2012 3rd International Conference on e-Education, e-Business, e-Management and e-Learning
IPEDR vol.27 (2012) © (2012) IACSIT Press, Singapore

Statistical analysis and consideration of personal information leakage from the viewpoint of locality

Hitoshi Fumikura [1], Tetsuro Kobayashi [2], Ryoichi Sasaki [1]

[1] Graduate School of Advanced Science and Technology, Tokyo Denki University
[2] National Institute of Informatics

Abstract. Recently, due to the improvement of information and communication technologies, almost all information is treated as digital data. Computerization can lead to a huge improvement of business efficiency. However, information leakage to a network has become a new problem. Retrieving information that has flowed out into a network is very difficult, because replication of digital data is easy. The information leakage can lead to many serious consequences for the responsible holder, such as repayment of much money and loss of trust of the people affected. This study deals with a statistical analysis and consideration of personal information leakage using reported incidents of information leakage, especially from the viewpoint of local characteristics.

Keywords: Personal data leak, Locality, Statistical analysis, Epidemiology, Clustering

1. Introduction

Recently, due to the improvement of information and communication technologies, almost all information is treated as digital data. Computerization can lead to a huge improvement of business efficiency. However, information leakage to a network has become a new problem.

Retrieving information that has flowed out into a network is very difficult, because replication of digital data is easy. Information involved in leakage incidents is classified as either personal information or confidential organizational information. This paper deals with personal information leakage. The leakage of personal information puts the holder of the information into many serious situations, such as legal responsibility and responsibility for compensations for damage and even fatal injury. Therefore, analysis of past information leakages and enacting countermeasures are very important.

This study deals with the statistical analysis of personal information leakage in Japan from the viewpoint of locality, such as districts. Incidents of information leakage were collected from the homepage of a continually updated security website. Although an analysis of personal information leakage was carried out by the Japan Network Security Association (JNSA) [1], an analysis by city and district was not performed. In the medical field, an analysis by city and district was carried out as an epidemiologic survey; however, such studies have not been performed in other fields, except for our studies [2] [3]. The present study increased the range of gathered data and added statistical analyses, such as cluster analysis.

2. Study method

We surveyed the website called “Security NEXT” [4], which reports the leakage incidents in Japan in the form of articles. Security NEXT investigates all reports on personal information leakages reported to the proper authorities from companies and institutes in Japan. Security NEXT shows all the incidents on its homepage.

After collecting the information, we listed such details as the place leaking the information, the medium, the cause of the leak, the leak year, and the number of people affected. Then, a static analysis such as the cluster analysis using these data was carried out. In this study, we used the static analysis methods used in epidemiology [5] [6].

3. Surveyed results

3.1 Results of the simple analysis

According to the reports of Security NEXT, the number of leak incidents and the total number of people with leaked information are shown in Figure 1. Although the total number of leak incidents decreased after the peak of 2007, the number of leaks per person increased after 2008. The data in all the figures and tables in this paper were collected for the years 2005–2010.

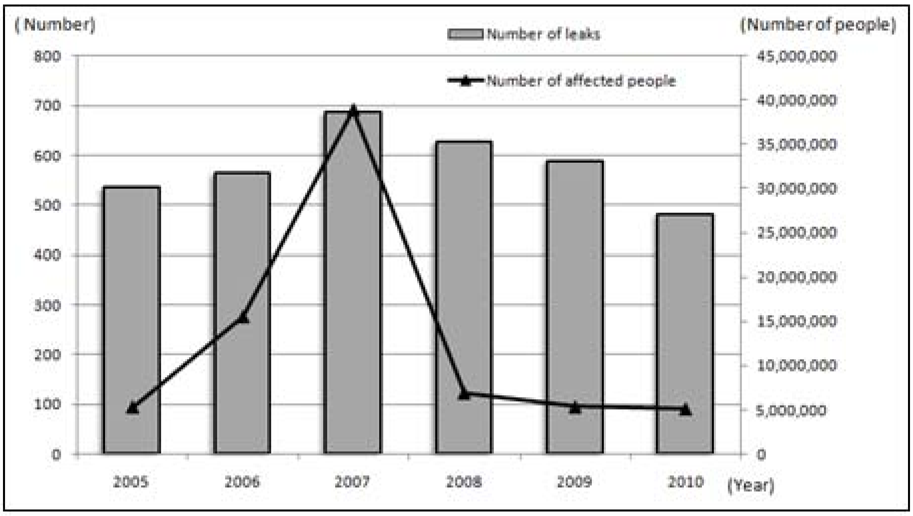


Figure 1: Cases of personal data leakages

Figure 2 shows the fraction (%) of mediums causing personal information leakage incidents. We know that paper, movable media, and PCs are the main mediums causing information leaks.

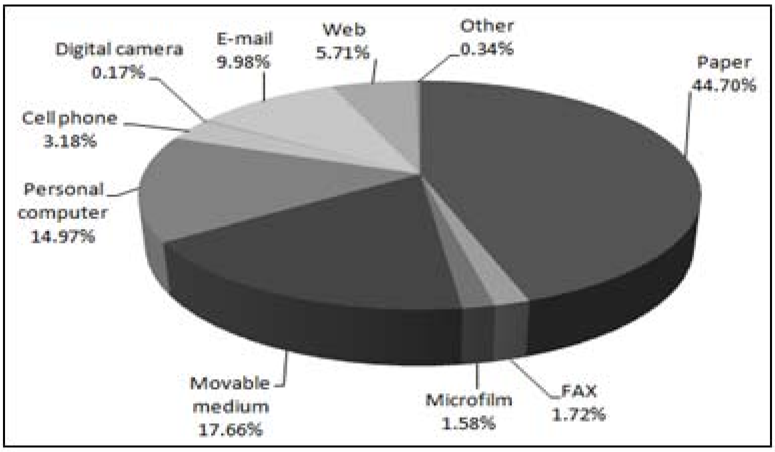


Figure 2: Ratio of mediums causing all information leakages in the years 2005–2010

The ratio (%) of categorized number of the leaked personal information per one leak incident of years 2005-2010 is illustrated in Figure 3.

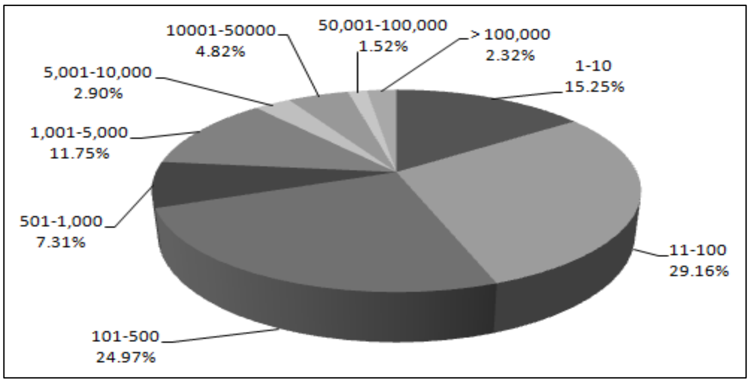


Figure 3: Number of items of leaked personal information per one incident in the years 2005–2010

Figure 4 shows the number of items of leaked information by type of industry [7].

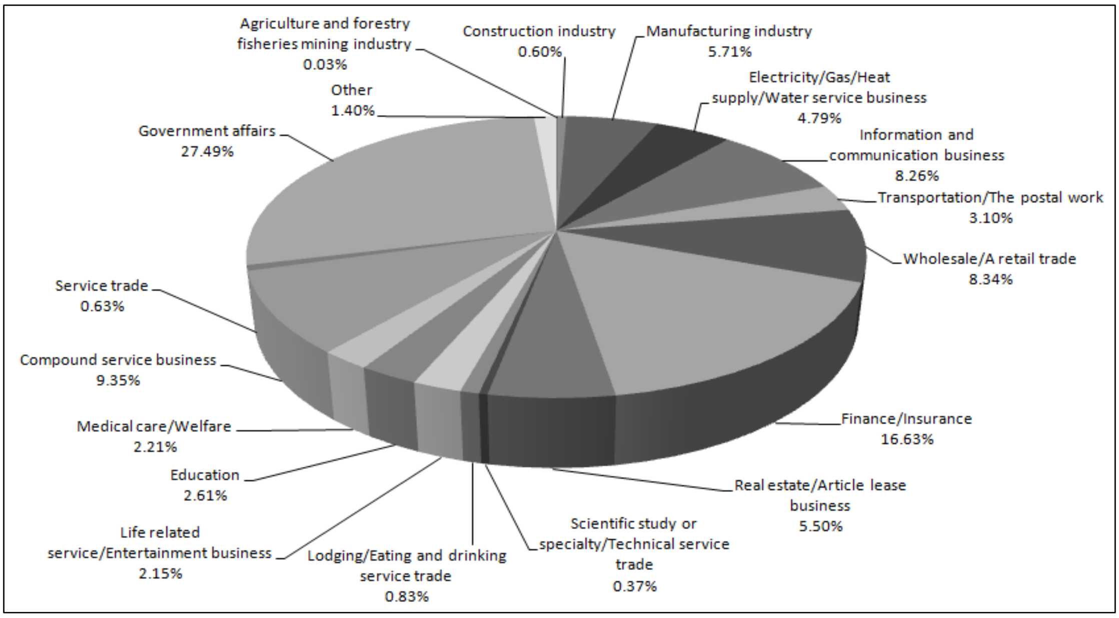


Figure 4: Number of items of leaked information by type of industry in the years 2005–2010

Figure 5 shows the number of information leakage incidents by cause. The main causes of leaks are "loss" and "theft.

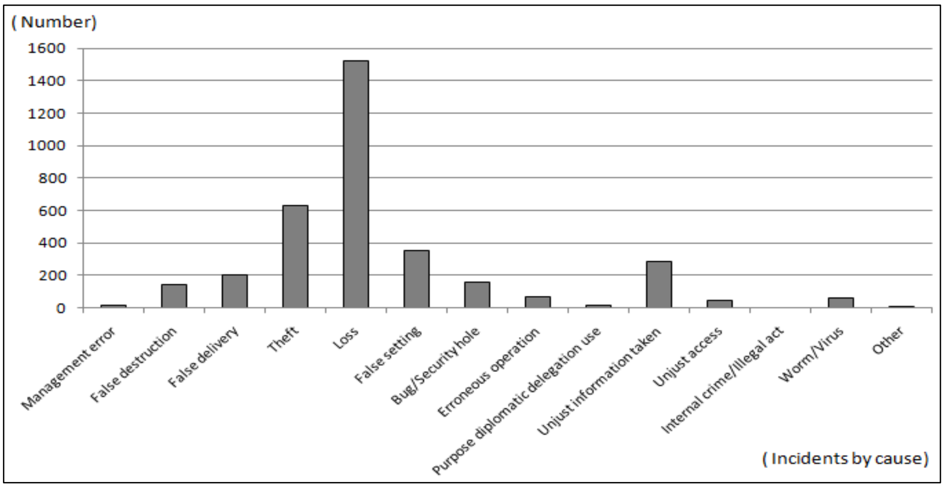


Figure 5: Number of information leakage incidents by cause

Figures 6 and 7 show the number of items of leaked information per 100,000 person-months by distinguishing cities and district and the number of people. Information leakage analyses distinguishing cities and districts have not previously been carried out.

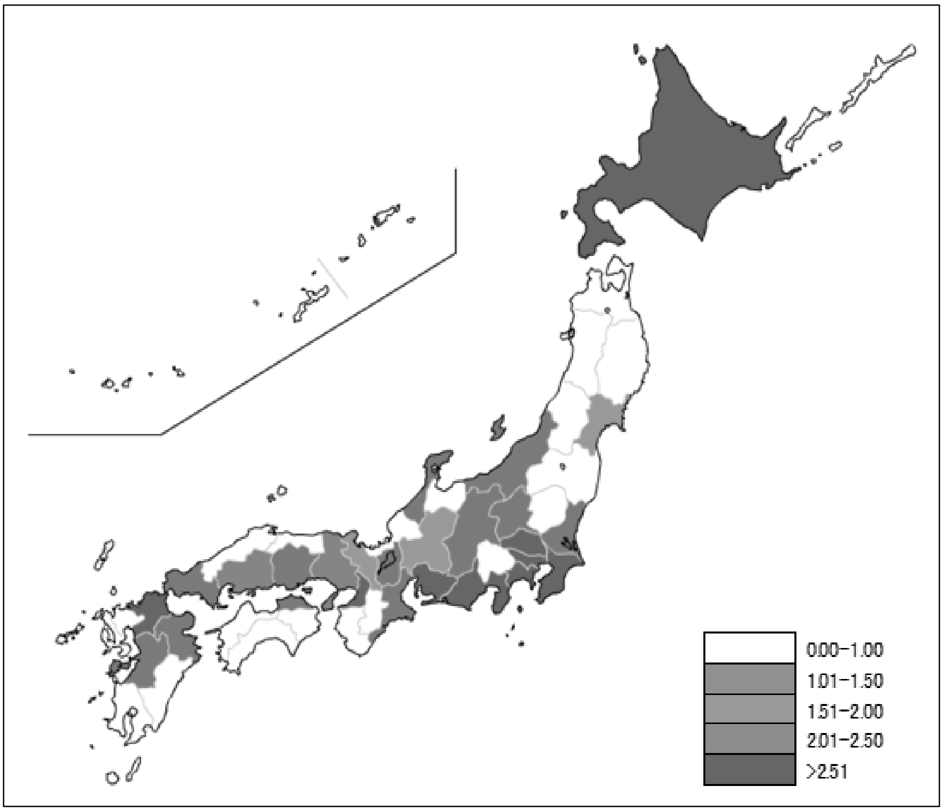


Figure 6: The information leakage number per 100,000 person-months by city and district

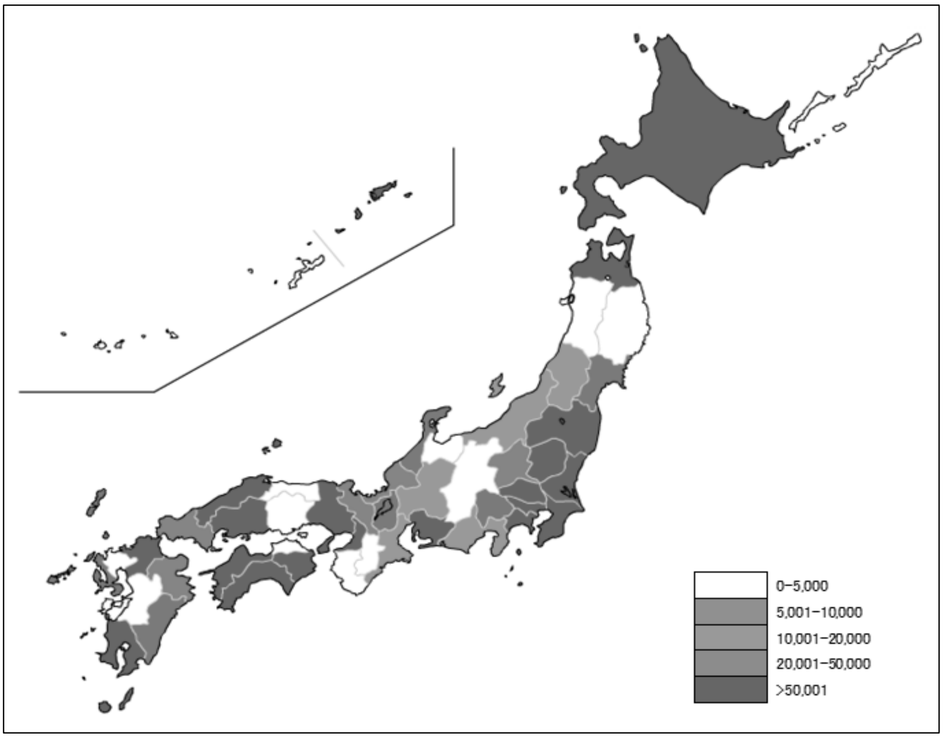
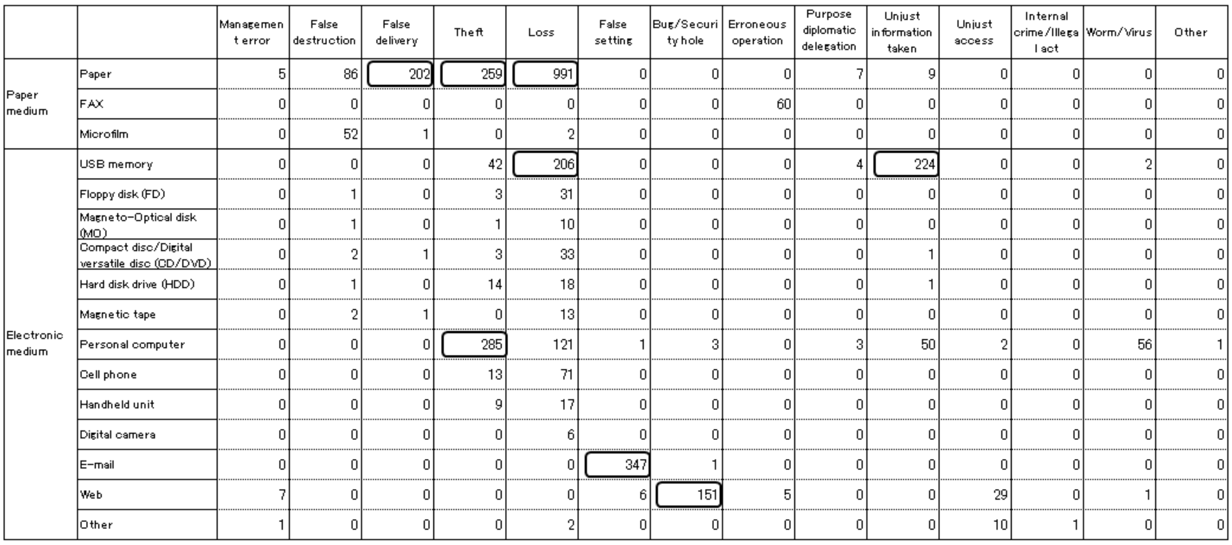


Figure 7: The information leakage number of people per 100,000 person-months by city and district

3.2 Result of cross-tabulation analysis

Table 1 is a cross-tabulation list between the medium (paper/electronic) and the cause of leaked personal information. We know that the relationships of the items circled are strong.

Table 1: Cross-tabulation between paper/electronic medium and cause of leaked personal information



		Management error	False destruction	False delivery	Theft	Loss	False setting	Bug/Security hole	Erroneous operation	Purpose diplomatic delegation	Unjust information taken	Unjust access	Internal crime/Illegal act	Worm/Virus	Other
Paper medium	Paper	5	86	202	259	991	0	0	0	7	9	0	0	0	0
	FAX	0	0	0	0	0	0	0	60	0	0	0	0	0	0
	Microfilm	0	52	1	0	2	0	0	0	0	0	0	0	0	0
Electronic medium	USB memory	0	0	0	42	206	0	0	0	4	224	0	0	2	0
	Floppy disk (FD)	0	1	0	3	31	0	0	0	0	0	0	0	0	0
	Magneto-Optical disk (MO)	0	1	0	1	10	0	0	0	0	0	0	0	0	0
	Compact disc/Digital versatile disc (CD/DVD)	0	2	1	3	33	0	0	0	0	1	0	0	0	0
	Hard disk drive (HDD)	0	1	0	14	18	0	0	0	0	1	0	0	0	0
	Magnetic tape	0	2	1	0	13	0	0	0	0	0	0	0	0	0
	Personal computer	0	0	0	285	121	1	3	0	3	50	2	0	56	1
	Cell phone	0	0	0	13	71	0	0	0	0	0	0	0	0	0
	Handheld unit	0	0	0	9	17	0	0	0	0	0	0	0	0	0
	Digital camera	0	0	0	0	6	0	0	0	0	0	0	0	0	0
	E-mail	0	0	0	0	0	347	1	0	0	0	0	0	0	0
	Web	7	0	0	0	0	6	151	5	0	0	29	0	1	0
	Other	1	0	0	0	2	0	0	0	0	0	10	1	0	0

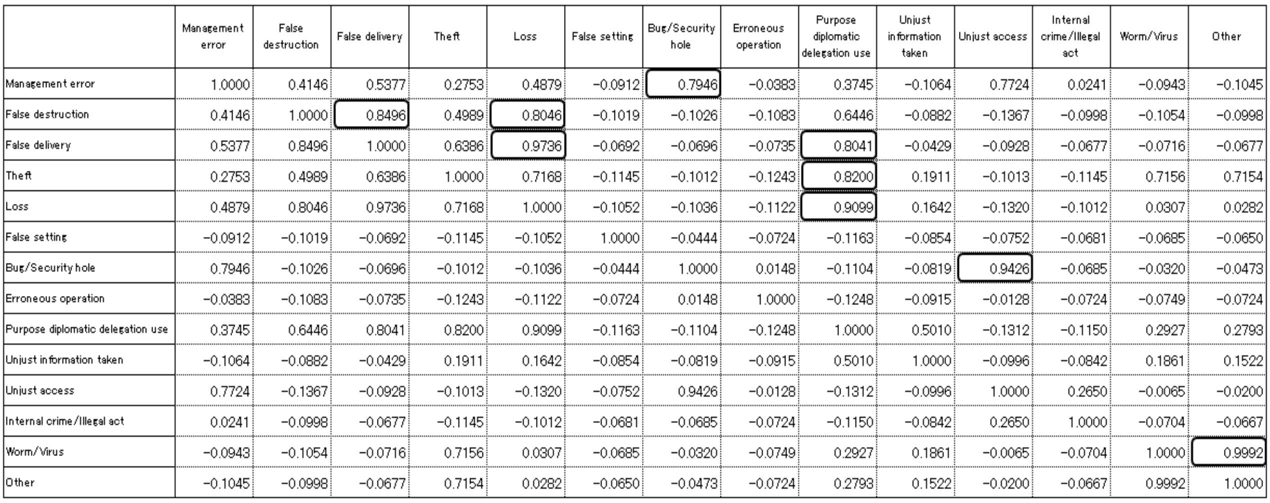
4. Statistical analysis

4.1 Correlation analysis

Table 2 represents the results of the correlation analysis for the causes of personal information leakage. The following correlation values are high.

- The correlation of false delivery and false destruction is high (0.8496).
- The correlation value of false delivery and loss of data, and the correlation value of false destruction and loss of data are high (0.8046 and 0.9736, respectively).
- The correlation value of a management error and a security hole caused by a bug is high (0.7946).
- The correlation values of false delivery, theft, and loss related to usage out of objective are high (0.8041, 0.8200, and 0.9099, respectively).
- The correlation of unjust access and a security hole caused by a bug is high (0.9426).

Table 2: As a result of correlation analysis



	Management error	False destruction	False delivery	Theft	Loss	False setting	Bug/Security hole	Erroneous operation	Purpose diplomatic delegation use	Unjust information taken	Unjust access	Internal crime/Illegal act	Worm/Virus	Other
Management error	1.0000	0.4146	0.5377	0.2753	0.4879	−0.0912	0.7946	−0.0383	0.3745	−0.1064	0.7724	0.0241	−0.0943	−0.1045
False destruction	0.4146	1.0000	0.8496	0.4989	0.8046	−0.1019	−0.1026	−0.1083	0.6446	−0.0882	−0.1367	−0.0998	−0.1054	−0.0998
False delivery	0.5377	0.8496	1.0000	0.6386	0.9736	−0.0692	−0.0696	−0.0735	0.8041	−0.0429	−0.0928	−0.0677	−0.0716	−0.0677
Theft	0.2753	0.4989	0.6386	1.0000	0.7168	−0.1145	−0.1012	−0.1243	0.8200	0.1911	−0.1013	−0.1145	0.7156	0.7154
Loss	0.4879	0.8046	0.9736	0.7168	1.0000	−0.1052	−0.1036	−0.1122	0.9099	0.1642	−0.1320	−0.1012	0.0307	0.0282
False setting	−0.0912	−0.1019	−0.0692	−0.1145	−0.1052	1.0000	−0.0444	−0.0724	−0.1163	−0.0854	−0.0752	−0.0681	−0.0685	−0.0650
Bug/Security hole	0.7946	−0.1026	−0.0696	−0.1012	−0.1036	−0.0444	1.0000	0.0148	−0.1104	−0.0819	0.9426	−0.0685	−0.0320	−0.0473
Erroneous operation	−0.0383	−0.1083	−0.0735	−0.1243	−0.1122	−0.0724	0.0148	1.0000	−0.1248	−0.0915	−0.0128	−0.0724	−0.0749	−0.0724
Purpose diplomatic delegation use	0.3745	0.6446	0.8041	0.8200	0.9099	−0.1163	−0.1104	−0.1248	1.0000	0.5010	−0.1312	−0.1150	0.2927	0.2793
Unjust information taken	−0.1064	−0.0882	−0.0429	0.1911	0.1642	−0.0854	−0.0819	−0.0915	0.5010	1.0000	−0.0996	−0.0842	0.1861	0.1522
Unjust access	0.7724	−0.1367	−0.0928	−0.1013	−0.1320	−0.0752	0.9426	−0.0128	−0.1312	−0.0996	1.0000	0.2650	−0.0065	−0.0200
Internal crime/Illegal act	0.0241	−0.0998	−0.0677	−0.1145	−0.1012	−0.0681	−0.0685	−0.0724	−0.1150	−0.0842	0.2650	1.0000	−0.0704	−0.0667
Worm/Virus	−0.0943	−0.1054	−0.0716	0.7156	0.0307	−0.0685	−0.0320	−0.0749	0.2927	0.1861	−0.0065	−0.0704	1.0000	0.9992
Other	−0.1045	−0.0998	−0.0677	0.7154	0.0282	−0.0650	−0.0473	−0.0724	0.2793	0.1522	−0.0200	−0.0667	0.9992	1.0000

4.2 Cluster analysis

4.2.1 Overview of the cluster analysis

Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Cluster analysis can be classified into hierarchical cluster analysis and non-hierarchical cluster analysis. In our study, we used the hierarchical cluster analysis, which shows the distance between clusters by using a visual tree-like chart called the dendrogram (tree diagram).

4.2.2 Results of the cluster analysis

Figure 8 is a dendrogram of the results of the hierarchical cluster analysis [8] [9] [10] for the cause of personal data leaks by city and district. Because the number of clusters was 10 and the distance of the second stage was large, as shown in Figure 9, the number of clusters was decided to be 3. In Figure 9, the classification is as follows: G7-G8 in group 1, G1-G6 in group 2, G9-G10 in group 3.

As shown in Figure 10, cluster 2 is a group with many unjust outflows via electronic mediums. In contrast, cluster 3 is a group with many outflows via the paper medium, and cluster 1 is located in the middle between cluster 2 and cluster 3.

In addition, Figure 10 shows the cities and districts included in each cluster on a Japanese map. Cluster 2 with many unjust outflows via electronic mediums includes small prefectures such as Toyama, Fukui, Shiga, Wakayama, Kagawa, and districts including big cities such as Tokyo. Cluster 3 with many outflows via the paper medium includes rural prefectures such as Aomori, Akita, Iwate, Yamanashi, Nara, and Tokushima.

Cluster 1, which has no special characteristics, includes many districts including big prefectures such as Osaka and Aichi.

We consider that the prefectures included in cluster 2 may have had personal information computerized at an earlier stage and many outflows by the electronic medium occurred by slight mistakes.

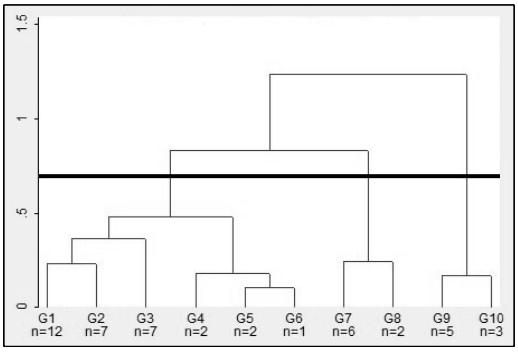


Figure 8: Dendrogram obtained from the cluster analysis for the cause of personal data leaks

by city district

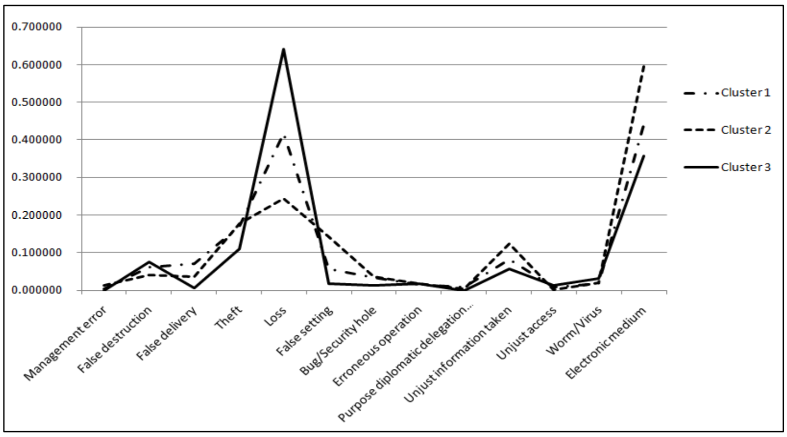


Figure 9: Results of the cluster analysis for the cause of personal data leaks by city and district

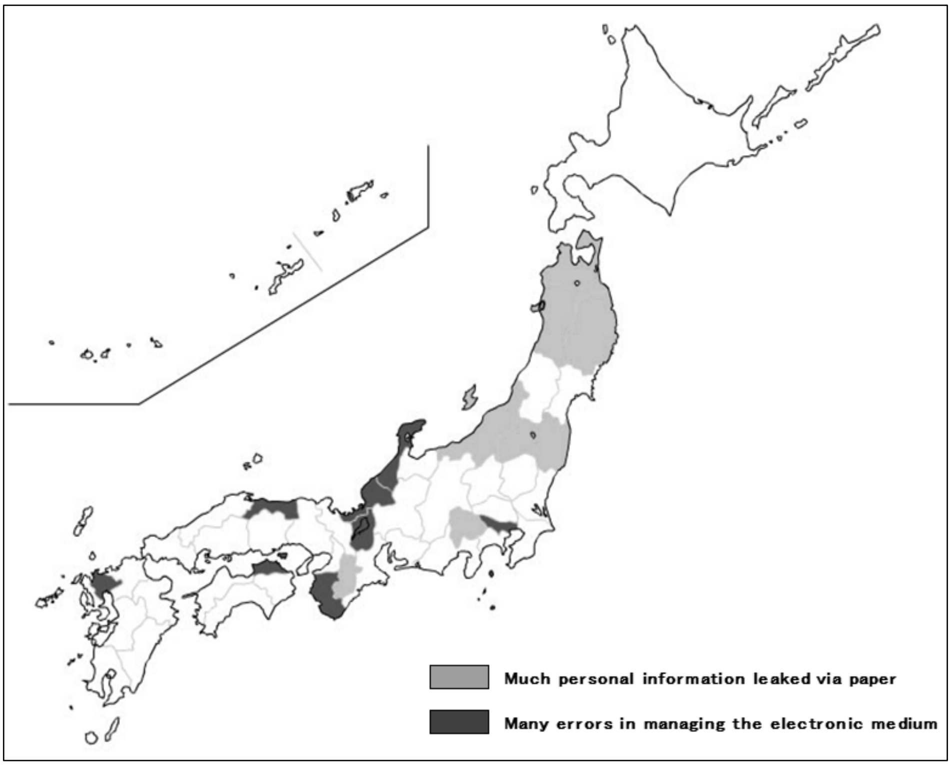

Much personal information leaked via paper
Many errors in managing the electronic medium

Figure 10: Japanese map based on cluster analysis for the cause of personal data leaks by city and district

5. Conclusion

We collected personal information leakage incident data reported on the Security Next website and performed a static statistical analysis. Then, a statistical analysis focusing on personal information leakage incidents by city and district was carried out. This type of analysis had previously been not carried out conventionally, at least not in Japan.

In future work, we will investigate the characteristics by city and district in more detail. In addition, we would like to apply the analyzed results to enact personal data leak countermeasures.

6. References

[1] Japan Network Security Association http://www.jnsa.org/

[2] H.Fumikura, R.Sasaki "Analysis on Personal Information Leakage from the Epidemiology Aspect", in Computer Security Symposium 2007, Information Processing Society of Japan

[3] H.Fumikura, T.Kobayashi, R.Sasaki "Statistical analysis and consideration on personal Information leakage and prefecture characteristics", in CSEC52/DPS146, Information Processing Society of Japan

[4] Security NEXT http://www.security-next.com

[5] Japan Epidemiological Association "the epidemiology - To learn from the basics -", Nankodo, 1996

[6] M.Walker "Way of thinking / how to lead epidemiologic studies", Shinko medical publishing company, 1996

[7] Ministry of Public Management, Home Affairs, Posts and Telecommunications Statistics Bureau statistics data http://www.stat.go.jp/data/index.htm

[8] I.Ishiguro, T.Kobayashi, M.Aida "Analysis of the social survey data by Stata", Kitaoji bookshop, 2008

[9] J.Yoshida, H.Hasegawa, M.Kasuga, "Selection of evaluation terms for video game contents using cluster analysis" ITE Technical Report Vol.32, No.21, PP.9-12

[10] T.Niimi, K.Imai, K.Kamegai, T.Hioki, A.Mano, "Classification of Observer-groups with Cluster Analysis and the Criteria of Visual Decision for Each Group", Japanese Society of Radiological Technology magazine, 2003