DEVELOPMENT OF SYNTHETIC MICRODATA FOR EDUCATIONAL USE IN JAPAN

MAKITA, Naoki[1], ITO, Shinsuke[1,2], HORIKAWA, Akiko[1],
GOTO, Takehiko[1], YAMAGUCHI, Kozo[3]
[1]National Statistics Center, Japan
[2]Meikai University, Japan
[3]Statistical Research and Training Institute Ministry of Internal Affairs and Communications, Japan

Contact email: ssitoh@meikai.ac.jp

ABSTRACT

Japan's new Statistics Act has come fully into effect in April 2009. The new law allows access to Anonymized microdata, and at the same time it requires users to go through an application process and imposes some restrictions. The National Statistics Center (NSTAC) has developed a type of microdata which can be accessed without an application process and used without restrictions. These data do not contain original microdata, but consist of Synthetic microdata. The absence of an application process and usage restrictions make Synthetic microdata particularly suitable for educational use.

This paper outlines the process for creating Synthetic microdata for educational use based on multi-dimensional tables derived from original microdata, and compares the characteristics of them.

1. BACKGROUND: THE LEGAL FRAMEWORK IN JAPAN

Japan's new Statistics Act has come fully into effect in April 2009, and allows the provision of Anonymized microdata (Article 36) and tailor-made tabulations (Article 34) for scientific purposes. Anonymized microdata are defined as "questionnaire information that is processed so that no particular individuals or juridical persons, or other organizations shall be identified." As Anonymized microdata are created using disclosure limitation methods and therefore different from the original microdata, they allow a wider use of official microdata including for higher education and academic research. The new Statistics Act has thus expanded the role of statistics in education and research in Japan.

The National Statistics Center (NSTAC) compiles statistical tables for the surveys conducted by Japanese Ministries and Agencies such as the Statistics Bureau of Japan (SBJ). In addition, based on the new Statistics Act and a Cabinet order, NSTAC today plays a key role in the new framework for statistics education and research by operating a data archive that provides

Anonymized microdata and tailor-made tabulations for data collected by government offices and ministries, and cooperates with academic research organizations to promote these services.

In order to access Anonymized microdata, the new Statistics Act requires users to apply for permission, imposes some conditions on data usage and storage, and requires payment of access costs of approximately 10,000 JPY (100 USD) per file. These factors make the data less attractive for education and training. To provide an alternative that offers easier access, the NSTAC has developed Synthetic microdata for education and training that do not stem from Anonymized microdata and which can be accessed without an application process as well as used without restrictions.

2. SYNTHETIC MICRODATA FOR EDUCATIONAL USE

This paper describes how these Synthetic microdata for educational use were created using multi-dimensional tabulation on original microdata from the 2004 'National Survey of Family Income and Expenditure' conducted by the Statistical Bureau of Japan. Since the Synthetic microdata created stem from the original microdata in an indirect way, they are free from the application process or restrictions that apply to Anonymised microdata. Specifically, the Synthetic microdata were created by using microaggregation, which is one of the disclosure limitation methods adopted for microdata of official statistics. Characteristics of microaggregation are (1) creation of records with common values for all types of qualitative attributes based on multi-dimensional tabulation and (2) sorting and dividing records with common values for qualitative attributes into groups larger than a specific minimum size.[1] In order for the Synthetic microdata to achieve distributions that approximately replicate those of the original microdata, a multivariate normal random number that replicates average, variance and co-variance of the original microdata was used based on the assumption that records are normally (or log-normally in the case of monetary amounts etc.) distributed within each cell of the multi-dimensional tables. The Synthetic microdata created in this research have about 30,000 records.

3. CREATING SYNTHETIC MICRODATA

The detailed process for the creation of the Synthetic microdata is as follows: First, quantitative and qualitative attributes to be contained in the Synthetic microdata were selected. Second, records with common values for qualitative attributes were sorted into groups with a minimum size of 3. Third, tables were created in order to generate multivariate lognormal random numbers and records for which values for some quantitative attributes are 0. This process allows creating Synthetic microdata with characteristics similar to those of the original microdata. The detailed process is as follows:

[1] For a more detailed explanation see Ito (2009) and Ito and Takano (2011).

(1) Qualitative attributes were selected from the multi-dimensional statistical tables compiled based on the original microdata. Specifically, 14 qualitative attributes were selected based on the survey items that are used most frequently by researchers, including gender, age and employment status. 184 quantitative attributes were selected, including Yearly Household Income and Monthly Household Expenditures.

(2) Records with common values for qualitative attributes were sorted into groups with a minimum size of 3. For records that have common values for some qualitative attributes and that refer to groups with a size of 1 or 2, values for the other qualitative attributes were transformed to 'unknown' (V) in order to create groups with a minimum size of 3. Figure 1 illustrates this process in the case of gender and employment status.

(3) Two types of tables were created in order to generate 1) multivariate lognormal random numbers and 2) records where values for some quantitative attributes are 0. Tables of 'Type 1' contain frequency, mean, variance and covariance of quantitative attributes not including 0. The records on which these tables are based were classified by qualitative attributes in order to generate multivariate lognormal random numbers. Tables of 'Type 2' are tables created by sorting records based on whether values for quantitative attributes are 0 or not 0, and on this basis the values for some quantitative attributes in the records were transformed to 0.

Figure 2 illustrates the creation of the Synthetic microdata and compares the frequency of the Synthetic microdata with that of the original microdata. To create the Synthetic microdata, logarithmic transformation was used for the original microdata. Then the above two types of tables were used to generate multivariate lognormal random numbers and transform the values for some quantitative attributes to 0. Lastly, exponential transformation was conducted.

4. COMPARISON OF ORIGINAL MICRODATA AND SYNTHETIC MICRODATA

To establish the usability of the Synthetic microdata, their characteristics were compared to the original Microdata. Table 1 presents the comparison of average values between the two microdata for several quantitative attributes such as Yearly Income, Receipts, Income, Receipts other than Income, Expenditure, Living Expenditure and Non-Living Expenditure. The difference between the two microdata was calculated by deducting the averages of the attributes of the original microdata from that of the Synthetic microdata, and dividing the deduction by the average of the attributes of the original microdata. The results show that the averages of the attributes contained in the Synthetic microdata are quite similar to those in the original microdata.

Table 2 presents a comparison of standard deviation for several quantitative attributes of the original microdata and the Synthetic microdata. The difference between the two was calculated by deducting the standard deviation of the attributes of the original microdata from that of the Synthetic microdata, and dividing the deduction by the standard deviation of attributes of the original microdata. The results show that the standard deviation for the Synthetic microdata is

similar to that for the original microdata.

Figure 3 shows histograms of ‘Receipts other than Income’ for the Synthetic microdata and the original microdata. The histograms of ‘Receipts other than Income’ are similar to each other.

Figure 4 are scatter diagrams of Yearly Income and Non-Living Expenditure for both the Synthetic microdata and the original microdata. The scatter diagrams of Yearly Income and Non-Living Expenditure resemble each other, although the Synthetic microdata have more outliers than the original microdata because of the influence of multivariate lognormal random numbers on the frequency of the Synthetic microdata.

Table 3 contains correlation matrices calculated for several attributes of records contained in the original microdata and the Synthetic microdata. By and large the correlation matrices appear similar. Therefore, the results show that the relationship between attributes of the two microdata is maintained.

5. CONCLUSION

This paper focuses on the characteristics of Synthetic microdata created by NSTAC for statistics education and training. The new Statistics Act allows the provision of Anonymized microdata, but at the same time requires users to go through an application process and imposes some restrictions, and as a result makes Anonymized microdata less attractive for use in statistics education and training.

The Synthetic microdata created by NSTAC are a suitable alternative as they are available without application process or restrictions. This paper details the creation of these Synthetic microdata with characteristics similar to those of the original microdata.

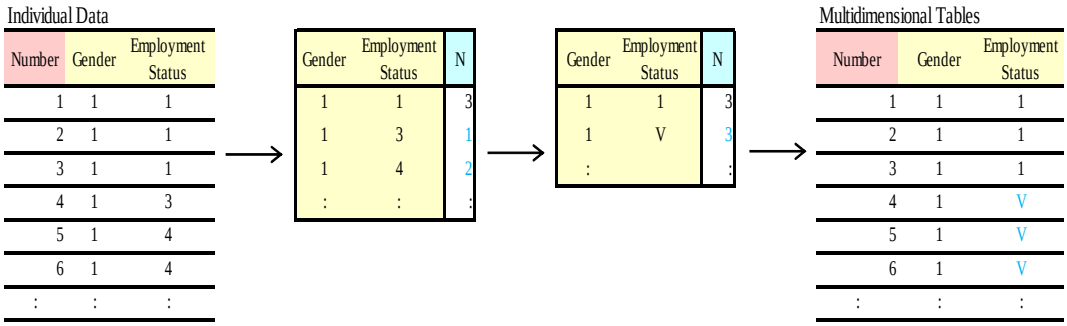
These Synthetic microdata are particularly suitable for educational purposes as they allow teachers to demonstrate the creation of graphs and tables, and students to practice regression analysis and the calculation of correlation coefficients without restrictions in the use of the data. These Synthetic microdata provide students with increased opportunities of handling actual microdata, and therefore increase their skills and practical experience. As a result, these Synthetic microdata have significant potential to enhance and expand statistics education and training to complement Anonymised microdata.

REFERENCES

Ito, S. (2009) “On Microaggregation as Disclosure limitation methods”, *Journal of Economics, Kumamoto Gakuen University*, Vol.15, No.3・4, pp.197-232 (in Japanese).

Ito, S. and Takano, M. (2011) “A Method to Quantitatively Assess Confidentiality and Potential Usage of Official Microdata in Japan”, Paper presented at the 58th World Statistics Congress of the International Statistical Institute at the Convention Centre Dublin.

APPENDIX



Note: "V" stands for "unknown".

Figure 1: Processing records with common values for qualitative attributes into groups with a minimum size of 3

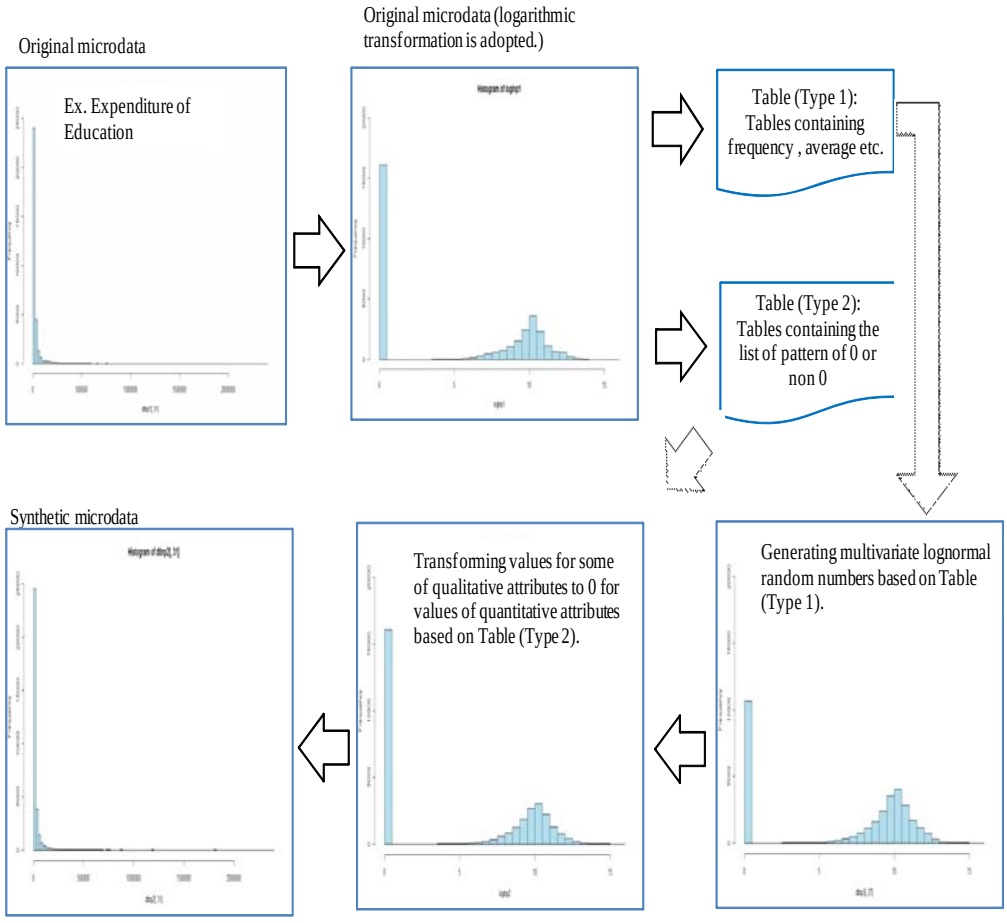


Figure 2: Creation of Synthetic microdata

Table 1: Average Values for Original Microdata and Synthetic Microdata

	Original Microdata	Synthetic Microdata	Difference
Yearly Income(ten thousand yen)	740	730	-0.01
Receipts(yen)	971,789	946,779	-0.03
Income(yen)	502,134	497,656	-0.01
Receipts other than Income(yen)	391,824	372,130	-0.05
Carry-over from previous month(yen)	77,832	76,993	-0.01
Disbursements(yen)	971,789	946,779	-0.03
Expenditure(yen)	415,809	403,747	-0.03
Living expenditure(yen)	339,199	328,140	-0.03
Food(yen)	73,739	72,883	-0.01
Housing(yen)	19,388	17,687	-0.09
Fuel, light and water charges(yen)	19,395	19,238	-0.01
Furniture and household utensils(yen)	9,784	9,204	-0.06
Clothes and footwear(yen)	14,649	14,138	-0.03
Medical care(yen)	11,936	11,366	-0.05
Transportation and communication(yen)	50,741	47,961	-0.05
Education(yen)	22,332	22,270	0.00
Reading and recreation(yen)	32,473	31,389	-0.03
Other living expenditure(yen)	84,762	82,003	-0.03
Non-living expenditure(yen)	76,610	75,607	-0.01
Disbursements other than expenditure(yen)	475,948	464,318	-0.02
Carry-over to month(yen)	80,032	78,714	-0.02

Note: “Difference” between the original microdata and the Synthetic microdata was calculated by deducting the averages of the attributes of the original microdata from that of the Synthetic microdata, and dividing the deduction by the average of the attributes of the original microdata.

Table 2: Standard Deviation for Original Microdata and Synthetic Microdata

	Original Microdata	Synthetic Microdata	Difference
Yearly Income(ten thousand yen)	358	338	-0.06
Receipts(yen)	541,291	473,481	-0.13
Income(yen)	280,696	261,558	-0.07
Receipts other than Income(yen)	353,922	263,446	-0.26
Carry-over from previous month(yen)	87,036	98,947	0.14
Disbursements(yen)	541,291	473,481	-0.13
Expenditure(yen)	224,420	219,291	-0.02
Living expenditure(yen)	194,501	192,447	-0.01
Food(yen)	30,149	28,064	-0.07
Housing(yen)	52,962	60,587	0.14
Fuel, light and water charges(yen)	8,009	7,690	-0.04
Furniture and household utensils(yen)	15,978	14,933	-0.07
Clothes and footwear(yen)	18,837	19,823	0.05
Medical care(yen)	19,763	19,284	-0.02
Transportation and communication(yen)	85,022	84,654	0.00
Education(yen)	51,990	64,157	0.23
Reading and recreation(yen)	32,162	32,723	0.02
Other living expenditure(yen)	95,899	102,041	0.06
Non-living expenditure(yen)	56,200	66,378	0.18
Disbursements other than expenditure(yen)	394,805	334,227	-0.15
Carry-over to month(yen)	96,421	118,056	0.22

Note: “Difference” between the original microdata and the Synthetic microdata was calculated by deducting the standard deviation of the attributes of the original microdata from that of the Synthetic microdata, and dividing the deduction by the standard deviation of attributes of the original microdata.

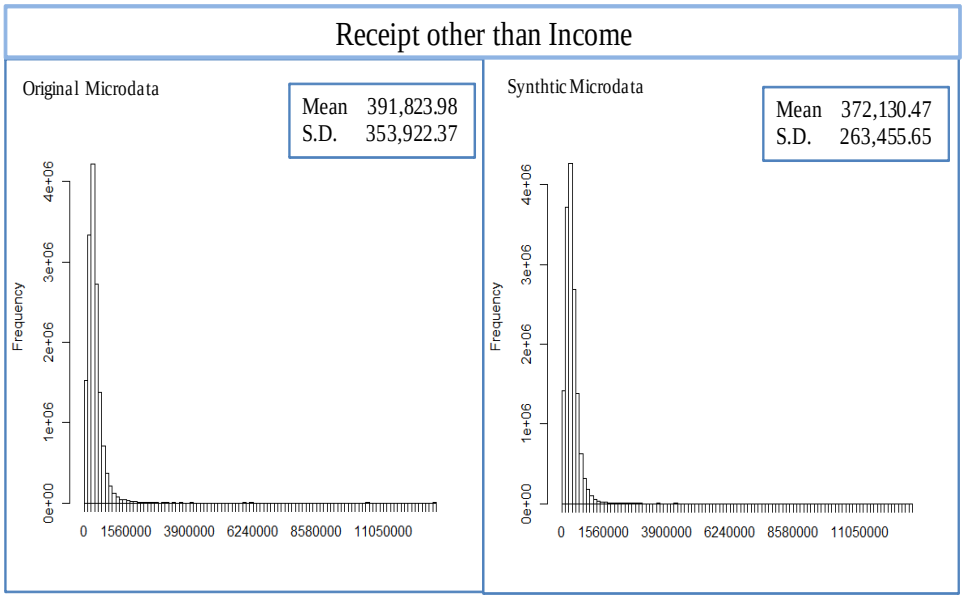


Figure 3: Histograms for Original Microdata and Synthetic Microdata

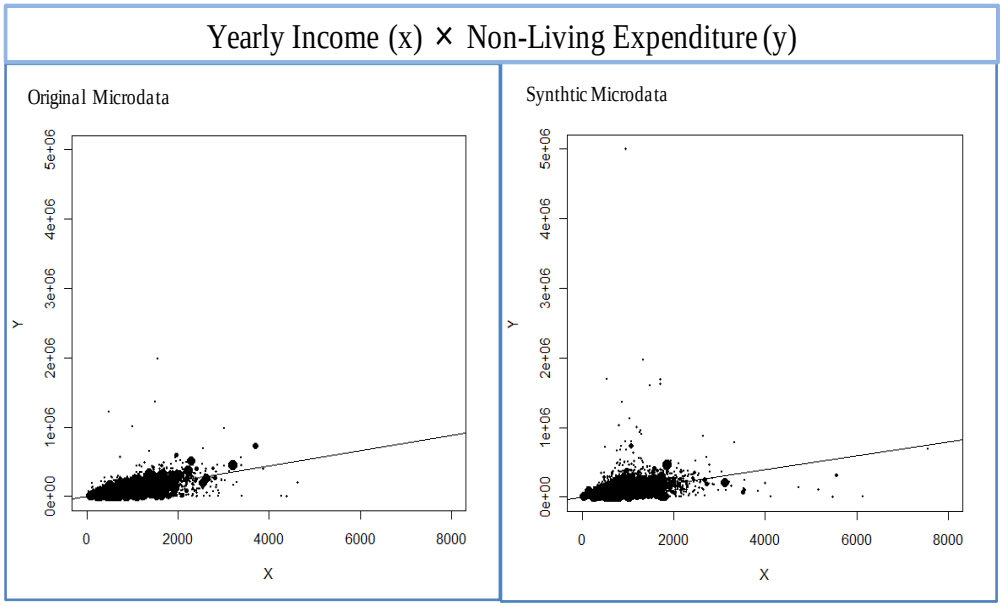


Figure 4: Scatter Diagrams for Original Microdata and Synthetic Microdata

Table 3 Correlation Coefficient for Original Microdata and Synthetic Microdata

Original Microdata

	Yearly Income	Receipts	Income	Receipts other than Income	Carry-over from previous month	Disbursem-ents	Expenditure	Living expenditure	Food	Housing	Fuel, light and water charges	Furniture and household utensils	Clothes and footwear	Medical care	Transport-ation and communicat-ion	Education	Reading and recreation	Other living expenditure	Non-living expenditure	Disbursem-ents other than expenditure	Carry-over to month
Yearly Income	1.00																				
Receipts	0.60	1.00																			
Income	0.66	0.78	1.00																		
Receipts other than Income	0.35	0.85	0.36	1.00																	
Carry-over from previous month	0.19	0.26	0.14	0.04	1.00																
Disbursements	0.60	1.00	0.78	0.85	0.26	1.00															
Expenditure	0.60	0.73	0.56	0.63	0.17	0.73	1.00														
Living expenditure	0.49	0.66	0.45	0.61	0.16	0.66	0.97	1.00													
Food	0.47	0.42	0.37	0.31	0.17	0.42	0.52	0.50	1.00												
Housing	-0.02	0.11	0.00	0.16	0.01	0.11	0.24	0.28	-0.03	1.00											
Fuel, light and water charges	0.32	0.24	0.22	0.16	0.11	0.24	0.28	0.27	0.44	-0.07	1.00										
Furniture and household utensils	0.15	0.25	0.12	0.26	0.09	0.25	0.26	0.27	0.17	0.07	0.10	1.00									
Clothes and footwear	0.30	0.30	0.24	0.24	0.10	0.30	0.39	0.38	0.29	0.02	0.12	0.16	1.00								
Medical care	0.11	0.16	0.10	0.15	0.07	0.16	0.24	0.25	0.15	0.01	0.07	0.08	0.09	1.00							
Transportation and communication	0.14	0.33	0.15	0.37	0.04	0.33	0.54	0.57	0.12	0.01	0.05	0.05	0.10	0.06	1.00						
Education	0.18	0.23	0.15	0.23	0.03	0.23	0.37	0.39	0.24	-0.03	0.19	0.02	0.09	0.04	0.07	1.00					
Reading and recreation	0.32	0.35	0.27	0.30	0.12	0.35	0.44	0.42	0.32	0.02	0.10	0.15	0.26	0.10	0.10	0.09	1.00				
Other living expenditure	0.39	0.46	0.38	0.37	0.12	0.46	0.66	0.66	0.21	0.01	0.13	0.12	0.19	0.11	0.12	0.04	0.16	1.00			
Non-living expenditure	0.70	0.63	0.70	0.38	0.12	0.63	0.62	0.43	0.35	-0.02	0.19	0.12	0.26	0.08	0.17	0.14	0.29	0.34	1.00		
Disbursements other than expenditure	0.44	0.90	0.72	0.79	0.04	0.90	0.40	0.32	0.25	0.01	0.14	0.18	0.17	0.08	0.14	0.11	0.22	0.23	0.49	1.00	
Carry-over to month	0.16	0.24	0.13	0.06	0.86	0.24	0.13	0.12	0.13	0.02	0.10	0.07	0.07	0.05	0.02	0.02	0.08	0.10	0.10	0.01	1.00

Table 3 (Cont'd)

Synthetic microdata

	Yearly Income	Receipts	Income	Receipts other than Income	Carry-over from previous month	Disbursem-ents	Expenditure	Living expenditure	Food	Housing	Fuel, light and water charges	Furniture and household utensils	Clothes and footwear	Medical care	Transport-ation and communicat-ion	Education	Reading and recreation	Other living expenditure	Non-living expenditure	Disbursem-ents other than expenditure	Carry-over to month
Yearly Income	1.00																				
Receipts	0.58	1.00																			
Income	0.63	0.85	1.00																		
Receipts other than Income	0.38	0.83	0.48	1.00																	
Carry-over from previous month	0.12	0.32	0.15	0.05	1.00																
Disbursements	0.58	1.00	0.85	0.83	0.32	1.00															
Expenditure	0.52	0.71	0.59	0.64	0.14	0.71	1.00														
Living expenditure	0.42	0.63	0.49	0.60	0.14	0.63	0.96	1.00													
Food	0.46	0.40	0.36	0.32	0.13	0.40	0.45	0.43	1.00												
Housing	-0.05	0.08	0.04	0.09	0.03	0.08	0.24	0.28	-0.06	1.00											
Fuel, light and water charges	0.32	0.25	0.23	0.18	0.09	0.25	0.26	0.25	0.44	-0.07	1.00										
Furniture and household utensils	0.12	0.15	0.11	0.14	0.04	0.15	0.19	0.19	0.15	0.00	0.10	1.00									
Clothes and footwear	0.21	0.23	0.19	0.20	0.06	0.23	0.29	0.28	0.20	0.01	0.08	0.12	1.00								
Medical care	0.07	0.13	0.09	0.13	0.04	0.13	0.19	0.20	0.11	0.00	0.06	0.05	0.05	1.00							
Transportation and communication	0.12	0.30	0.17	0.35	0.04	0.30	0.50	0.54	0.10	-0.01	0.05	0.03	0.06	0.04	1.00						
Education	0.14	0.24	0.18	0.24	0.02	0.24	0.38	0.41	0.18	-0.02	0.16	0.01	0.04	0.02	0.04	1.00					
Reading and recreation	0.26	0.30	0.24	0.28	0.06	0.30	0.35	0.34	0.26	-0.01	0.06	0.12	0.18	0.07	0.07	0.05	1.00				
Other living expenditure	0.33	0.44	0.38	0.37	0.11	0.44	0.63	0.65	0.17	-0.02	0.11	0.07	0.11	0.06	0.09	0.04	0.10	1.00			
Non-living expenditure	0.50	0.50	0.52	0.35	0.07	0.50	0.53	0.26	0.24	-0.04	0.14	0.07	0.14	0.05	0.09	0.07	0.18	0.21	1.00		
Disbursements other than expenditure	0.45	0.85	0.77	0.74	0.07	0.85	0.32	0.25	0.25	-0.05	0.15	0.08	0.13	0.05	0.09	0.09	0.18	0.18	0.35	1.00	
Carry-over to month	0.10	0.28	0.14	0.05	0.82	0.28	0.07	0.07	0.09	0.00	0.08	0.03	0.03	0.02	0.01	0.00	0.02	0.06	0.04	0.00	1.00