MEAT & LIVESTOCK AUSTRALIA

UNIVERSITY OF TECHNOLOGY SYDNEY

Unlocking the value of data that currently exists and is accessible by MLA across the red meat industry value chain

Yi Zheng

07-03-2018

University of Technology Sydney

Contents

1. Project Background

1.1 Introduction

MLA is leading the development of a Digital Value Chain strategy with industry to enable the seamless capture, integration and interpretation of the vast and increasing range of data that's being generated through new and existing technology in the red meat industry. Maximizing information exchange will ensure the red meat industry produces what our markets need more sustainably and profitably. Improved communication will also increase the capacity of all industry players to embrace new technology – and the use of meaningful data in their own business. But to achieve this we need collaboration across industry and with the world's best innovation companies. Producers, processors, logistics companies, retailers and consumers are faced daily with a plethora of concepts and solutions that all fit within the auspices of a digital strategy. The feedback to MLA from stakeholders across the industry is they don't know:

(a) what is real?
(b) how it can be used?
(c) is it of value to them?
(d) which options do I choose?
(e) how do I use it within my business?
(f) who owns the data?
(g) what do I do with the data and do I need new analytical skills?
(h) will I be better off or is this just a passing fad?

This project will focus on conducting background research to address the questions above. There already exists a huge amount of information in our industry and a key outcome will be to identify all the data, who owns it and where it can add value. Current data sources include:

(a) Eating quality
(b) Carcase compliance (slaughtered livestock)
(c) Livestock movements and location history
(d) Animal health information
(e) Genetics

An analytical skillset is required with managing multiple, large datasets as well as an innovative approach to unlocking insights and value in data i.e., machine learning, data mining, advanced statistical analysis, and etc.

1.2 Research to be conducted

(a) Identify the core datasets, along with Master Data & Quality strategies to bring those datasets together for analysis.
(b) Identify any missing attributes that may need to be added to the datasets to achieve the project outcomes.
(c) Develop innovative models and analytics that demonstrate how the datasets can create additional value that does not currently exist.

2. Research Method

2.1 Statistical Methods

2.1.1. Descriptive Statistics

Descriptive statistics are summary statistics quantitatively describe or summarize features of a collection of information/data [1]. The aim of descriptive statistics is to get a basic understanding of the distribution of data, e.g., the mean, maximum, frequency, standard deviation and etc. Only when we get the distribution of the data, can we develop better and more suitable models to analyze the data.

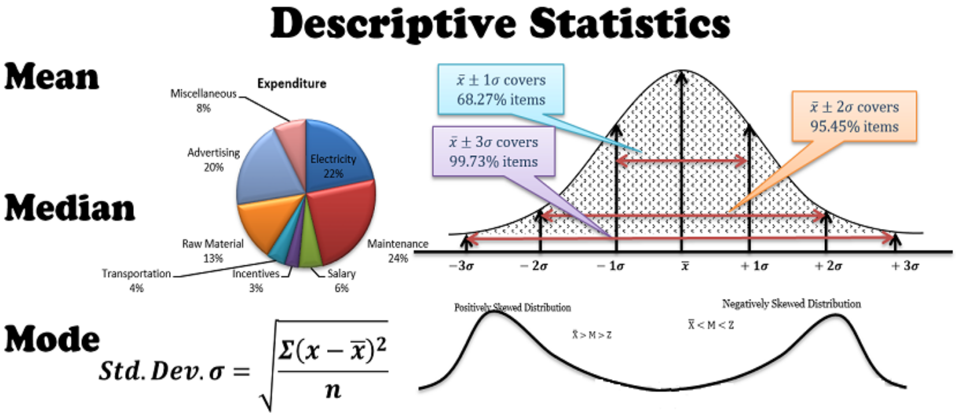


Figure 1. Schematic diagram of descriptive statistics

2.1.2. Variable Combination

We combine several variables together to show the trends/patterns one variable changes with other variables. We use two examples to illustrate the method. The first example is to tally the number of normal transfers from 1999 to 2017 (Figure 2). In this example, the normal transfer number is combined with the time (year). Then the trend which shows how the normal transfer number changes with time is drawn. The second example is to tally the number of historical missing transfers in every state. The missing transfer number and the state are combined, and the percentages of missing transfer among states are drawn.

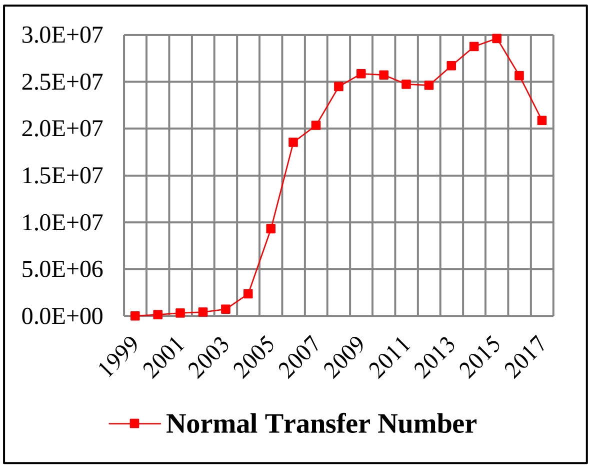


Figure 2. Trend of normal transfer from 1999 to2017

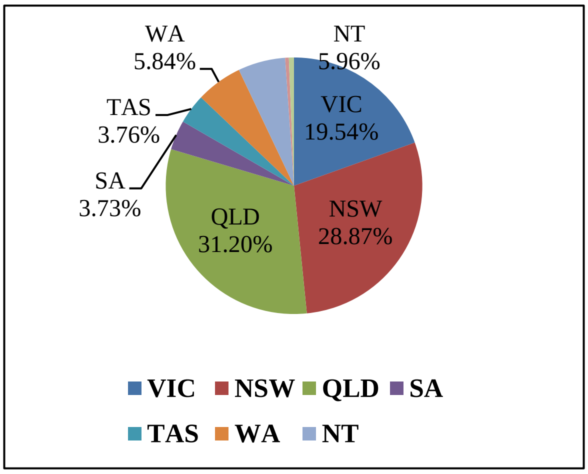


Figure 3. Missing transfer distribution among states

2.2 Simple Linear Regression

The simple linear regression is used in the analysis of the second business scenario, i.e., identification of PICs which show continually growing trends in current holdings in recent years. It concerns two-dimensional sample points with one independent variable and one dependent variable. Its goal is to find a linear function which can predict the dependent variable values as accurately as possible by using the independent variable as input [2]. A linear regression line has an equation of the form Y = a + b*X, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0). We employ the method least-square to fit the regression line. It calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line.

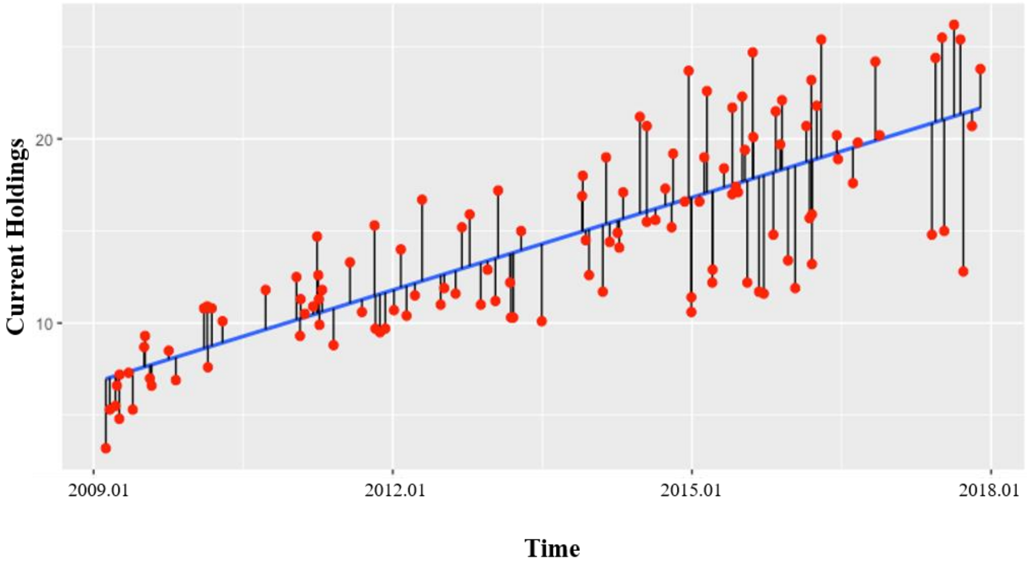


Figure 4. Example of simple linear regression between current holdings and time

2.3 Geographic Information System (GIS) Map Visualization

We visualize the distribution of transfers among postcode areas on the GIS map, so that the hotspots and patterns can be identified easily. GIS map provides direct visualization of the analysis results, even users without domain knowledge can get the key points quickly. The foundation API of the GIS map is supported by Leaflet and OpenstreetMap [3].

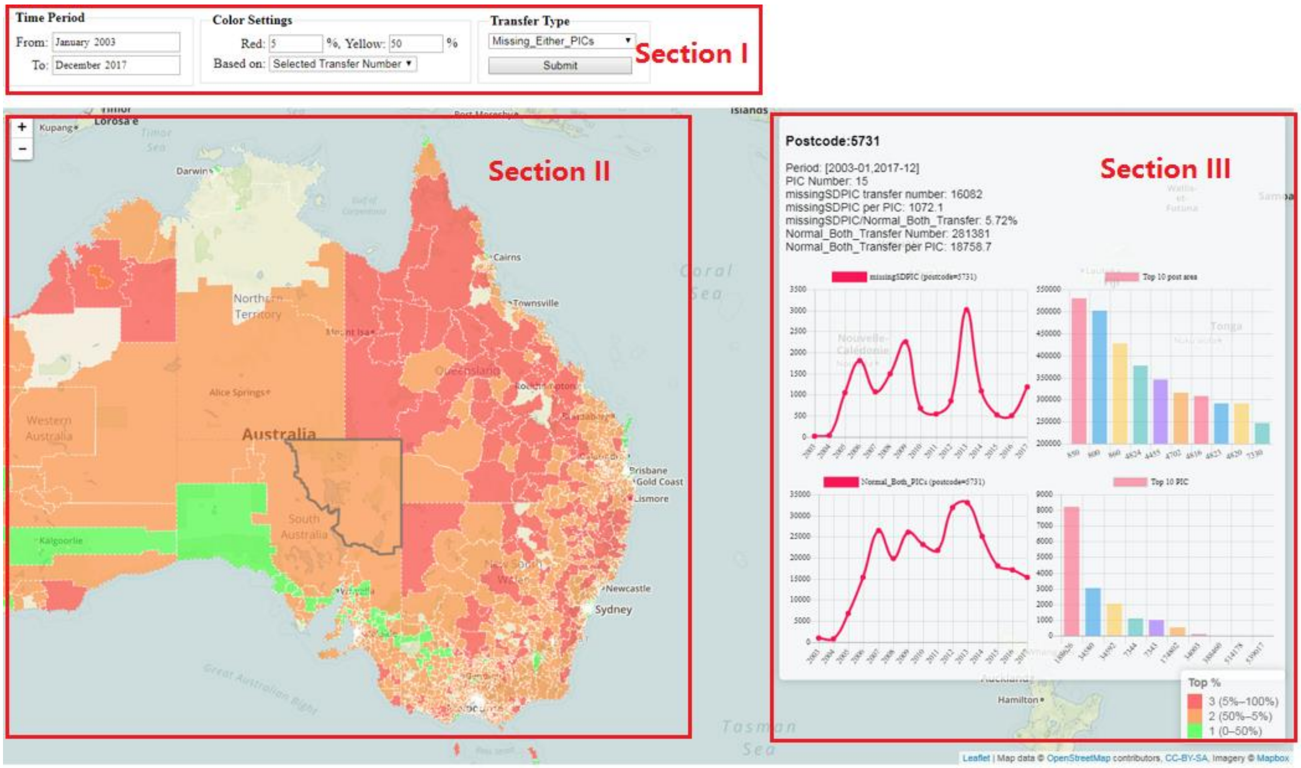
To implement the visualization, we first downloaded boundary coordinates of postcode areas in Australia from the Australian Bureau of Statistics (shape file). Then we converted the shape file into KML format using Google Earth. Following that, we extracted the boundary coordinates from the KML file and stored them into GeoJson file. Next, we coded using Leaflet API to map information of corresponding postcode areas to their coordinates. The postcode information is stored in the table PICDetail from the database NLISMirror. 561,771 PIC-postcode pairs are mapped successfully, while 4,582 PICS doesn’t have records in PICDetails.

The map is consisted of three sections, namely the setting section (Section I), the map section (Section II), and the information display section (Section III).

In the setting section (Section I), we can set the time period we want to visualize simply by choosing the start month&year, and the end month&year. We can also set the base to colour the map (by transfer number or transfer ratio of the selected transfer type) and the segmentation percentages. The postcode areas are ranked in descending order based on the selected transfer number/ratio first. Then the top X% (e.g., 5%) areas are coloured red, areas between (Y%, X%) are coloured yellow, and the remaining areas are coloured green. Besides, selection of the type of transfers to be visualized is provided via the ComboBox.

Section II shows the whole map of Australia. The postcode areas are coloured red, yellow and green according to their ranks. Those postcode areas which don't have data are of raw colour of the map. We can zoom in and zoom out with our mouse wheel or clicking the "+/-" buttons in the top left corner. When hovering over the map, the corresponding information of the located postcode area will be displayed in Section III.

Section III illustrates the information of the selected postcode area. The information includes the postcode, the visualization time period, the PIC number, the missing transfer number, the missing transfer number per PIC, the missing transfer ratio, the normal transfer number, the normal transfer number per PIC. Apart from the above information, two line charts which show the trends of the selected transfer and normal transfer, and two bar charts illustrate the top 10 post areas and top 10 PICs are provided.



Time Period
From: January 2003
To: December 2017

Color Settings
Red: 5 %, Yellow: 50 %
Based on: Selected Transfer Number

Transfer Type
Missing_Either_PICs

Postcode:5731

Period: [2003-01,2017-12]
PIC Number: 15
missingSDPIC transfer number: 16082
missingSDPIC per PIC: 1072.1
missingSDPIC/Normal_Both_Transfer: 5.72%
Normal_Both_Transfer Number: 281381
Normal_Both_Transfer per PIC: 18758.7

Figure 5. GIS visualization of transfers. Section I shows the map settings, Section II shows the colored map, and Section III shows information of the selected area.

2.4 One-Class Support Vector Machine (SVM)

One-class SVM is an unsupervised algorithm that learns a decision function for novelty detection: classifying new data as similar or different to the training set [4]. It is trained on data that has only one class, which is the "normal" class. And it has been demonstrated to be useful in anomaly detection [5-6]. Thus, we employ it to identify abnormal killing patterns of abattoirs. We didn't provide the theory and reasoning of one-class SVM in this report due to its complexity. Please find them from literature [4].

To identify the abnormal kill patterns, we first separated the kill data into weeks for each abattoir. Then we selected 200 weekly kill number sequences which belong to the normal killing pattern, as the training set. Next, we trained the one-class SVM with the selected training set. Finally, we tested the sequence of the killed cattle number across each week for each abattoir and identify abattoirs which have anomalous sequences. We also visualized those weeks of abnormal patterns using line chart.

3. Intern's Contribution

In this internship project, I combined different datasets for analysis, built models to tackle the real-world problems, and developed different visualizations to present analysis results. My contributions are detailed as follows:

(1) Explored two large-scale databases, namely "NLISMirror" and "NLISDW", obtained their structures and got a basic understanding of all tables. Extracted, cleaned and integrated different datasets from the two databases for analysis.

(2) Built models to analyse the above processed datasets, mined hidden trends, patterns and knowledge from them and finally derived insights from the analysis results.

(3) Developed different visualizations, i.e., figures, tables, GIS map, to present analysis results.

(4) Drafted project documents to report the project background, research methods, research impacts, and research results.

4. Research Results and Outcomes

4.1 Summary of NLISMirror

4.1.1. 31 tables in total, 261.2 GB allocated

(1) Transfer related: Transfer, PICDetail, Property, PICERPStatus

(2) Carcase: CarcaseID, CarcaseHeader, CarcaseBodyInfo, CarcaseSideInfo, CarcaseMeasurements, CarcaseValue

(3) Device: Device, DeviceStatus

(4) Kill: Kill, ManualKillUpload, MobBasedKill

(5) NVD: NVDAgent, NVDNumber, NVDToPIC, NVDOtherFromPIC, NVDPostBreeder

(6) Upload: Upload, ManualKillUpload, lkpUploadStatus

(7) Error: ErrorLog, lkpError

(8) Lkp: lkpStatus, lkpError, lkpTransactionType, lkpLossOfLTReason, lkpUploadStatus

(9) Others: SightedCattle, IndexAudit, EstablishmentNumberPICLink

4.1.2. UML Class Diagram

See additional file 1.

4.2 Analysis on Cattle Movements/Transfers

4.2.1 Definition

(1) PIC: Property Identification Codes, e.g., Farm/Abattoir/Saleyard

(2) Transfers: records of an animal transfer from one PIC to another

PIC1
(e.g.,35687)
PIC2
(e.g., 542)

Figure 6. Transfer from PIC1 to PIC2

(3) Missing Transfer

Transfer records whose Source PIC or Destination PIC is 8X (i.e., PICID=192596). In this project, we focus on those missing transfers whose source PIC is missing.

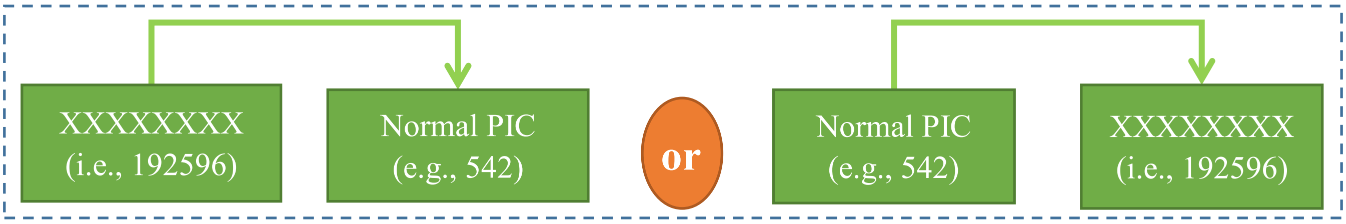


Figure 7. Examples of missing transfers

(4) Normal Transfer

Transfer records whose Source PIC and Destination PIC both are not 8X.

Table 1. Examples of missing transfer and normal transfer

Category	TransferID	TagID	SourcePICID	DestinationPICID	TransferDate	…
Missing Transfer	1	915093	192596 (PIC=8X)	131851	2000-01-27	…
	2	936819	364330	192596 (PIC=8X)	2000-02-11	…
Normal Transfer	3	933214	542	605	2017-08-11	…

4.2.2 Findings

(1) Summary of Transfer

A. Total transfers: 337,614,731

B. Missing transfer

(a) SourcePIC=XXXXXXXX: 14,228,611 records, 184,211 distinct Destination PICs

(b) DestinationPIC=XXXXXXXX: 14,228,629 records, 160,658 distinct Source PICs

(c) SourcePIC=XXXXXXXX or DestinationPIC=XXXXXXXX: 28,457,237 records, 210,445 distinct PICs

C. Normal transfer

(a) Records: 309,157,494 records,

(b) Distinct PICs: 247,857

(c) Distinct Source PICs: 232,692

(d) Distinct Destination PICs: 198,026

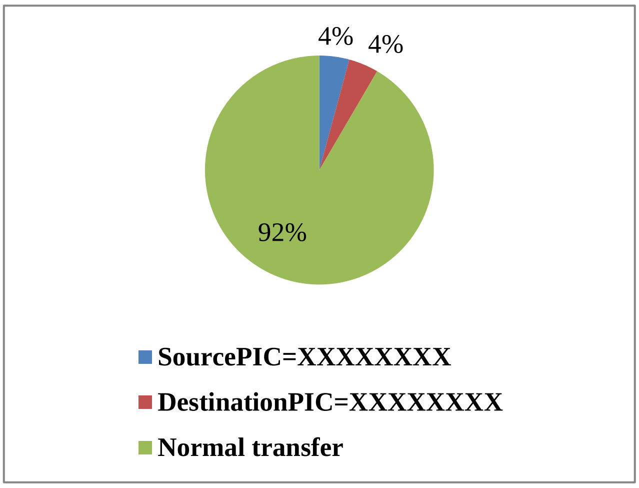


Figure 8. Percentage of missing transfers and normal transfers

(2) Missing transfers and normal transfers share similar trends with time.

Both the normal transfers and missing transfers have similar trends, increasing rapidly from 2003 to 2008 and decreasing dramatically from 2015/2008 (Figure 9 and Figure 10). It indicates the more transfers made, the more missing transfers occurred.

(3) Transfer peak season is autumn, while slack season is summer.

It seems the peak season of both normal transfers and missing transfers is autumn (from March to May), while the slack season is summer (from December to February, see Table 2, Table 3, Figure 11 and Figure 12).

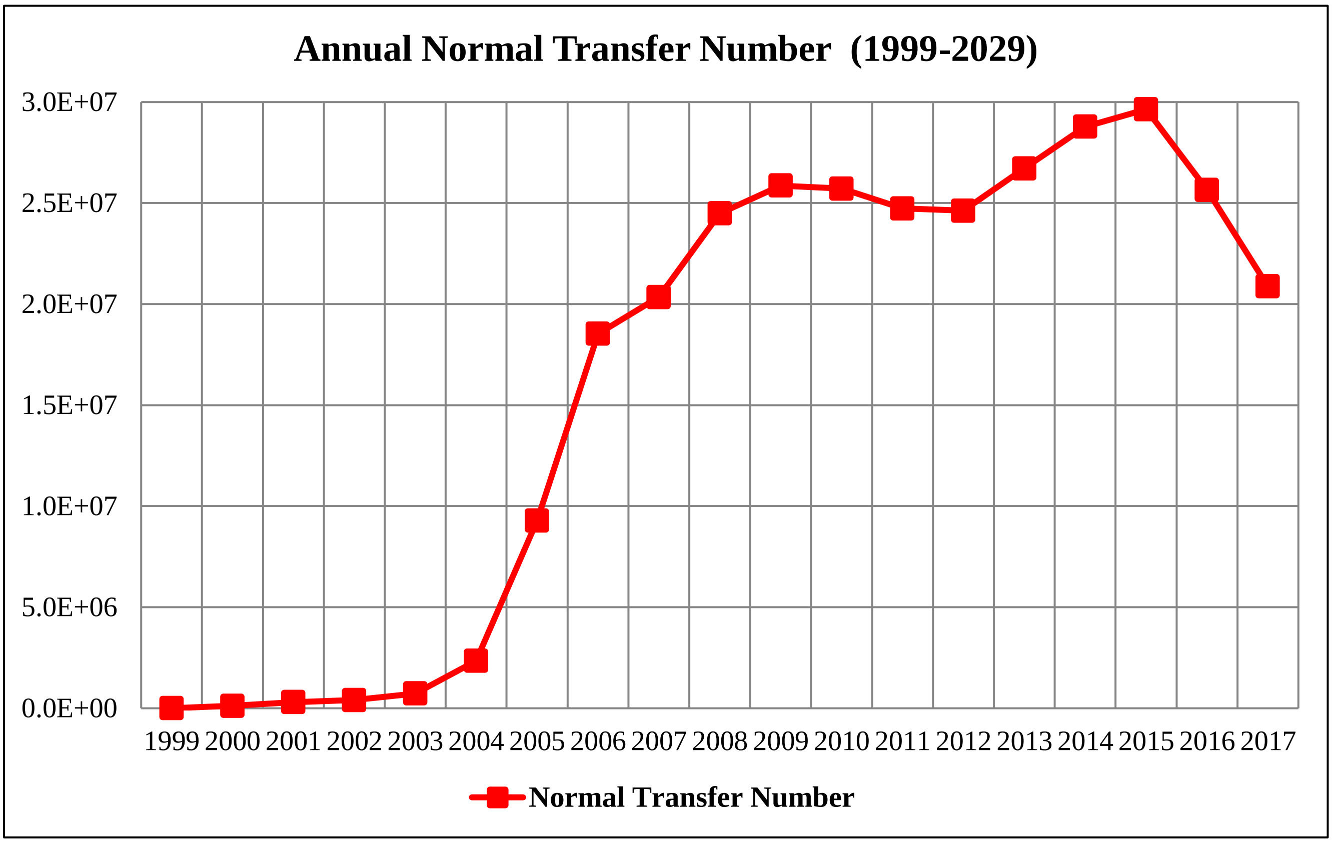


Figure 9. Annual Normal Transfer Number from 1999 to 2017

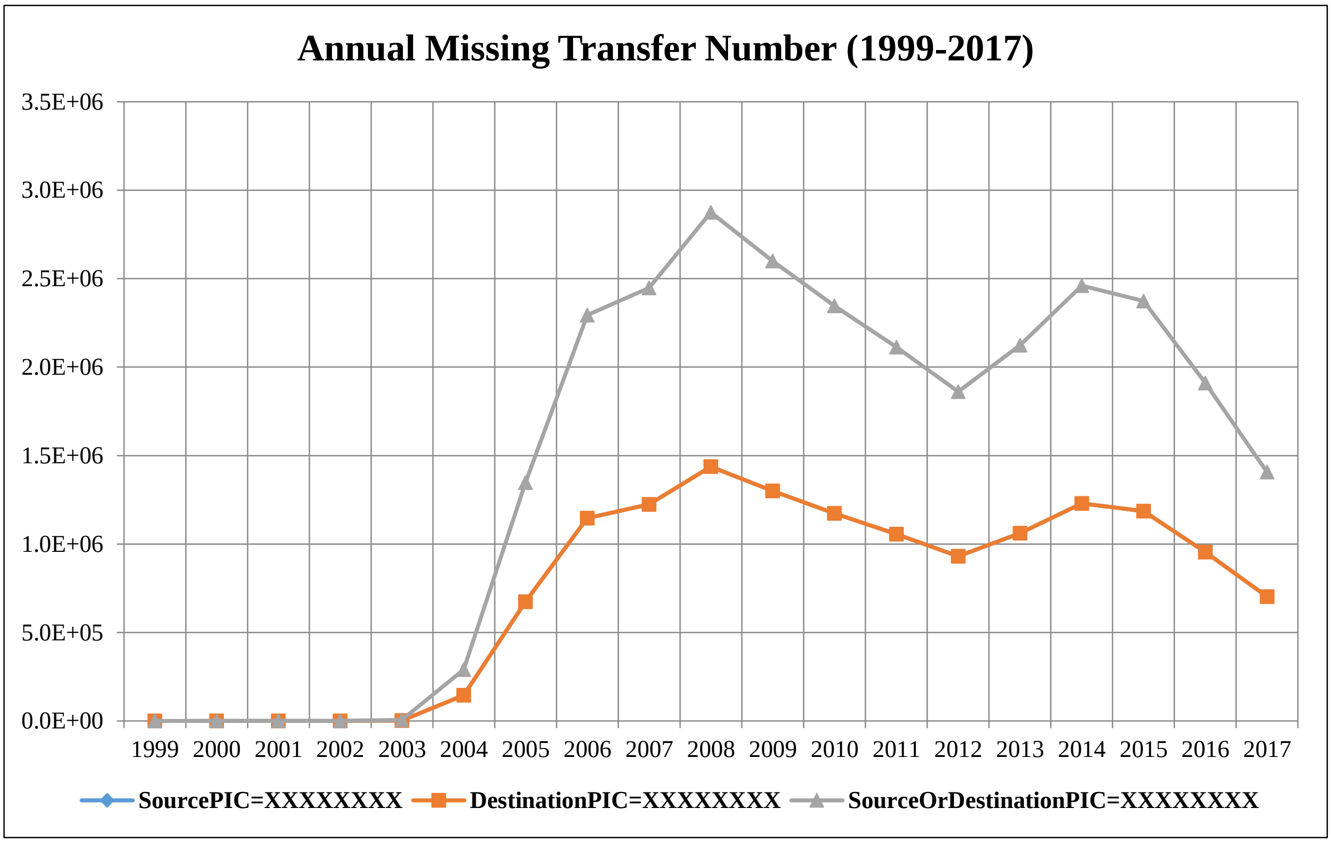


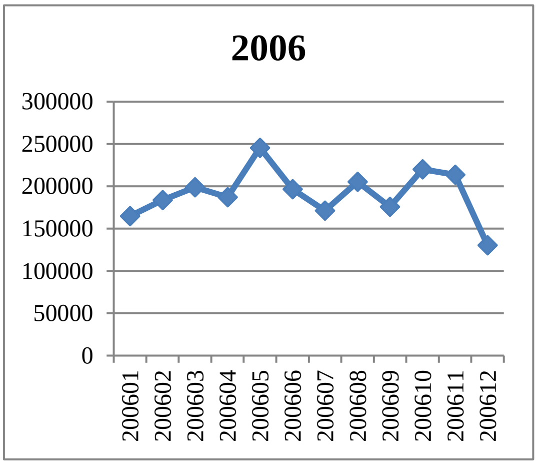
Figure 10. Annual Missing Transfer Number from 1999 to 2017

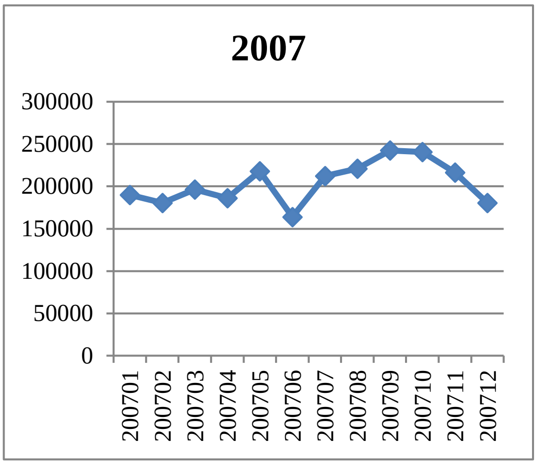
Table 2. Season distribution of missing transfers

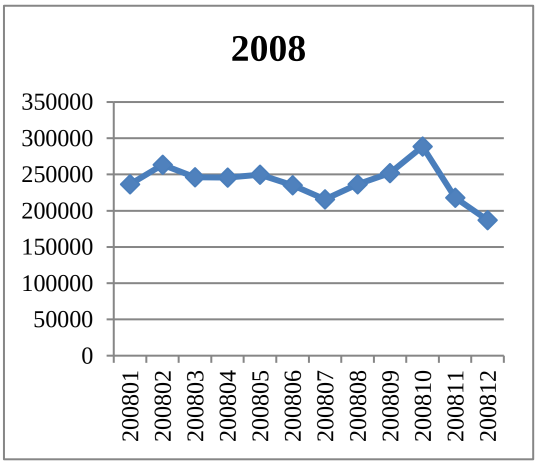
	autumn	summer	winter	spring
2004	26970	7510	90796	131851
2005	137442	129756	364330	582294
2006	631941	516904	573256	609500
2007	600342	500448	597170	699704
2008	741772	680370	687634	758666
2009	715808	570358	674460	658520
2010	673788	516290	602508	571623
2011	569266	509382	492170	554570
2012	500780	428350	480034	470466
2013	562566	421518	537370	549918
2014	646328	540778	624352	665992
2015	698220	564512	586496	543884
2016	550152	463364	463954	447992
2017	441602	372348	430790	282686

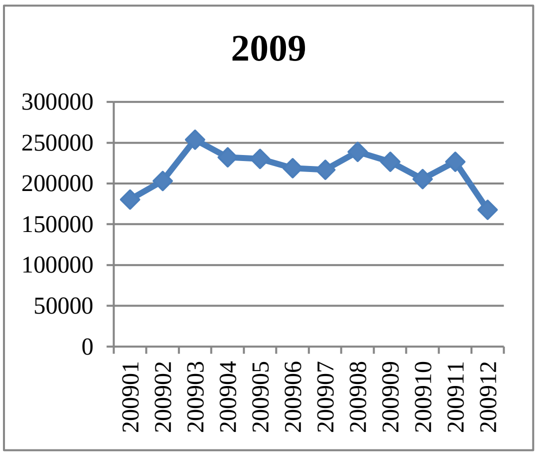
Table 3. Season distribution of normal transfers

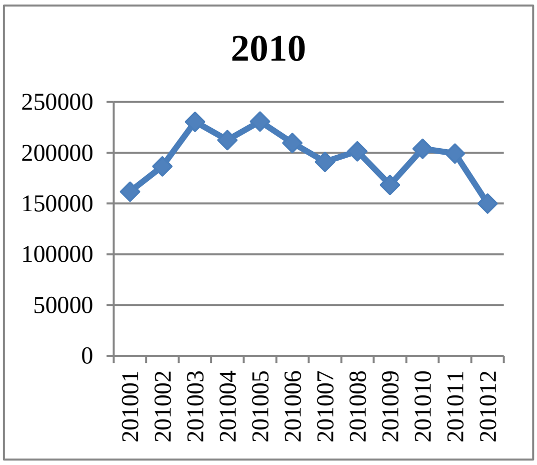
	autumn	summer	winter	spring
2004	561114	489302	517541	648009
2005	1029392	907877	2540202	3967082
2006	5121480	3820114	4588055	4994308
2007	5496880	4240182	5068936	5310661
2008	6409187	5104164	5935128	6687770
2009	7085881	5487975	6851333	6401015
2010	7424899	5211782	6873449	6331885
2011	6629536	5260643	6288146	6551395
2012	6663434	5156672	6626403	6284015
2013	7271667	5152175	6991677	6903681
2014	7863723	5963095	7388029	7587936
2015	8154382	6336705	7890537	7235178
2016	7324313	5747682	6639726	6177442
2017	6576122	5233950	6618034	4057291

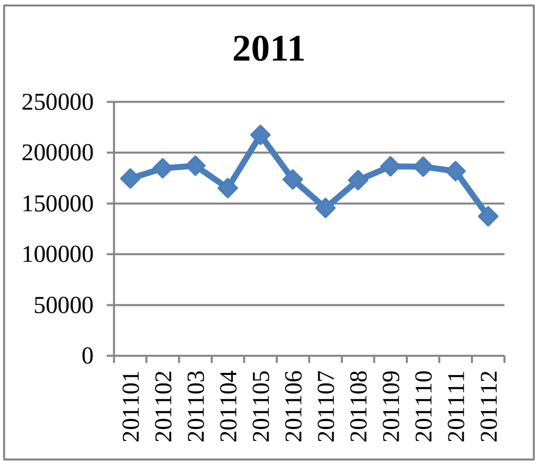

2006
300000
250000
200000
150000
100000
50000
0
200601
200602
200603
200604
200605
200606
200607
200608
200609
200610
200611
200612

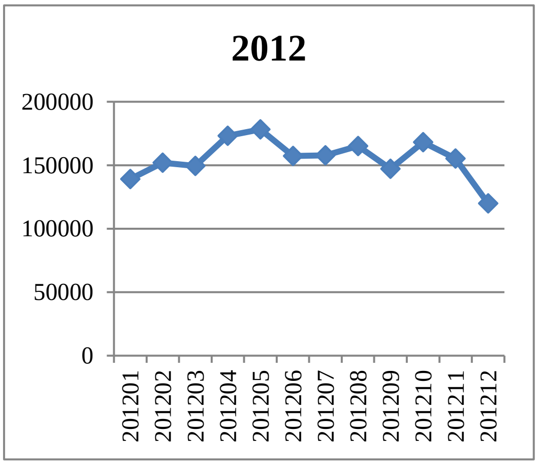

2007
300000
250000
200000
150000
100000
50000
0
200701
200702
200703
200704
200705
200706
200707
200708
200709
200710
200711
200712

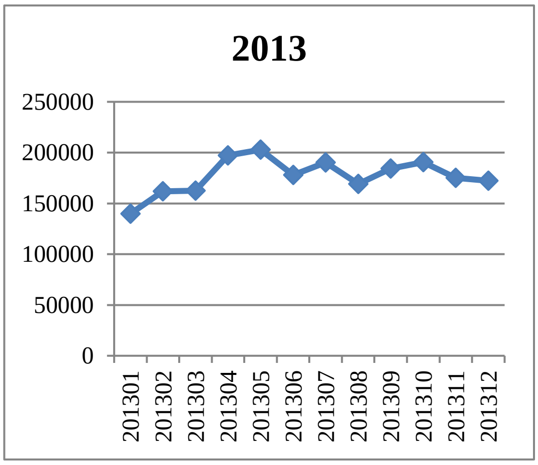

2008
350000
300000
250000
200000
150000
100000
50000
0
200801
200802
200803
200804
200805
200806
200807
200808
200809
200810
200811
200812


2009
300000
250000
200000
150000
100000
50000
0
200901
200902
200903
200904
200905
200906
200907
200908
200909
200910
200911
200912


2010
250000
200000
150000
100000
50000
0
201001
201002
201003
201004
201005
201006
201007
201008
201009
201010
201011
201012


2011
250000
200000
150000
100000
50000
0
201101
201102
201103
201104
201105
201106
201107
201108
201109
201110
201111
201112


2012
200000
150000
100000
50000
0
201201
201202
201203
201204
201205
201206
201207
201208
201209
201210
201211
201212


2013
250000
200000
150000
100000
50000
0
201301
201302
201303
201304
201305
201306
201307
201308
201309
201310
201311
201312

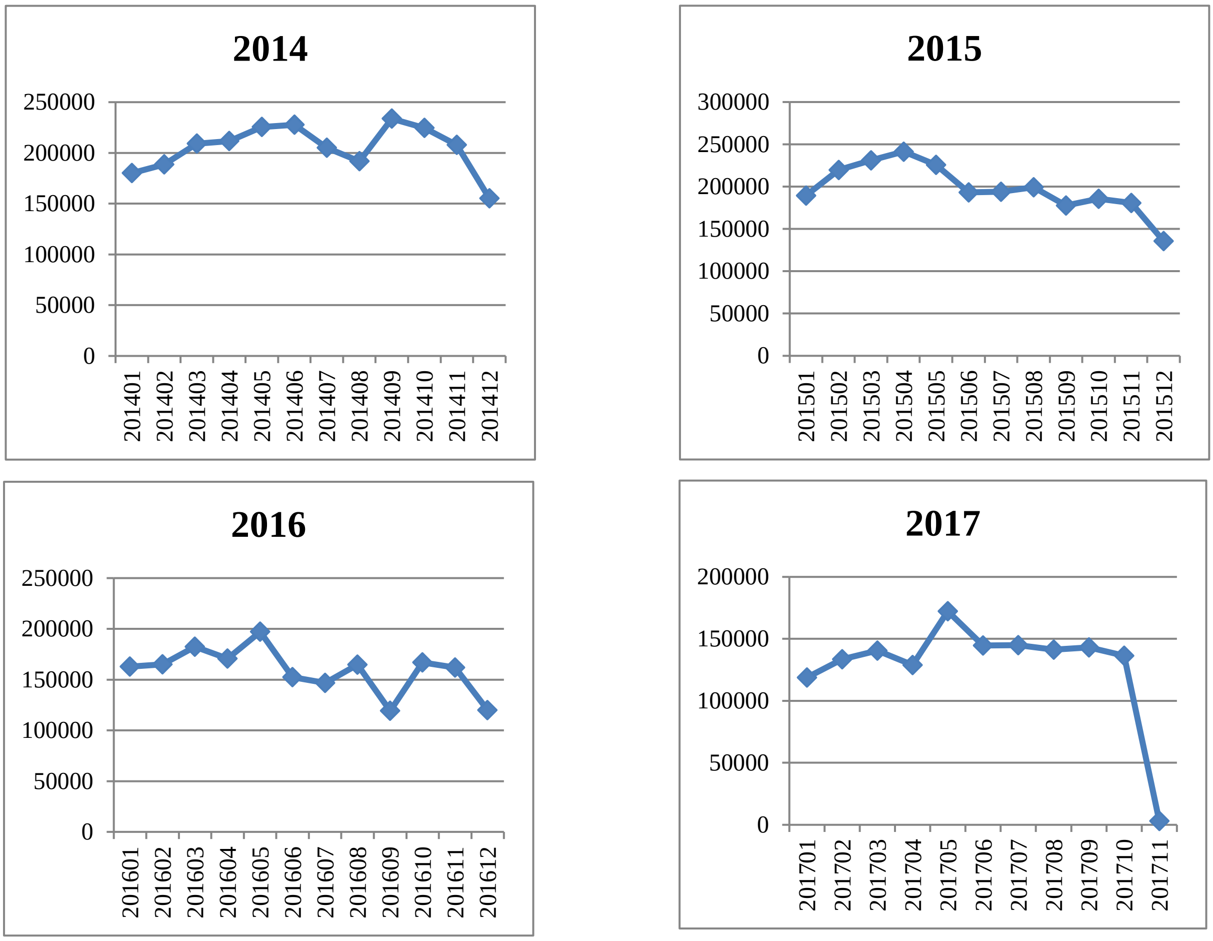
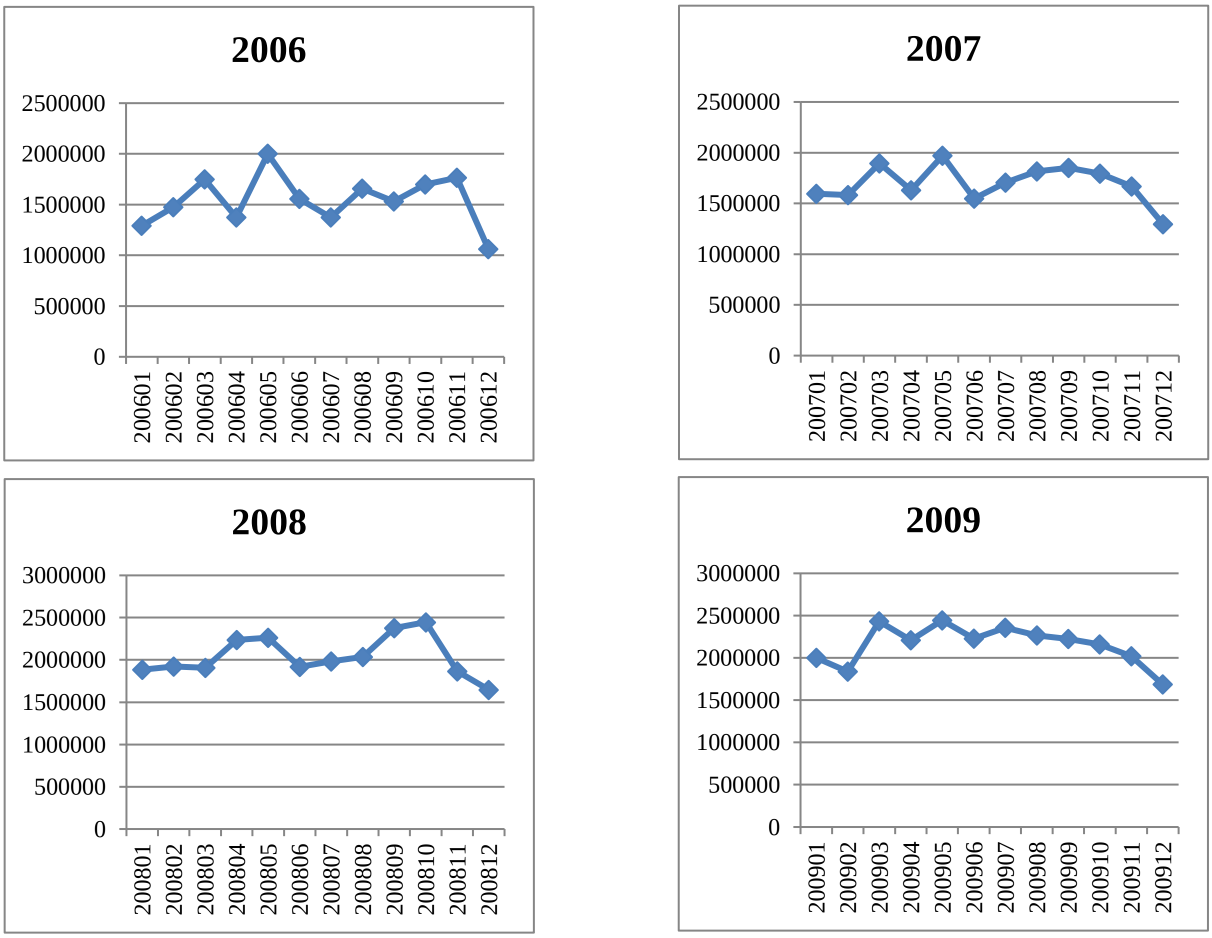


Figure 11. Monthly Missing Transfer Number Distribution
(2006-2017, SourceOrDestinationPIC=XXXXXXXX)



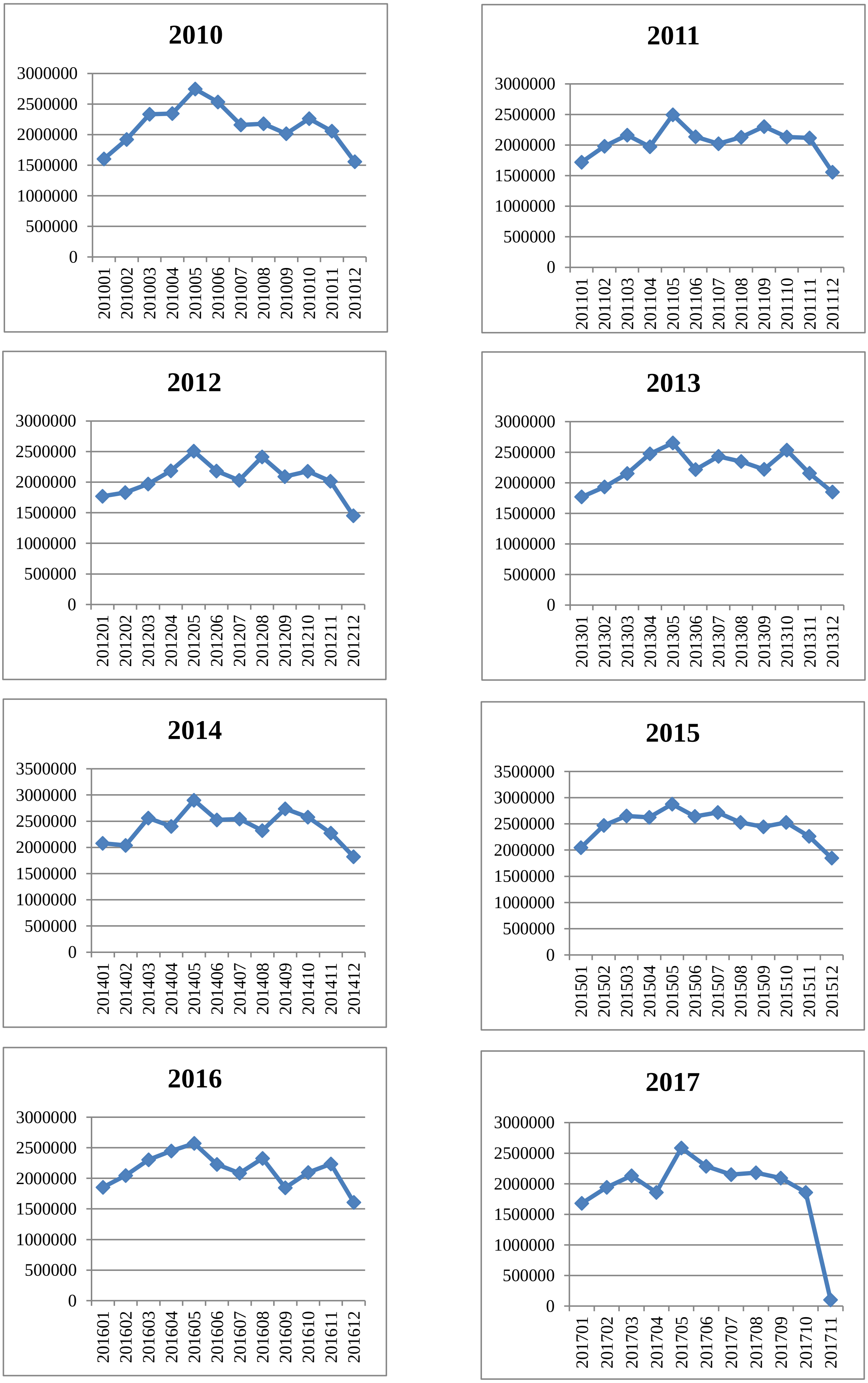


Figure 12. Monthly Normal Transfer Number Distribution (2006-2017)

(4) TAS, VIC, NSW, and NT are more likely to have missing transfers.

TAS, VIC, NSW, and NT are more likely to have missing transfers. Their normal transfers occupy 1.32%, 10.95%, 18.6% and 3.68% among all normal transfers, while their missing transfers occupy 3.76%, 19.54%, 28.87% and 5.96% among missing transfers (Figure 13 and Figure 14).

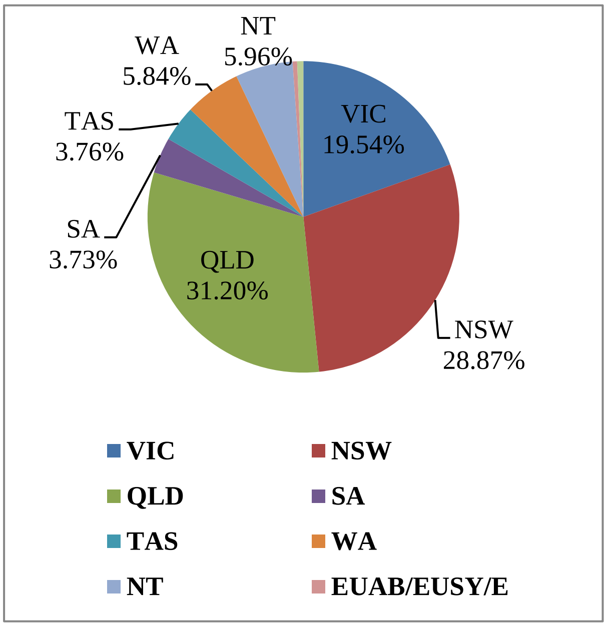


Figure 13. Missing Transfer Distribution Among States

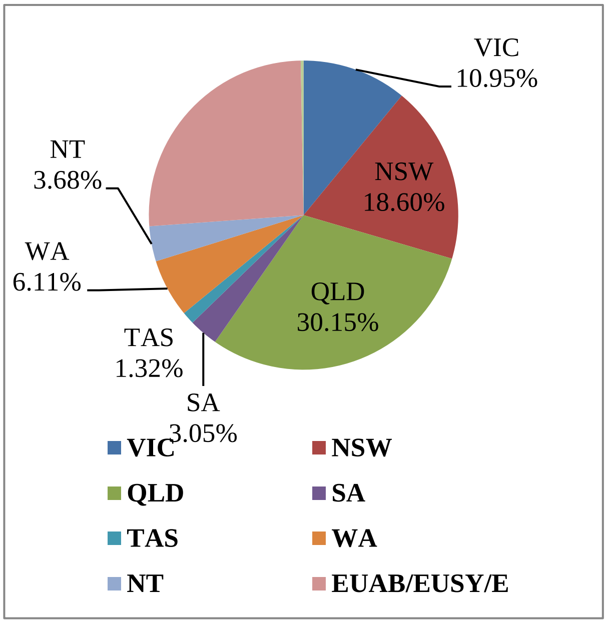


Figure 14. Normal Transfer Distribution among States

(5) TA has the largest missing ratio among all states.

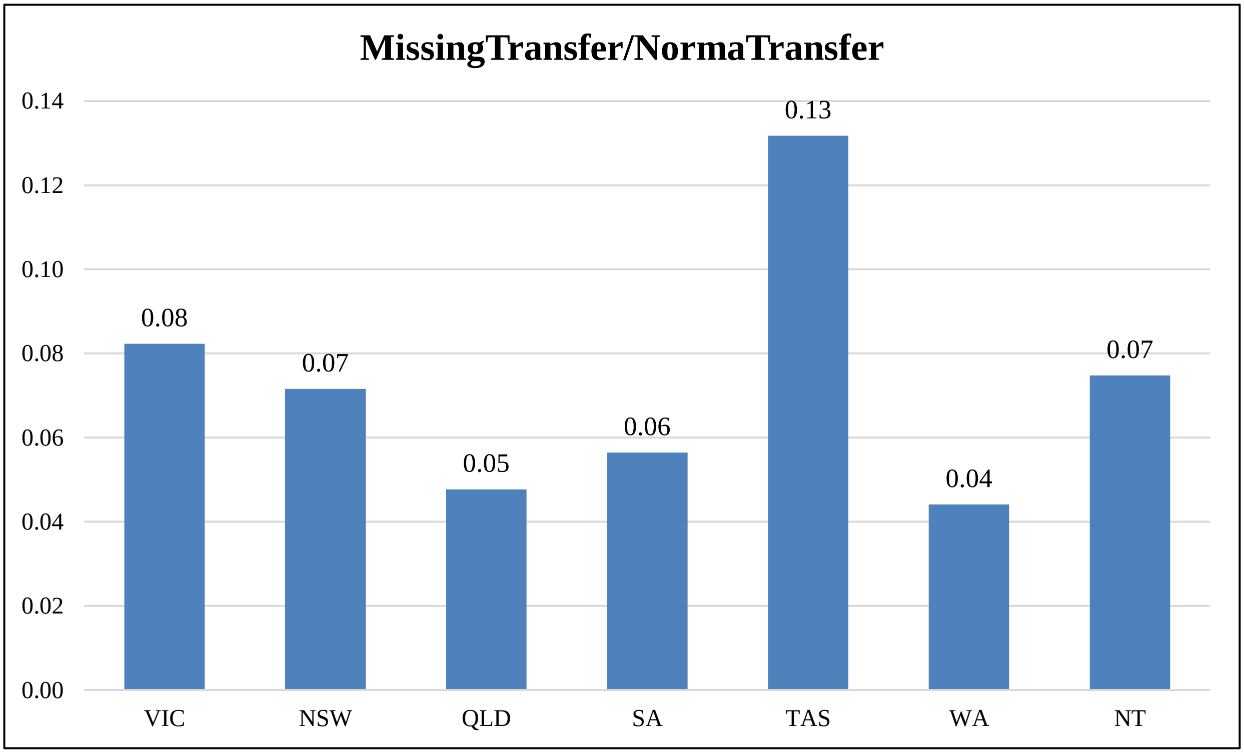


Figure 15. The ratio of missing transfers to normal transfers among different states.

(6) PIC number distributes similarly in missing transfers and normal transfers.

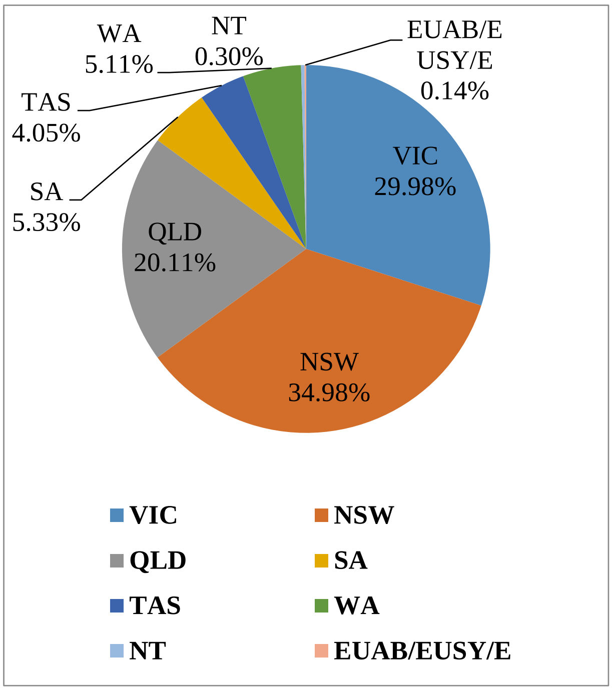


Figure 16. PIC Distribution among States in Missing Transfers

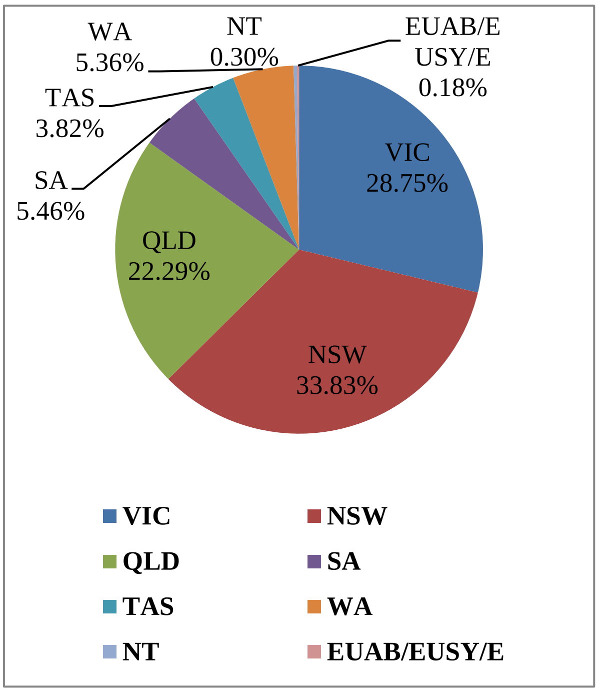


Figure 17. PIC Distribution among States in Normal Transfers

(7) NT (T) has the most normal and missing transfers per PIC.

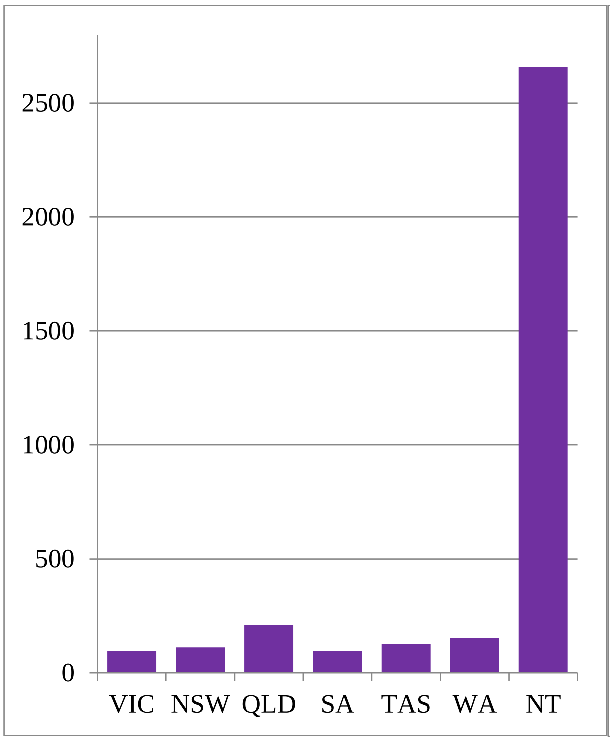


Figure 18. Missing Transfer Number per PIC among states.

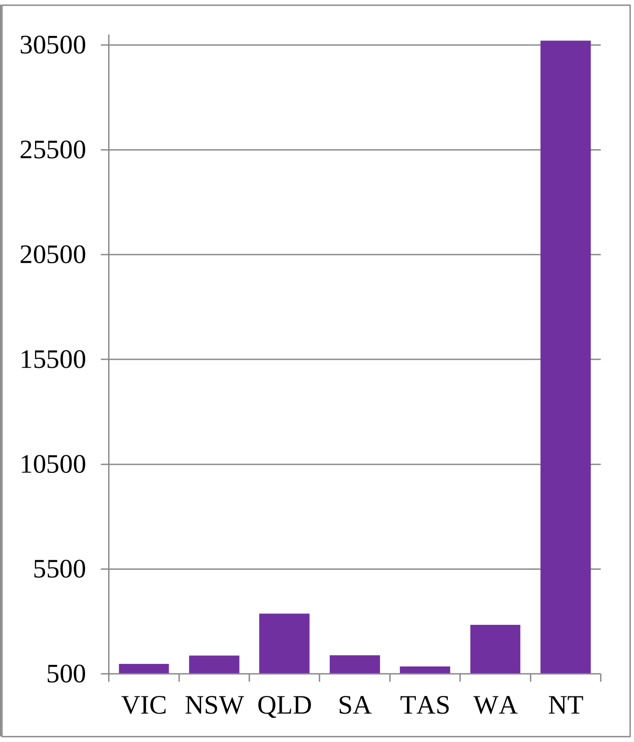


Figure 19. Normal Transfer Number per PIC among states

(8) The percentage of missing transfers keeps declining from 2005. It indicates that the data quality has been improved in recent 12 years.

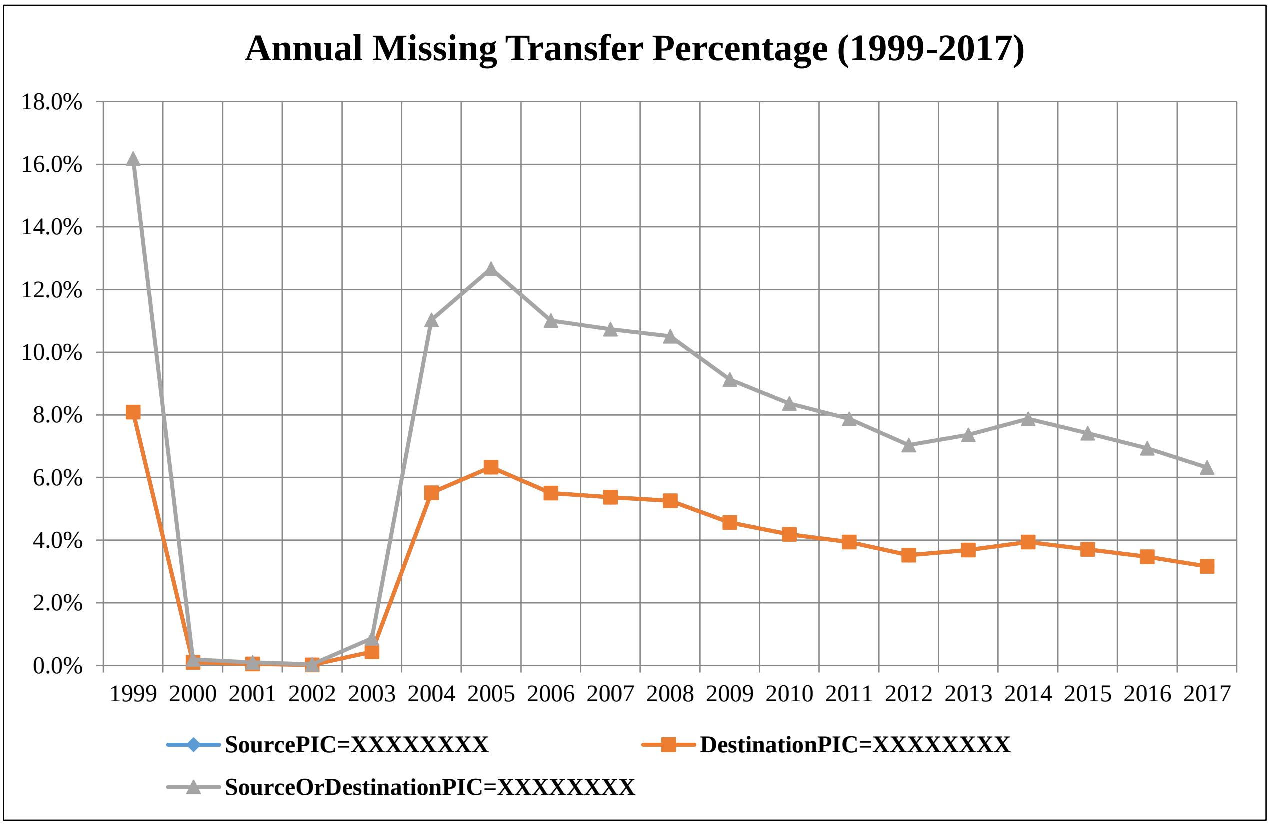


Figure 20. Annual Missing Transfer Percentage from 1999 to 2017

(9) 99% of the source PICs or destinations PICs of missing transfers are PICType 1. By comparison, 75% of the PICs of normal transfer are PICType 1.

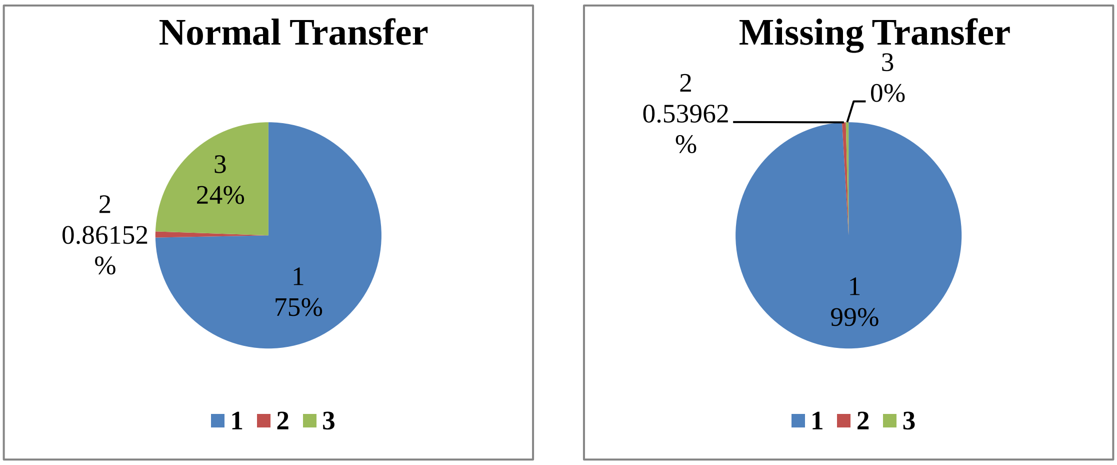


Figure 21. The PICType Distribution of Normal Transfers and Missing Transfers

(10) 95% of the source PICs or destinations PICs of missing transfers have the EUStatus false.

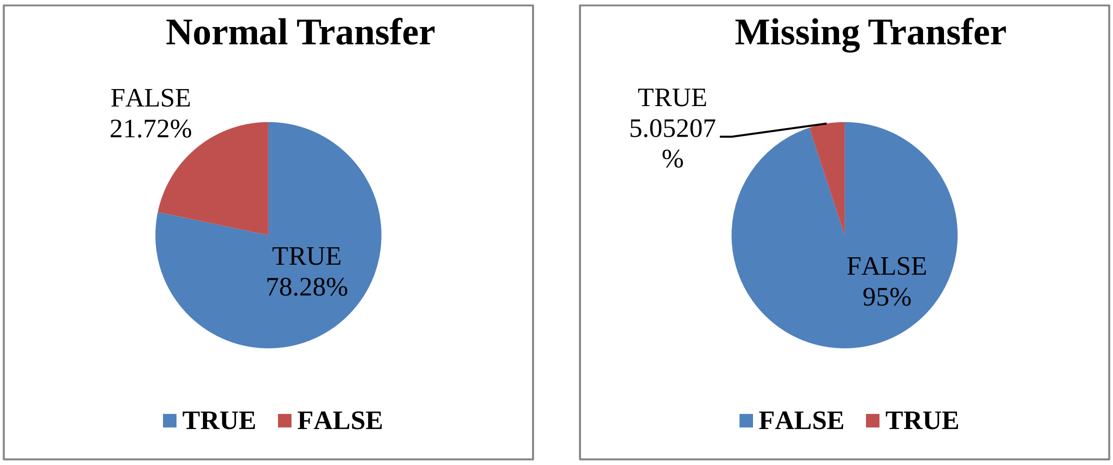


Figure 22. The EUStatus Distribution of Normal Transfers and Missing Transfers

(11) 95% of the source PICs or destinations PICs of missing transfers have the PICRegisterStatus A.

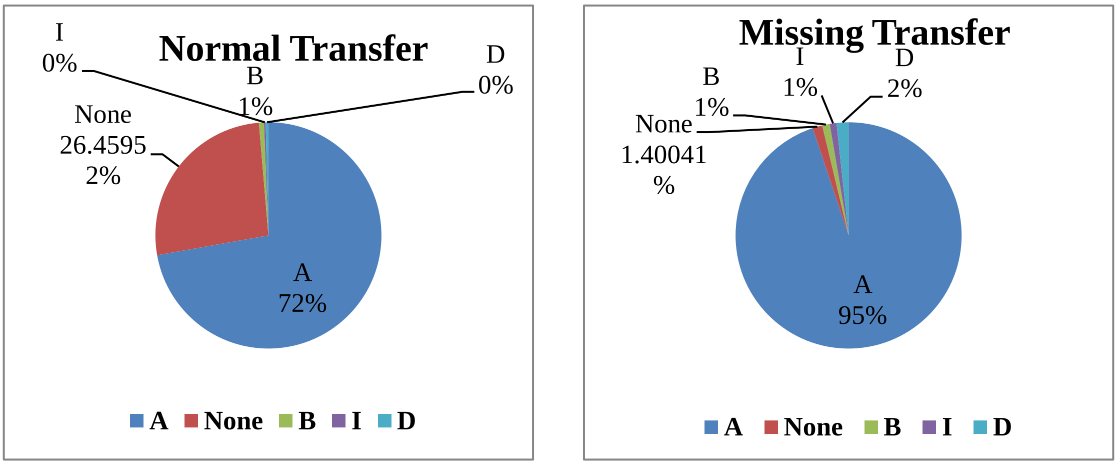


Figure 23. The PICType Distribution of Normal Transfers and Missing Transfers

(12) The top 10 postcode areas which have the most historical SourcePIC missing transfers from 2013.01 to 12.2017 are as follows. 6 of the postcodes are in Queensland, which has more large PICs.

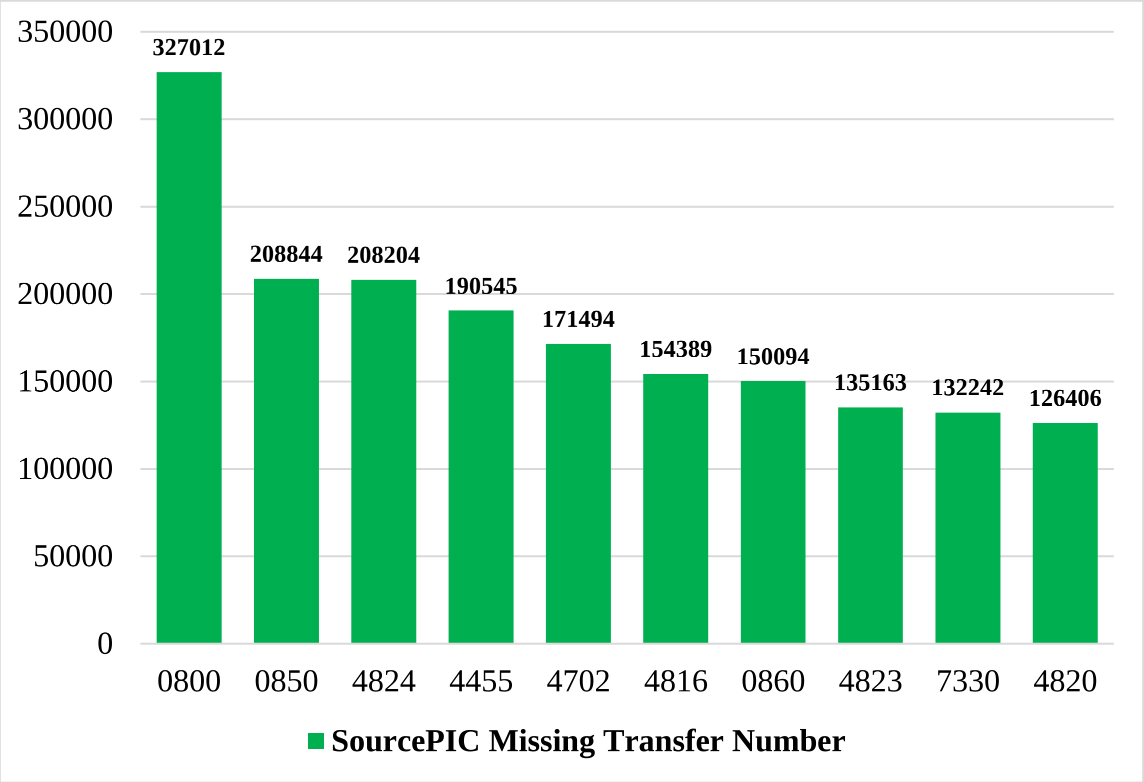


Figure 24. The top 10 postcode areas which have the most historical SourcePIC missing transfers from 2013.01 to 12.2017

4.3 Three Business Scenarios

4.3.1 Scenarios1: Identification of postcode areas and PICs which contribute most to the total missing transfers.

First, we tally the normal transfer number and the missing transfer number in each postcode/PIC in each month/year. Then we calculate the missing transfer ratio for the i^{th} postcode/PIC p_i in the j^{th} month/year in the following two ways:

$$MissingTransferRatioLocal(\mathrm{p}_i, \mathrm{j}) = \frac{\text{Missing Transfer}(\mathrm{p}_i, \mathrm{j})}{\text{Normal Transfer}(\mathrm{p}_i, \mathrm{j})} \quad (1)$$

$$MissingTransferRatioWhole(\mathrm{p}_i, \mathrm{j}) = \frac{\text{Missing Transfer}(\mathrm{p}_i, \mathrm{j})}{\sum_{i=1}^{n} \text{Missing Transfer}(\mathrm{p}_i, \mathrm{j})} \quad (2)$$

Where p_i is the i^{th} PIC/postcode; j is the j^{th} month/year, n is the total number of PICs/postcodes. Next, we report those postcodes/PICs with higher MissingTransferRatioLocal and MissingTransferRatioWhole.

Specifically, we calculated MissingTransferRatioLocal and MissingTransferRatioWhole for each postcode/PIC in each month/year from 2003.01 to 2017.12. A plenty of results have been generated (Additional File 2). Here we take the statistical results for the year 2017 as an example. 20,123 PICs and 102 postcodes have missing transfers only (no normal transfers) in 2017. The top

10 PICs and postcodes are listed Table 4 and Table 5. The top 10 PICs and postcodes in terms of MissingTransferRatioLocal are listed in Table 6 and Table 7. And the top 15 PICs and postcodes on MissingTransferRatioWhole are listed in Table 8 and Table 9.

Table 4. Top 15 PICs which have missing transfers only in 2017

PIC	Missing Transfer Number	Normal Transfer Number	Missing Transfer Local (%)	Missing Transfer Whole (%)
383467	2727	0	Infinity	0.013063
355334	1612	0	Infinity	0.007722
70483	1185	0	Infinity	0.005676
6938	756	0	Infinity	0.003621
25579	736	0	Infinity	0.003526
93486	679	0	Infinity	0.003253
275580	674	0	Infinity	0.003229
73025	671	0	Infinity	0.003214
211811	624	0	Infinity	0.002989
575497	555	0	Infinity	0.002659

Table 5. Top 15 postcodes which have normal transfers only in 2017

Postcode	Missing Transfer Number	Normal Transfer Number	Missing Transfer Local (%)	Missing Transfer Whole (%)
4495	325	0	Infinity	0.046183
2838	61	0	Infinity	0.008668
2153	59	0	Infinity	0.008384
5725	52	0	Infinity	0.007389
6716	42	0	Infinity	0.005968
3133	29	0	Infinity	0.004121
6623	27	0	Infinity	0.003837
4504	23	0	Infinity	0.003268
4566	20	0	Infinity	0.002842
3752	18	0	Infinity	0.002558

Table 6. Top 15 PICs in MissingTransferRatioLocal in 2017

PIC	Missing Transfer Number	Normal Transfer Number	Missing Transfer Local (%)	Missing Transfer Whole (%)
215814	744	1	74400	0.105708
496084	415	1	41500	0.058964
221865	360	1	36000	0.051149
21163	1714	5	34280	0.243526
19885	303	1	30300	0.04305
133895	796	3	26533.33	0.113096
488860	512	2	25600	0.072745
89673	225	1	22500	0.031968
90399	568	3	18933.33	0.080702
247517	163	1	16300	0.023159

Table 7. Top 15 postcodes in MissingTransferRatioLocal in 2017

Postcode	Missing Transfer Number	Normal Transfer Number	Missing Transfer Local (%)	Missing Transfer Whole (%)
6707	102	1	10200	0.014494
6572	37	1	3700	0.005258
7186	62	2	3100	0.00881
6560	289	10	2890	0.041068
3412	25	1	2500	0.003553
4874	120	5	2400	0.017052
5650	24	1	2400	0.00341
6751	775	48	1614.583	0.11013
6336	27	2	1350	0.003837
4575	26	2	1300	0.003695

Table 8. Top 15 PICs in MissingTransferRatioWhole in 2017

PIC	Missing Transfer Number	Normal Transfer Number	Missing Transfer Local (%)	Missing Transfer Whole (%)
170058	8076	14101	57.27253	1.147444
496009	4650	16543	28.10857	0.660676
560996	4088	24506	16.68163	0.580826
81531	3372	5834	57.79911	0.479096
9951	3140	9374	33.49691	0.446134
16931	3114	17324	17.97506	0.44244
383467	2727	0	0	0.387454
219581	2456	1080	227.4074	0.34895
394266	2062	31696	6.505553	0.292971
508198	2018	7902	25.53784	0.286719

Table 9. Top 15 postcodes in MissingTransferRatioWhole in 2017

Postcode	Missing Transfer Number	Normal Transfer Number	Missing Transfer Local (%)	Missing Transfer Whole (%)
4824	14421	60021	24.02659	2.049264
4455	12076	200142	6.033716	1.716033
800	12027	503556	2.388414	1.70907
4816	11538	205886	5.604072	1.639582
850	9652	166441	5.799052	1.371576
4822	9182	68470	13.41025	1.304788
4702	9108	228279	3.989855	1.294272
4730	7620	35629	21.38707	1.082823
4470	7464	22968	32.49739	1.060655
4823	7325	78335	9.350865	1.040903

We implemented two versions of the statistical program, namely the Spark version and the stand-alone version. The statistical results can be found in Additional File 2. The folder "picids" and "postcodes" contain the results of the PIC and postcode area respectively. Within them, both the yearly and monthly statistical results are included in the format of csv. We can easily sort the PICs according to the MissingTransferLocal or MissingTransferWhole using the Excel ranking functions. Attention shall be paid to PICs which contribute more to the total missing transfers.

4.3.2 Scenarios2: Identification of hotspots(PICs) which show continually growing trends in current holdings in recent years, e.g., 9 years.

As described in the Method section (2.2), the simple linear regression is employed to identify trends of the current holdings in recent years. Specifically, we use simple linear regression to approximate the current holding trends of 75, 450 PICs from January 2009 to December 2017. The trend of each PIC can be expressed in the form Y = *a* + *b**X, where X is the time and Y is the current holdings. To show the trends directly and clearly, we visualize the real current holding trend and the approximate line using figures. We classify all the PICs into three categories, namely PICs have positive, negative, and even trends. There are 113,207 PICs have growing trends in current holdings (positive), 105,378 PICs have decreasing trends (negative), 374,437 PICs have the same current holdings across time (even).

The results are attached in Additional File 3. The folder named "Figures" contains all visualized figures, and the figures are stored in the subfolders which are named by their categories. Figure 25 shows three typical PICs which have growing, decreasing, and even trends respectively. The estimated parameters including *a* and *b,* and the statistical results are also included in the file names

"TrendList.csv". The PICs are ranked in descending order of the slope value, i.e., *b*. We list the top 10 PICs whose current holdings grew at the fastest speed in recent 9 years.

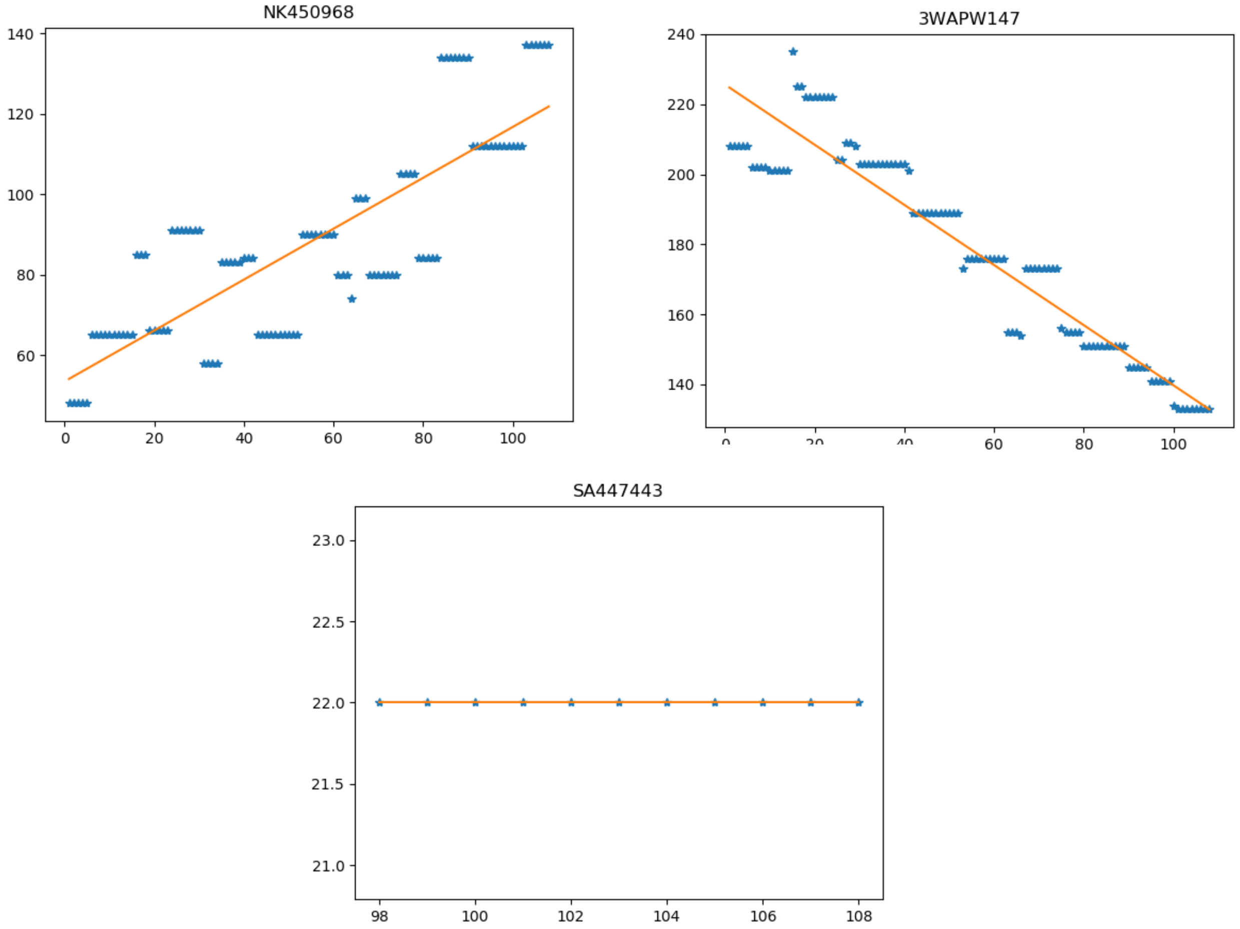


Figure 25. Examples of the current holding trends. The x-axis is the index of the month (from 2009.01 to 2017.12), and the y-axis is the number of current holdings. The blue dots are the real current holdings, and the yellow line is the fitted curve. The LocationID of the PIC is shown on the top of the figure. The PIC NK450968, 3WAPW147 and SA447443 have growing, decreasing, and even trends respectively.

Table 10. Top 10 PICs which grew the fastest in recent 9 years on the current holdings.

PIC	Slope (b)	Intercept (a)
EEEEEEEE	70605.789	1038517
QKCX0272	1280	-135308
3ABEG180	1034.931	-108361
QIBM0761	1000	-105741
QFEE0247	863.7	-89824.8
TCDG0103	810.44544	12791.84
TIBT0023	747.34297	17698.84
TABT0002	709.02659	38964.93

NB201877	690	-72582
TKBT0115	674.38261	116537.9

4.3.3 Scenarios3: Identification of abattoirs which have anomalous killing patterns.

The experiment flow is illustrated in Figure 27. According to the priori domain knowledge, the kill pattern should be weekly periodicity. To identify the abnormal kill patterns, we obtained the current holding sequence across time for each abattoir first (2017.01.01-2017.10.31). Then we split the time sequences into weeks. Next, 244 weekly sequences of normal patterns were randomly selected as training samples to train the One-Class SVM. Finally, all weekly sequences of each abattoir are fed into the trained One-Class SVM to predict their labels (+1 for normal, -1 for anomalous). In order to better show the predicted result, we visualize the time sequence of each abattoir using a line chart. The weekly sequences which belong to normal kill patterns and anomalous kill patterns are colored black and red respectively. Figure 26 shows examples of the colored line charts.

Since the kernel of SVM plays a key role in the prediction task, we tried different SVM kernels for prediction first (linear, poly, RBF, and sigmoid). Extensive comparison experiments show that the linear kernel outperformed other kernels, is more capable to capture the data characteristics.

The results are packaged in the Additional File 4. Specifically, abattoirs which have anomalous weekly sequences are stored in the folder “anomalous”, and abattoirs which have normal weekly sequences are stored in the folder “normal”.

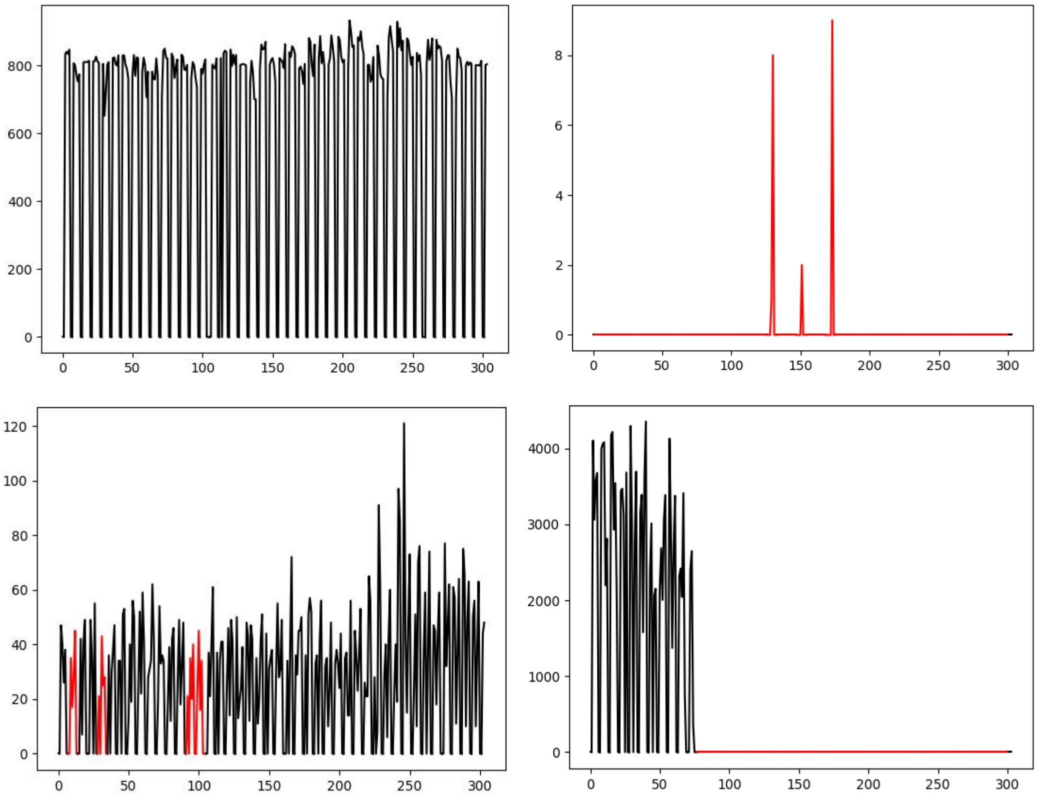


Figure 26. Examples of visualized current holdings trends. The black line denotes the weekly killing pattern is normal, while the red line means the killing pattern is anomalous. The first and the second sub-figures all are normal and anomalous killing patterns. The remaining sub-figures contain both normal and anomalous killing patterns.

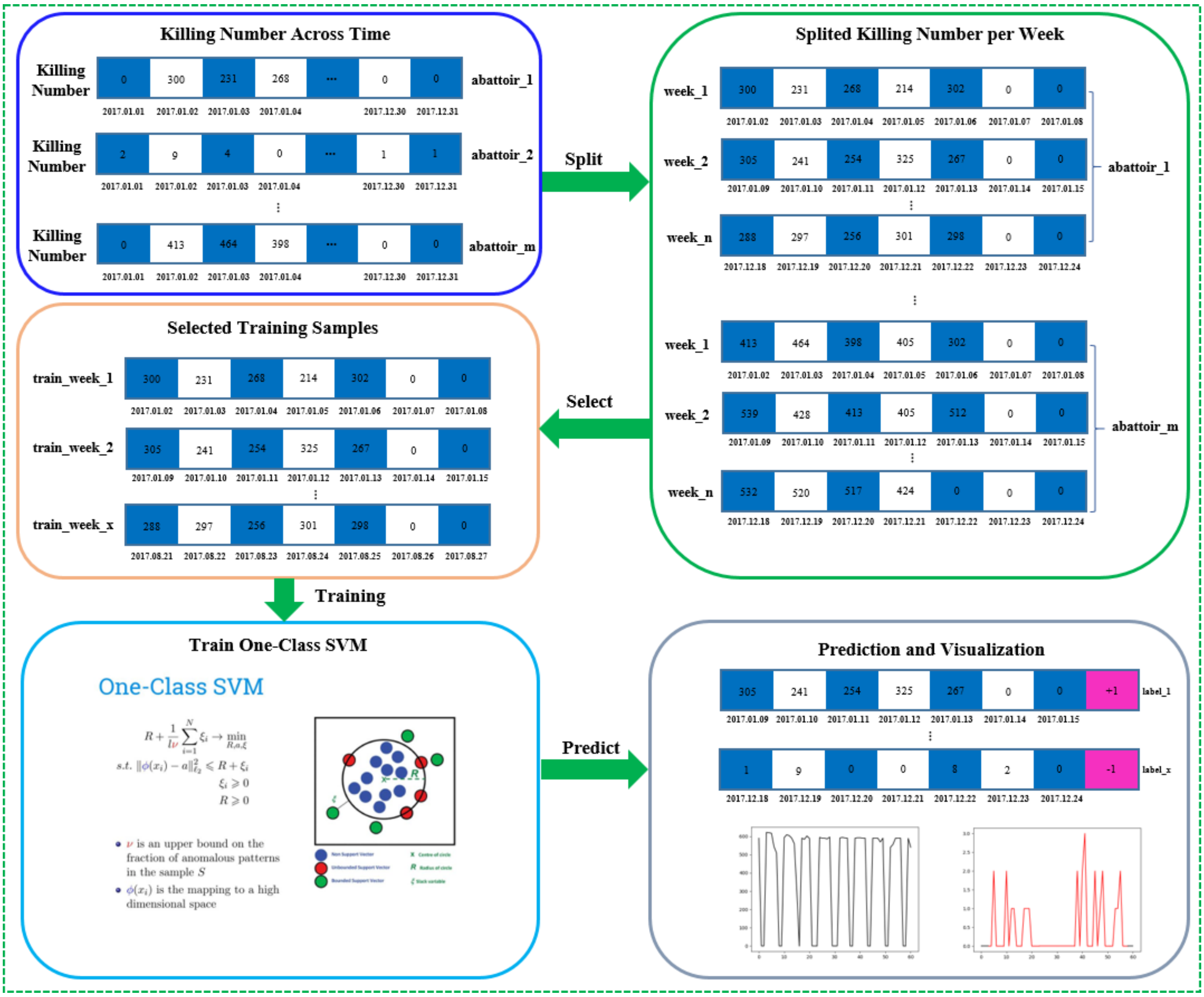


Figure 27. Flow chart of the identification of abattoirs which have anomalous killing patterns using One-Class SVM.

Reference

[1] Wikipedia contributors. "Descriptive statistics." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 10 Oct. 2017. Web. 7 Feb. 2018.

[2] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. Vol. 821. John Wiley & Sons, 2012.

[3] Haklay, Mordechai, and Patrick Weber. "Openstreetmap: User-generated street maps." IEEE Pervasive Computing 7.4 (2008): 12-18.

[4] Chen, Yunqiang, Xiang Sean Zhou, and Thomas S. Huang. "One-class SVM for learning in image retrieval." Image Processing, 2001. Proceedings. 2001 International Conference on. Vol. 1. IEEE, 2001.

[5] Chandola, Varun. Anomaly detection for symbolic sequences and time series data. University of Minnesota, 2009.

[6] Ma, Junshui, and Simon Perkins. "Time-series novelty detection using one-class support vector machines." Neural Networks, 2003. Proceedings of the International Joint Conference on. Vol. 3. IEEE, 2003.