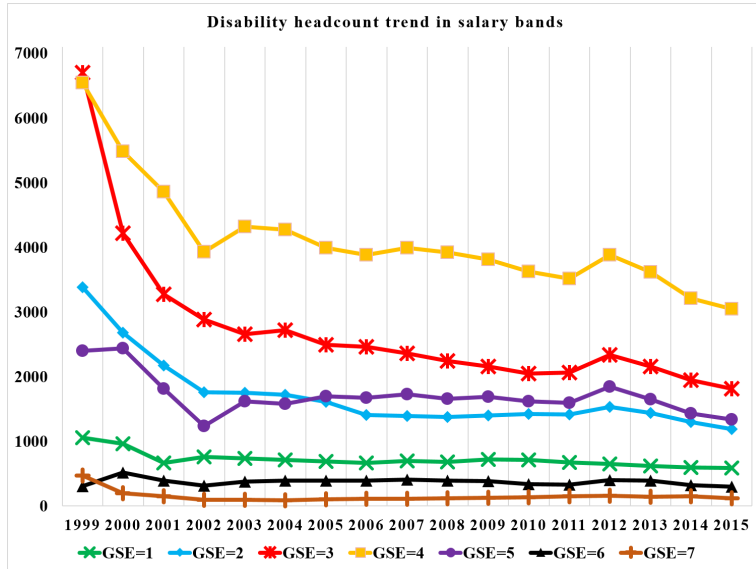


Fig. 6. Percentage of employees in each subpopulation with a GSE6 salary

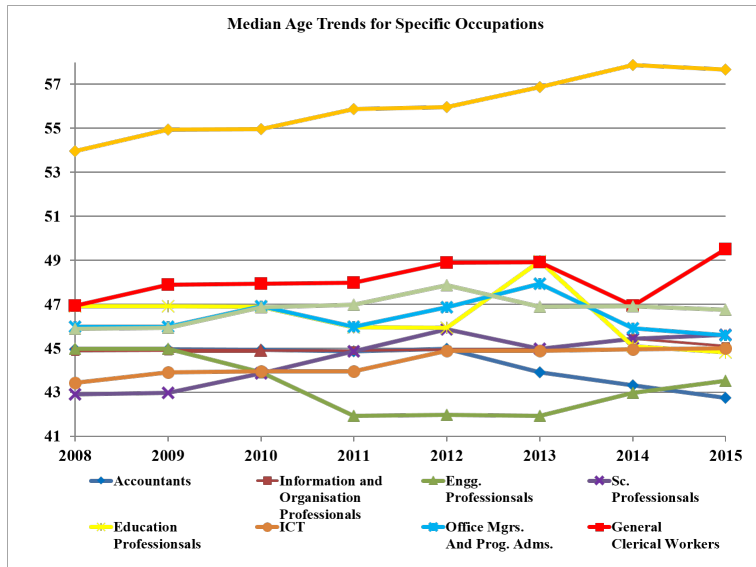


Fig. 7. Percentage of employees in each subpopulation with a GSE6 salary

counts in the salary bands 6 and 7 are maintained over recent years.

Hypothesis 2: Within the leadership cohort, there exists a comparatively strong negative correlation between leave variables and age. This negative correlation

Table 5. Top 10 Ranked Workforce Variables for the Leadership Population using Pearsons Correlation Coefficient.

Name of the Variable	Correlation with Age (PCC)
Unpaid Sick Leave Taken Ann Ref	-0.446
Unpaid Leave taken Ann Ref	-0.442
Paid Sick Leave Taken Ann Ref	-0.437
Extended Lave Taken on full pay Ann Ref	-0.427
Extended Lave Taken on half-pay Ann Ref	-0.352
Mat Pater Parent Leave Taken at full pay Ann Ref	-0.198
Special Leave Taken during the reference period	-0.187
Mat Pater Parent Leave Taken at half pay Ann Ref	-0.165
Paid Sick Leave Taken as Careers Leave Ann Ref	-0.152
Unpaid Mat Pater Parent Leave Taken Ann Ref	-0.125

does not exist for the general population.

The senior leaders in the workforce have taken much less sick leave (of different types including paid and unpaid) than the other senior staff.

Hypothesis 3: There does not exist adequate Aboriginal representation in leadership cohort, which both includes and affects the profession of "Community and Personal Service Workers".

The Community and Personal Service Workers (ANZSCO major 4) occupation ranks among the top 5 occupations for both ethnic and Aboriginal subpopulations in terms of the number of employees working in it. Each year, more employees of ethnic and aboriginal populations choose a profession within Community and Personal Service Workers as their occupation. However, there exist significantly less aboriginal leaders in Community and Personal Service Workers, when compared to the whole population.

5 Conclusion

In the current study, a real life public sector workforce dataset was employed to perform trend analysis from 1999 to 2015. In recent years, there have been few research efforts that employed a combination of data analytic methods to derive insights about the Australian public sector workforce. Specifically, our study emphasises the importance of statistical analysis and the need to combine findings with knowledge discovery techniques for generating actionable outcomes that can be usable for workforce interventions.

References

1. A. Adams, C., Muir, S., Hoque, Z. (2014). Measurement of sustainability performance in the public sector. Sustainability Accounting, Management and Policy Journal, 5(1), 46-67.

2. Christensen, Tom, and Per Lgreid. "The wholeofgovernment approach to public sector reform." *Public Administration Review* 67, no. 6 (2007): 1059-1066.
3. MacCrory, F., Westerman, G., Alhammadi, Y., Brynjolfsson, E. (2014, May). *Racing With and Against the Machine: Changes in Occupational Skill Composition in an Era of Rapid Technological Advance*. In ICIS.
4. Bessen, J. E. (2015). *How Computer Automation Affects Occupations: Technology, Jobs, and Skills*. Boston Univ. School of Law, Law and Economics Research Paper, (15-49).
5. Woon, Wei Lee, Zeyar Aung, Wala AlKhader, Davor Svetinovic, and Mohammad Atif Omar. "Changes in Occupational Skills-A Case Study Using Non-negative Matrix Factorization." In *International Conference on Neural Information Processing*, pp. 627-634. Springer International Publishing, 2015.
6. Caselli, M., 2014. Trade, skill-biased technical change and wages in Mexican manufacturing. *Applied Economics*, 46(3), pp.336-348.
7. Wei, D., Varshney, K.R. and Wagman, M., 2015, June. Optigrow: People analytics for job transfers. In *2015 IEEE International Congress on Big Data* (pp. 535-542). IEEE.
8. Ramamurthy, K.N., Singh, M., Davis, M., Kevern, J.A., Klein, U. and Peran, M., 2015, November. Identifying Employees for Re-skilling Using an Analytics-Based Approach. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)* (pp. 345-354). IEEE.
9. Varshney, K.R., Chenthamarakshan, V., Fancher, S.W., Wang, J., Fang, D. and Mojsilovi, A., 2014, August. Predicting employee expertise for talent management in the enterprise. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1729-1738). ACM.
10. Amorim, J. A., Andler, S. F., Gustavsson, P. M., Agostinho, O. L. (2013, October). *Big Data Analytics in the Public Sector: Improving the Strategic Planning in World Class Universities*. In *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, 2013 International Conference on (pp. 155-162). IEEE.
11. Morabito, V. (2015). *Big data and analytics for government innovation*. In *Big Data and Analytics* (pp. 23-45). Springer International Publishing.
12. Trewin, D., Trewin, D.J. and Pink, B.N., 2006. ANZSCO: Australian and New Zealand standard classification of occupations. Australian Bureau of Statistics/Statistics New Zealand.
13. Datta, R., Hu, J., Ray, B. (2008). Sequence mining for business analytics: Building project taxonomies for resource demand forecasting. *Frontiers in Artificial Intelligence and Applications*, 133.
14. Hu, J., Ray, B. K., Singh, M. (2007). Statistical methods for automated generation of service engagement staffing plans. *IBM Journal of Research and Development*, 51(3.4), 281-293.
15. Mojsilovi, A., Connors, D. (2010). *Workforce analytics for the services economy*. In *Handbook of service science* (pp. 437-460). Springer US.
16. Richter, Y., Naveh, Y., Gresh, D. L., Connors, D. P. (2008). Optimatch: applying constraint programming to workforce management of highly skilled employees. *International Journal of Services Operations and Informatics*, 3(3-4), 258-270.
17. Singh, M., Varshney, K. R., Wang, J., Mojsilovic, A., Gill, A. R., Faur, P. I., Ezry, R. (2012, December). An analytics approach for proactively combating voluntary attrition of employees. In *2012 IEEE 12th International Conference on Data Mining Workshops* (pp. 317-323). IEEE.