

Document Information

Document Title: PARAT Walkthrough
Original Document Date: 25 October 2011
Document Version: Version 3
Copyright: Privacy Analytics Inc.
Author(s): Khaled El Emam, Grant Middleton, and Jesse Leeson
More Information: www.privacyanalytics.ca

Table of Contents

1	INTRODUCTION	2
2	GETTING STARTED.....	3
3	RISK ASSESSMENT RESULTS.....	5
4	SPECIFYING DE-IDENTIFICATION PARAMETERS.....	7
5	RESULTS OF AUTOMATED DE-IDENTIFICATION	9
6	MANUAL DE-IDENTIFICATION	12
7	LONGITUDINAL DATA	14
8	GEOSPATIAL DATA AGGREGATION.....	18
9	REPORTS	22
10	CONTACT INFORMATION.....	24

1 Introduction

The purpose of this walkthrough is to demonstrate through concrete examples how the PARAT tool can be used to perform a re-identification risk assessment and how to de-identify different types of data sets. The walkthrough will show how to de-identify cross-sectional data, longitudinal data, and geospatial data.

The scenario is that a researcher is requesting access to the data. However, the same walkthrough would apply if, for example, a public health unit has requested the data.

We will assume that the data has already been loaded into the PARAT workspace. Therefore, we will focus only on the risk assessment and de-identification functions, and not on how to go about importing data.

The analyst who is using the PARAT tool to perform the risk analysis and de-identification will be referred to as the “user”. The user may be an analyst in the data custodian’s privacy or compliance office, a statistician, or an individual assigned by a research ethics board (or institutional review board in the US) to perform risk assessments in order to inform their decision making.

Before using the PARAT tool the user needs to do an initial analysis of plausible re-identification attacks on the data set. This exercise would look at the plausible adversaries, what their capabilities are, and what background information they would get access to. Such threat modeling is critical and must be done thoughtfully to ensure that the risk assessment is realistic. The threat modeling exercise needs to be done for any re-identification risk assessment, irrespective of whether a tool like PARAT is used. We will assume that a risk assessment has been performed.

The assumption in putting this walkthrough together is that the reader has some basic understanding of re-identification risk, risk assessment, and de-identification concepts. The walkthrough then focuses on how to use the PARAT tool to automate the analysis that is needed.

Note that the screen shots that are used in this walkthrough were taken from version 2.X of the PARAT tool. Newer versions of the tools may have a slightly different layout for the buttons and icons, but the basic functions and behaviors will be the same as described here.

2 Getting Started

We will start off with a simple cross-sectional dataset; a birth registry. The sample birth registry data set contains approximately 126,000 records.

The first step in performing a re-identification risk assessment is to select the variables that will be used in this analysis. This is done in the “settings” tab of the main PARAT dialogue.

The screen shot of Figure 1 shows the dialogue that is used for selecting the variables (the “quasi-identifiers”) that are included in the risk assessment. On the left hand side are the variables that exist in the data set. Some of these variables will be quasi-identifiers that are included in the risk assessment. The user also needs to specify the re-identification threshold. This threshold indicates how much risk the data custodian is willing to take when disclosing the data to the researcher. The other important parameter is the sampling fraction. The sampling fraction specifies what proportion of the population is included in the data set. For example, if a data set represents all individuals in the population then the sampling fraction would be one, as in this example.

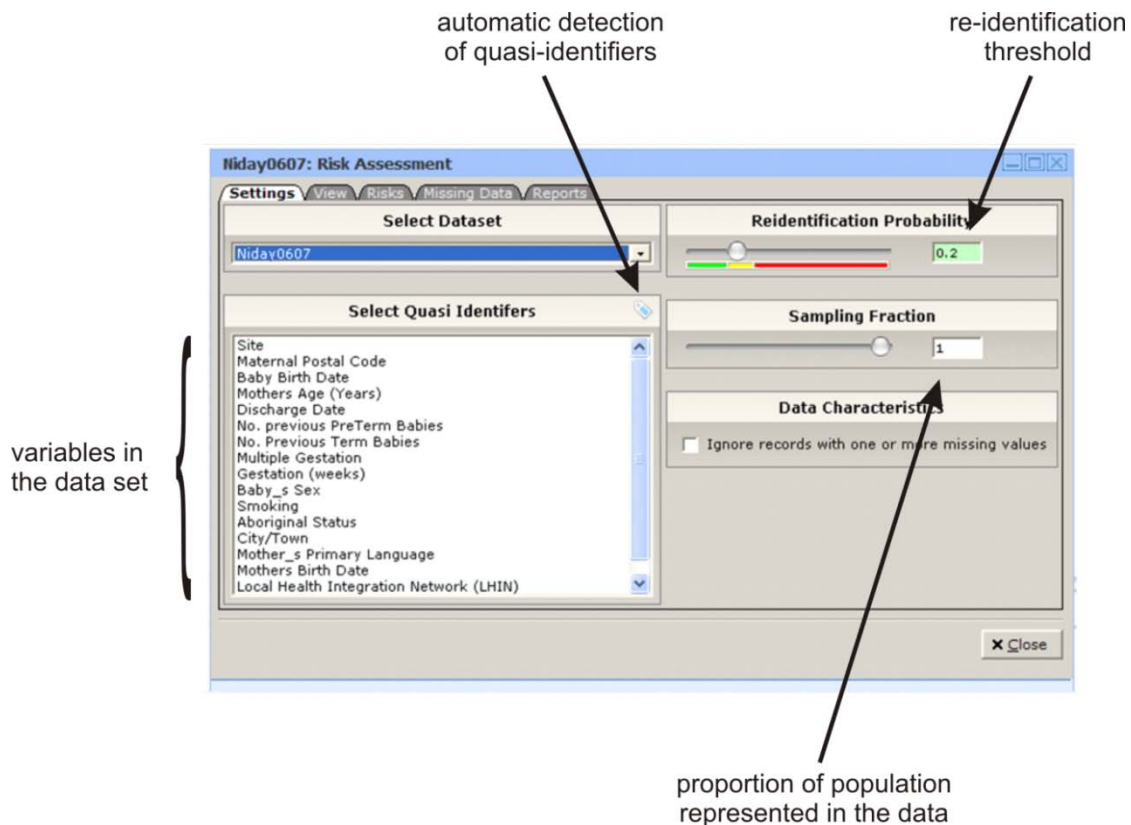


Figure 1: Dialogue for selecting the quasi-identifiers and setting other risk assessment parameters.

The settings dialogues also has the capability to auto detect quasi-identifiers. This uses a set of heuristics to look for fields that are often quasi-identifiers. The analysis is based on the contents of the fields and not on the names of the fields. The auto-detect feature is useful when there are a large number of variables and it may be time-consuming to go through them one-by-one. The auto detect feature will help the user focus on the most likely fields.

As shown in Figure 2, the user then selects the specific quasi-identifiers that will be used for risk assessment. The specific quasi-identifiers that are selected will depend on the threat modeling exercise that is performed before using PARAT. Threat modeling evaluates the plausible re-identification attacks that can be launched on the data set, and based on that, the quasi-identifiers are chosen.

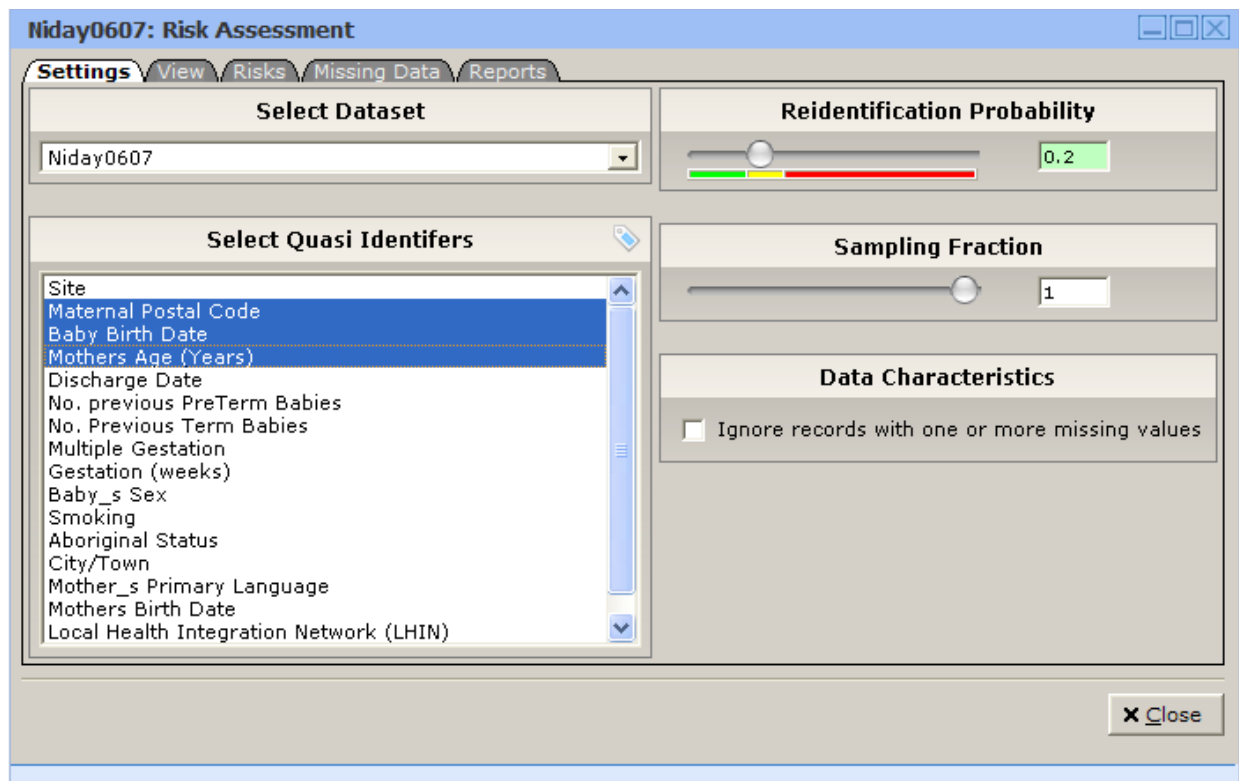


Figure 2: The settings dialogue screen after the user has selected the quasi-identifiers that will be used for analysis.

3 Risk Assessment Results

The results of the risk assessment on that set of quasi-identifiers are shown in Figure 3. Given the parameters that were selected by the user, the only two types of risk that are relevant are “prosecutor” and “marketer”. Prosecutor risk pertains to the highest risk for a single record, and marketer risk is the average risk across all of the records. The latter will be the same as or smaller than the former. In the dialogue shown in Figure 3 the user is interested in the results for marketer risk.

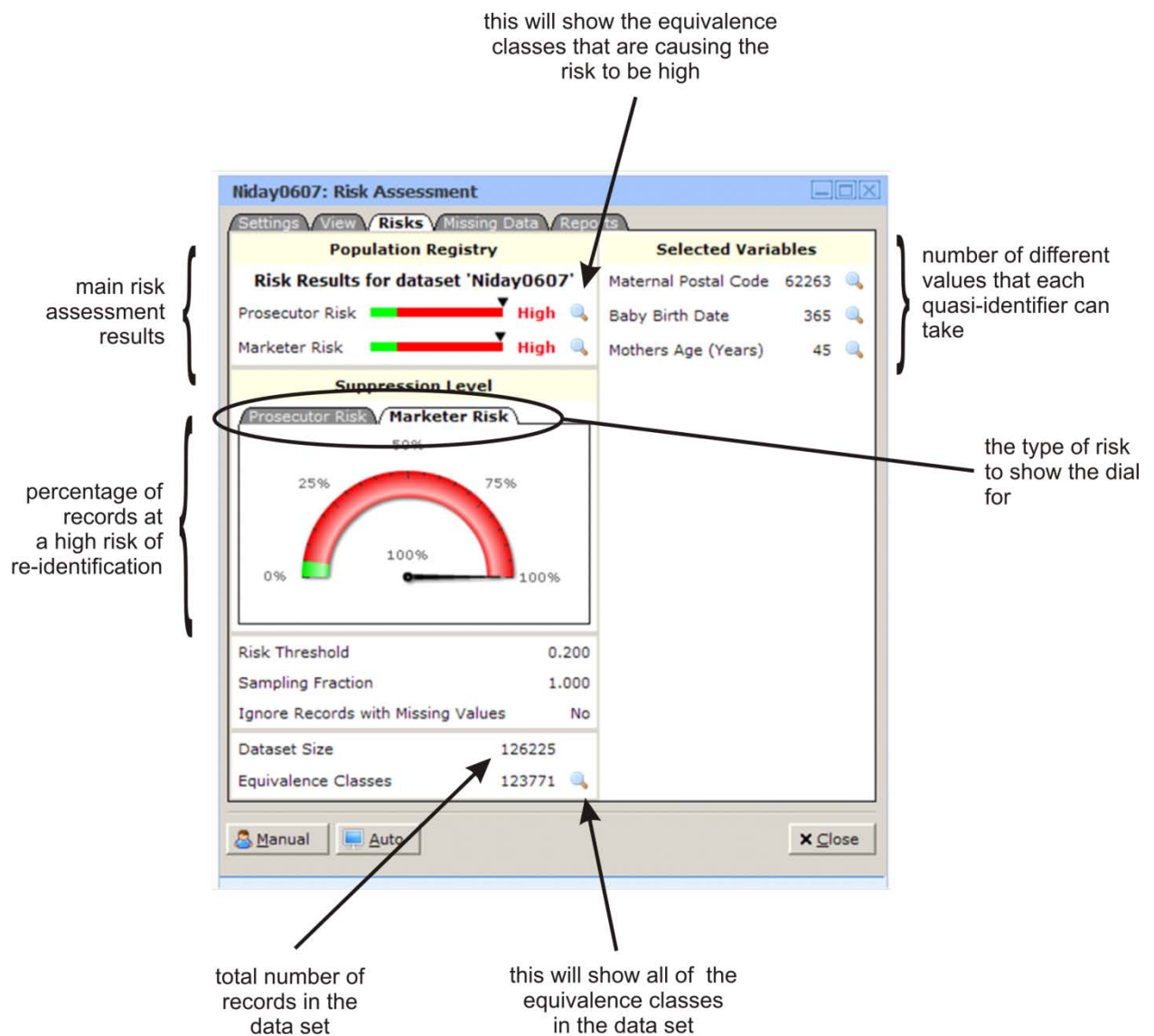


Figure 3: The dialogue showing the results of the risk assessment.

You can see the re-identification risk relative to the threshold for each type of risk at the top. On the green/red bar the green portion shows the risk range that is below the threshold and the red portion shows the risk range that is above the threshold. The arrow on top indicates where the actual risk on the data set is. As can be seen, in this case the risk is close to the maximum for both types of risk.

The dial shows the percentage of records that are high risk. The tool assigns a risk value to every record in the data set. If the risk is higher than the threshold then that is flagged as a high risk record. The dial shows the percentage of records that are considered high risk. In this example the risk is high for all of the records (100%). Of course, if the user changes the threshold (in the settings tab) then this will produce a different answer.

The dialogue also shows that there were 62,263 different values for the maternal postal code in this data set, and 45 different values for the mother's age. The magnifying glasses allow the user to view the equivalence classes themselves. The magnifying glass next to the high risk bars would allow the user to see which equivalence classes are causing the risk to be high.

The results make clear that the user should not release the data as-is because the risk of re-identification is high.

4 Specifying De-identification Parameters

Figure 4 shows the dialogue that is used to specify the parameters for automated de-identification. Automated de-identification finds the optimal de-identification solution for the data set. It uses a combination of generalization and suppression to do so, trying to keep both at a minimum but still find a solution that has a re-identification risk below the threshold.

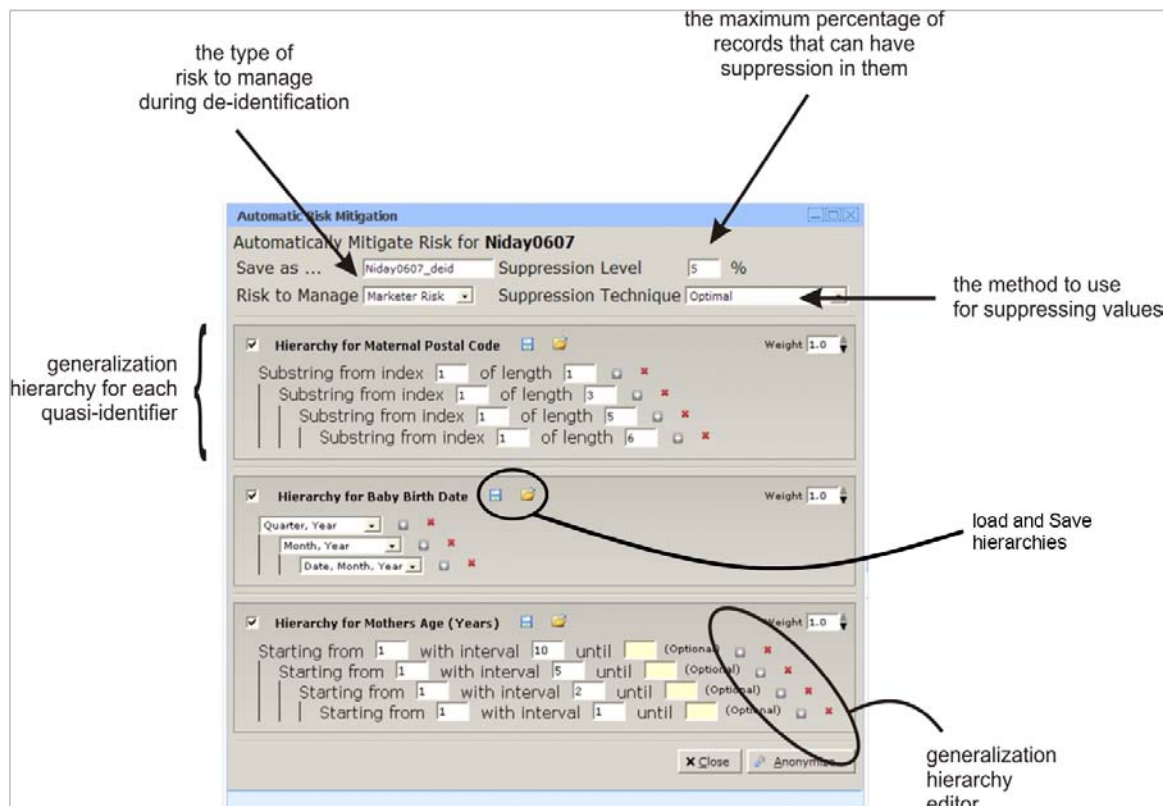


Figure 4: The dialogue for setting the parameters for automated de-identification.

The settings screen includes the definition of hierarchies as well as the ability to assign weights to specific variables. The “suppression level” is the maximum percentage of records that would be allowed to have suppression in them due to the de-identification. The suppression algorithm is “optimal”, which means it will only suppress cells in those records that are flagged for suppression and only on the quasi-identifiers selected. An optimization algorithm is used to find the fewest possible cells to suppress that will ensure that the re-identification risk is below the threshold.

The hierarchies can be edited by the user. Alternatively, there are a number of standard hierarchies that are included in the default library and that can be loaded by the user. The

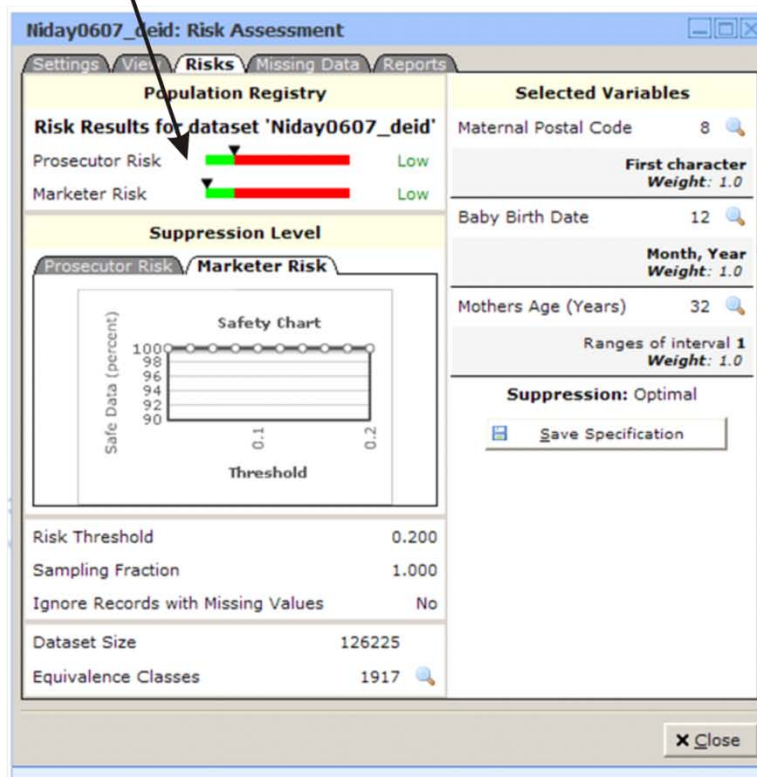
generalization hierarchies reflect the acceptable generalizations of the quasi-identifiers. They are defined for integers, real numbers, dates, and strings (or fields that can be treated as strings, such as postal codes or ZIP codes).

5 Results of Automated De-identification

The screen shot in Figure 5 shows the result of the de-identification. The right hand side shows the optimal de-identification (one that balances generalization with suppression). The “safety chart” is the cumulative distribution. In this example it basically shows that there is no reduction in risk if the threshold is reduced.

The optimal de-identification keeps the first character of the postal code, the month and year of birth for the baby, and the mother's age in years.

the re-identification risk is below the threshold for both types of risk



the optimal generalization for this quasi-identifier

Figure 5: The results of the automated de-identification showing the solution.

Sometimes such a solution is not completely satisfactory. One way to adjust the results is to change the weights of the quasi-identifiers. We can change the weights of the variables as shown in Figure 6, assigning lower weights to variables that can have more de-identification applied to them. The weights need to be obtained from the researcher because they reflect how the quasi-identifiers will be used in an analysis. A lower weight means that more distortion is acceptable on that quasi-identifier.

In this example the user maintained a high weight for the postal code. This makes sense if, for example, the researcher wishes to perform geospatial analysis and the location of the patients is important. The age was given the lowest weight. This is appropriate if the researcher was going to group the ages into intervals anyway during the analysis.

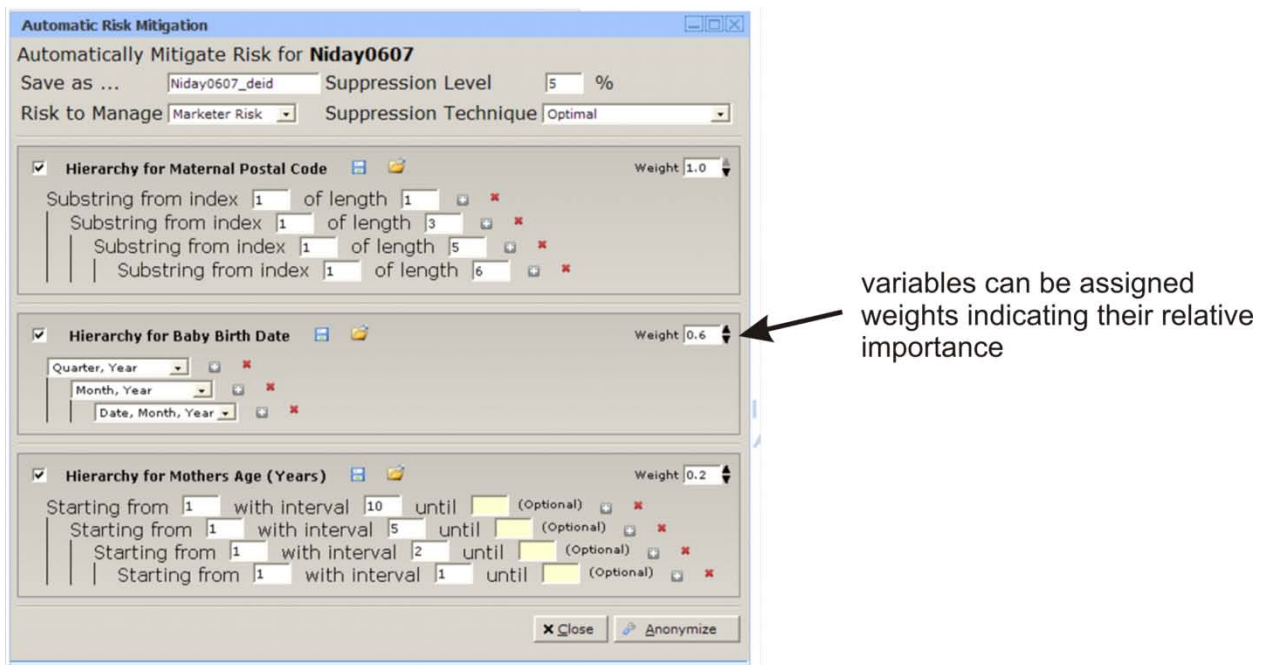


Figure 6: The weights of the variables can be changed to reflect their relative importance.

As shown in Figure 7, we will get a different optimal de-identification in this example because the postal code had the highest weight. In this example more geospatial information is retained because we assigned lower weights to some of the other variables. Therefore, three characters of the postal code were retained as opposed to a single character. Because the age had the lowest weight it was generalized to a 10 year interval.

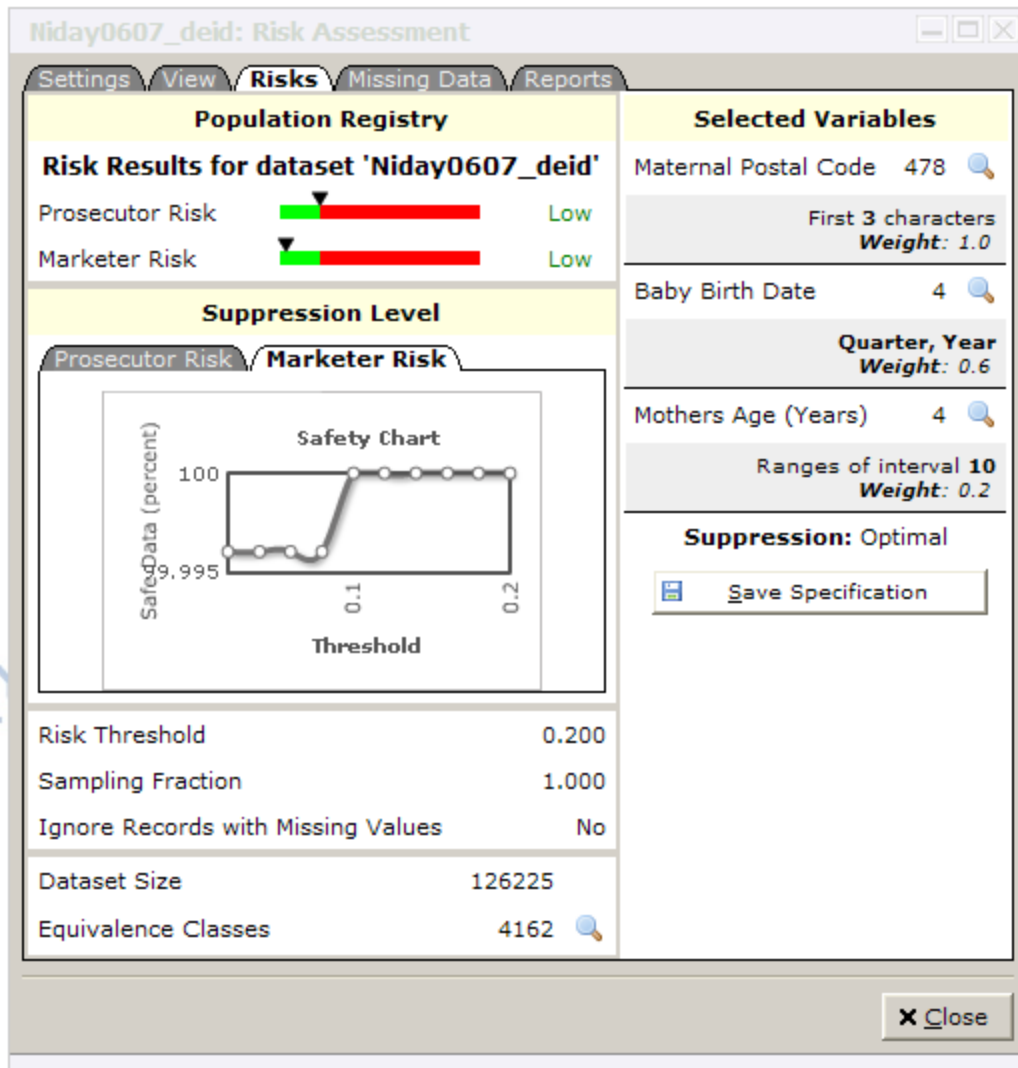


Figure 7: The results of automated de-identification when the weights are altered.

6 Manual De-identification

You can also do manual de-identification to perform “what-if” analyses of risk as illustrated in Figure 8. In this example we defined a specific generalization on our quasi-identifiers. The age was top-coded (ages 51 and above are included in the same category), and bottom-coded (ages below 19 are coded into a single category). The date of birth was generalized to month and year of birth, and the postal code was generalized to only the first three characters. No suppression was specified because we wanted to see what the risk would be with just the generalization.

The screenshot shows the 'Risk Mitigation for Niday0607' dialog box. The 'Save As...' field contains 'Niday0607_deid2'. The 'Risk to Manage' is set to 'Prosecutor Risk'. The 'Generalization Technique' is 'Manual' with a 'Suppression Limit' of 5.00. The 'Suppression Technique' is 'None'. There are three checked hierarchy sections:

- Hierarchy for Maternal Postal Code:** Substring from index 1 of length 3. Weight: 1.0.
- Hierarchy for Baby Birth Date:** Month, Year. Weight: 1.0.
- Hierarchy for Mothers Age (Years):** Starting from 19- with interval 5 until 51+ (Optional). Weight: 1.0.

Two arrows point to the '19-' and '51+' fields in the 'Mothers Age (Years)' section, labeled 'bottom-code specification' and 'top-code specification' respectively. At the bottom right are 'Close' and 'Anonymize' buttons.

Figure 8: Dialogue to specify the parameters for manual generalization and suppression.

The results of that risk assessment are shown in Figure 9. As can be seen, only 0.17% of the records have a high risk of re-identification if we are concerned with marketer risk, and approximately 25.4% are high risk if we are concerned with prosecutor risk. In the case of marketer risk, with such a small percentage at risk these records could be easily suppressed. However, with prosecutor risk some additional generalization would probably be needed, otherwise too many records would have suppression.

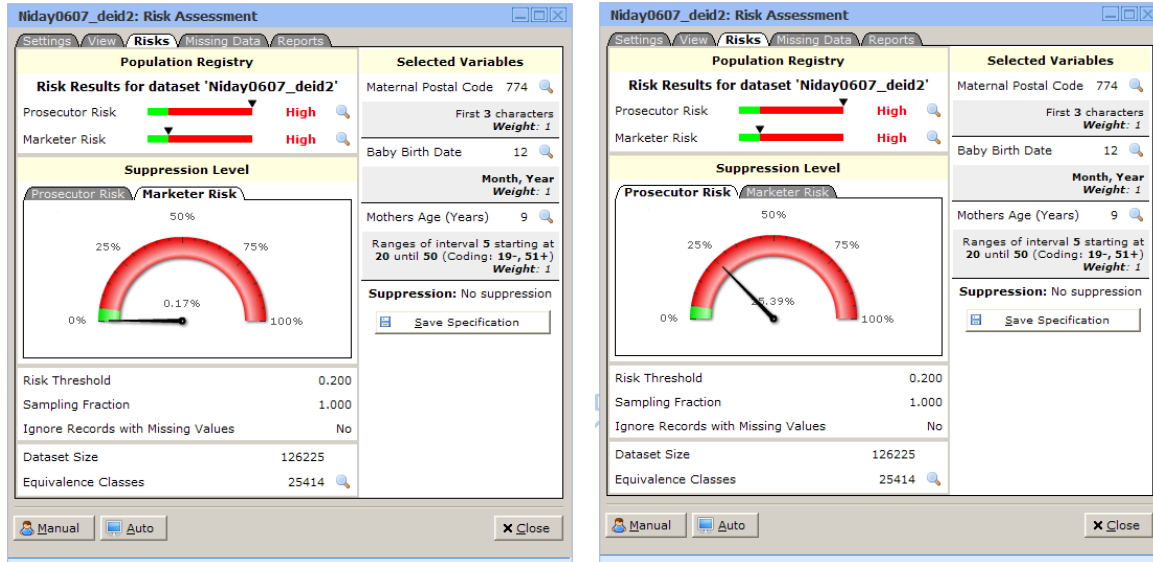


Figure 9: The results after manual generalization.

7 Longitudinal Data

PARAT also supports performing risk assessments and risk mitigation on longitudinal data. For this example we will use the longitudinal dataset that comes with PARAT in the samples folder. This data set contains patients, their demographics and their hospital visit information. Longitudinal data contains multiple levels of information. In this example, the patient demographic information data is at the first level. The patient's visit information is considered the second level of data.

To perform a risk assessment, you must open a New Risk Assessment window. After selecting the dataset, you need to click the "Use Levels" button (Figure 10). This button tells PARAT that the data is longitudinal.

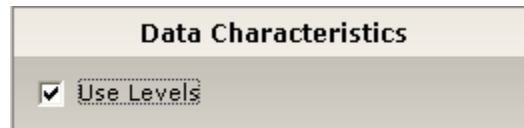


Figure 10: Use Levels Button

The next step is to specify which fields belong at which level, as well as indicating which fields are primary keys and which are quasi identifiers. Figure 11 and Figure 12 show the settings for both level 1 and level 2.

For level 1, you need to specify both a primary key and quasi identifiers. The primary key is "Patient_id" and the quasi-identifier is "Gender". At level 2 you only need to specify quasi identifiers. The level 2 quasi-identifiers are "Visit_Dates" and "Postal_Code".

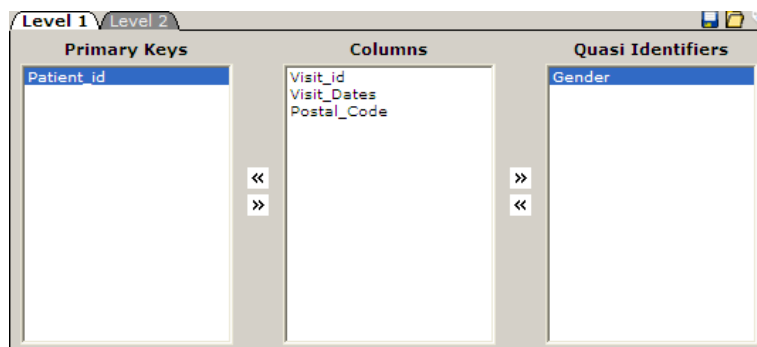


Figure 11: Level 1 Settings

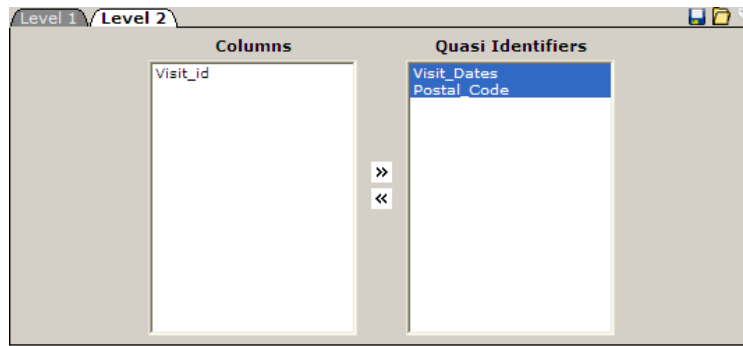


Figure 12: Level 2 Settings

Now that the settings are provided, we can go to the risk window. This window is almost identical to a cross sectional risk assessment. The two main differences are the additional information in the equivalence classes and the fact that you cannot view the records at risk.

Figure 13 shows the results of the risk assessment. In this example we can see that the dataset has both high prosecutor risk and high marketer risk. When we look at both prosecutor and marketer risk, we see that 100% of the records are at risk. Since the data is at a high risk, we will want to de-identify it.

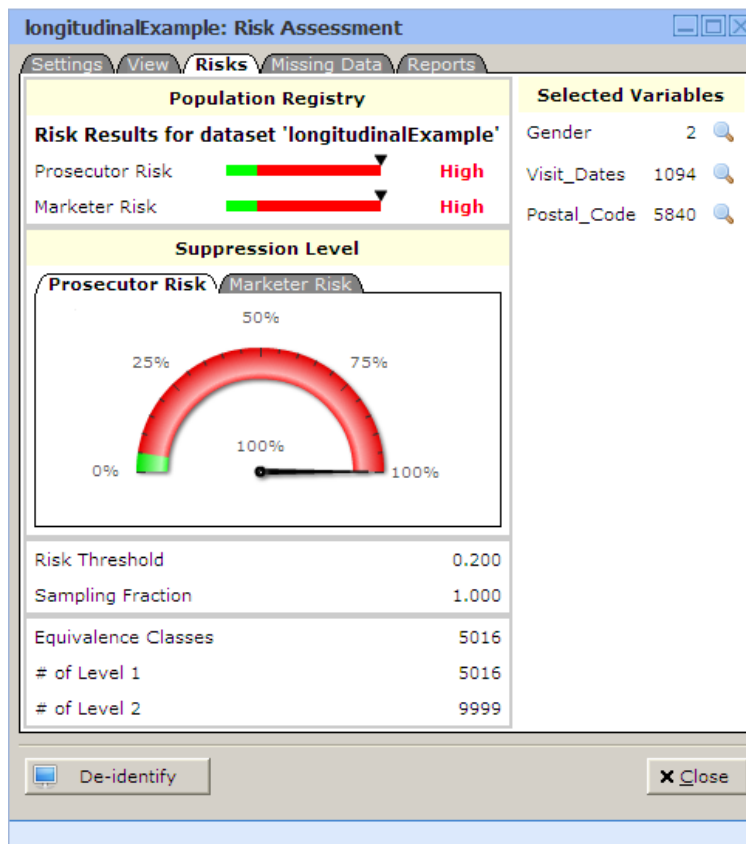


Figure 13: Longitudinal Risk Assessment

To perform the de-identification we need to specify the different de-identification settings and specify the hierarchies for each field. Figure 14 shows the settings used for this example. We are going to manage prosecutor risk, use optimal generalization and suppression, and use the automatic suppression limit. The hierarchies for all three variables can be seen in Figure 14. We then click anonymize to perform the risk mitigation.

Risk Mitigation for longitudinalExample

Save As...

Risk to Manage Auto Suppression Limit

Generalization Technique Suppression Limit

Suppression Technique

Hierarchy for Gender Weight

Hierarchy for Visit_Dates Weight

Year

Quarter, Year

Month, Year

Date, Month, Year

Hierarchy for Postal_Code Weight

Substring from index of length

Substring from index of length

Substring from index of length

Substring from index of length

Figure 14: Risk Mitigation Settings

Figure 15 shows the results of the risk mitigation. We see that both prosecutor and marketer risk are now low and that the percentage of records at risk is 0%. We see that the “gender” field was left unchanged, the “Visit_Dates” field was generalized to year, and that the “Postal_Code” field was generalized to the first character.

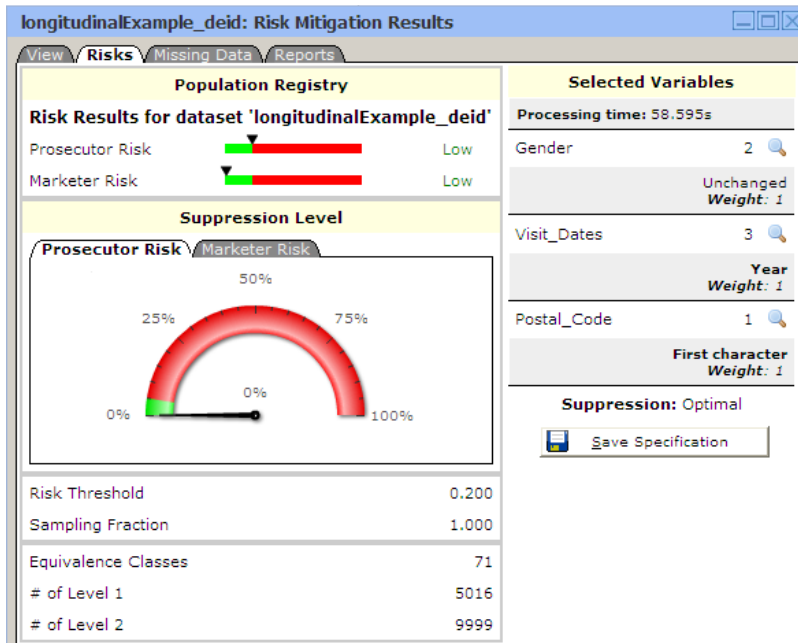


Figure 15: Risk Mitigation Results

8 Geospatial Data Aggregation

To generalize Canadian postal codes using PARAT's geospatial aggregation feature, you must first install the geospatial aggregation database. This database is included with PARAT on a USB key. Instructions for installing this database are found in PARAT's Installation Manual. Once the database is installed the aggregation feature will be available in PARAT.

In this example we will use a dataset that contains a single field, named "PostalCode". This dataset contains 82,828 records with 14,305 unique postal codes.

The first step is to open a de-identification window for the dataset. This can be done by opening a New Risk Assessment, selecting the appropriate dataset and quasi identifiers, opening the Risk tab and then finally clicking the De-identify button. Inside the hierarchy editor, next to the "PostalCode" field you will see the "Generalize as a Postal Code" button (Figure 16).



Figure 16: Generalize as Postal Code Button

This button has a mailbox image and when the button has been clicked, the mailbox on the button will become active (Figure 17). This shows that we are generalizing this field using our geographic aggregation technique. We will click this button for the "PostalCode" field.

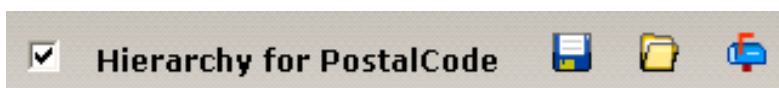


Figure 17: Activated Mailbox

We then click the anonymize button to perform the de-identification. The de-identification will appear just like any other de-identification; it will first generalize the data and then perform any necessary suppression. Where this feature differs is the generalization that is applied to the data. The data is aggregated into special groupings. These groupings consist of aggregated postal codes.

The resulting dataset, within the view tab, will look like Figure 18. The resulting data contains a set of numbers and may also contain 3 letter postal codes. The numbers represent the different postal code groupings. When the data is exported a lookup table will also be exported. This table will show you the set of postal codes that belong to each group.

Forward Sortation Areas (FSAs) may show up in the data. This happens when a postal code in the set is not found in the geographic aggregation database. The postal codes that are not in this database are automatically truncated to 3 letters (FSA), "KOA" in this example.

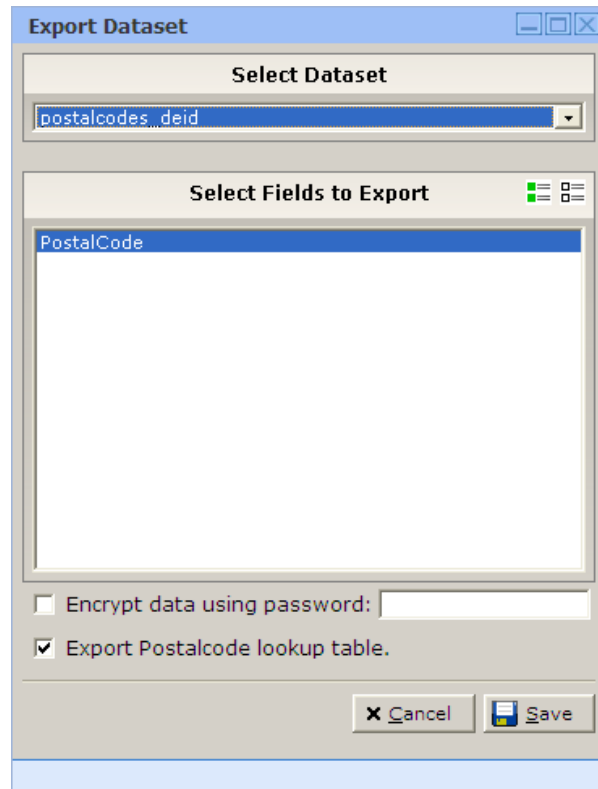
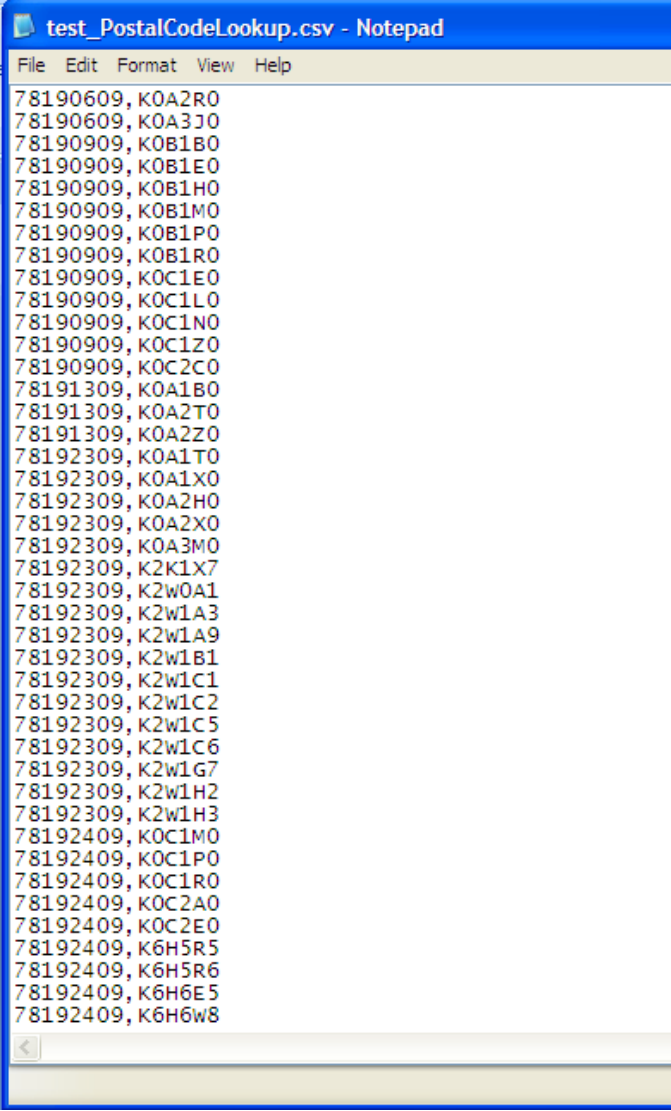


Figure 19: Export Data Window

Figure 20 shows the exported postal code lookup CSV file. Here we can see the group ID 78190609 that is shown in Figure 18. This shows that the following postal codes are a part of this group: K0A2R0, K0A3J0.



```
test_PostalCodeLookup.csv - Notepad
File Edit Format View Help
78190609, KOA2R0
78190609, KOA3J0
78190909, KOB1B0
78190909, KOB1E0
78190909, KOB1H0
78190909, KOB1M0
78190909, KOB1P0
78190909, KOB1R0
78190909, KOC1E0
78190909, KOC1L0
78190909, KOC1N0
78190909, KOC1Z0
78190909, KOC2C0
78191309, KOA1B0
78191309, KOA2T0
78191309, KOA2Z0
78192309, KOA1T0
78192309, KOA1X0
78192309, KOA2H0
78192309, KOA2X0
78192309, KOA3M0
78192309, K2K1X7
78192309, K2W0A1
78192309, K2W1A3
78192309, K2W1A9
78192309, K2W1B1
78192309, K2W1C1
78192309, K2W1C2
78192309, K2W1C5
78192309, K2W1C6
78192309, K2W1G7
78192309, K2W1H2
78192309, K2W1H3
78192409, KOC1M0
78192409, KOC1P0
78192409, KOC1R0
78192409, KOC2A0
78192409, KOC2E0
78192409, K6H5R5
78192409, K6H5R6
78192409, K6H6E5
78192409, K6H6W8
```

Figure 20: Postal Code Lookup File

9 Reports

Two basic reports are used quite often during the de-identification process: a missingness report and a summary of the de-identification itself.

The missingness table in Figure 21 shows the level of missingness in the data after de-identification. This would be a reasonably intuitive measure of data quality for the researcher. As can be seen, only 1% of the records would have suppression in the postal code and 1 record would have suppression in the age field.

Niday0607_deid2: Risk Assessment

Settings View Risks **Missing Data** Reports

Missingness result for **Niday0607_deid2**

Maternal Postal Code	Baby Birth Date	Mothers Age (Years)	
(1292: 1.023%)	(0: 0%)	(1: 0.000%)	(1293 cells: 0.34%)
			(1292: 1.023%)
			(1: 0.000%)
			(1293 rows: 1.02%)

Close

Figure 21: Missingness table for the de-identified data. This includes missingness due to suppression as well as original missingness in the data.

The Microsoft Word report in Figure 22 shows the results of the risk assessment. This example shows the risk for the manually de-identified data. The Word report template can be modified or included in other documentation, and can also be used as a certificate for a research ethics board demonstrating the risk in the data set that is being requested.

Summary of Dataset				
Table name	Niday0607_deid2			
No. of valid records	126225			
No. of valid equivalence classes	25414			
Sampling Fraction	1.00			
Ignore Records with missing values	False			

Risk Assessment Results		
Re-identification Threshold	0.20	
	Risk Level	Records at Risk (%)
Prosecutor Risk	High	32048 (25.38%)
Journalist Risk	-	-
Marketer Risk	High	212 (0.16%)

Quasi-identifiers Selected				
Name	Type	# Equiv. Classes	Generalization	Weight
Maternal Postal Code	Character String	774	First 3 characters	1
Baby Birth Date	Character String	12	Month, Year	1
Mothers Age (Years)	Character String	9	Ranges of interval 5 starting at 20 until 50 (Coding: 19-, 51+)	1

De-identification Settings	
Type	Manual
Maximum Suppression	5%
Suppression Type	-

Figure 22: The Word report summarizing the risk in the data set.

10 Contact Information

For more information contact us at:

Privacy Analytics Inc.
800 King Edward Ave. Suite 3042
Ottawa, Ontario K1N 6N5
Canada

Tel: +1 613.369-4313

Fax: +1 613 369 4312

Email: info@privacyanalytics.ca

www.privacyanalytics.ca