Strategies for Integrating Administrative Data into Business Survey Activities

Kimberly Fyfe

January, 2018

1. Introduction

As National Statistical Offices (NSOs) seek efficiencies within their work it becomes natural to start looking towards obtaining data from government and non-government entities that are collecting administrative data that could be used in place of survey responses. This is especially important in the Caribbean region where most NSOs tend to experience fairly high refusal rates for their business surveys. As such, the following sections outline some considerations that need to be made before embarking on attempting to obtain access to administrative data as well as some possible strategies for incorporating administrative data into a survey process.

It is noted that the government entities mentioned above consists of Ministries, Departments and Agencies which will be referred to as MDAs.

2. Advantages and Disadvantages of Administrative Data

As with any change in activity, there will be advantages and disadvantages to moving towards using administrative data in place of survey data. The two key advantages for Caribbean NSOs to move towards using administrative data are:

i) To reduce costs associated with personal interviews to collect the data and
ii) To increase data available for estimation.

With respect to cost reduction, personal interviews are expensive due to the high costs associated with hiring, training and paying interviewers; especially when sending interviewers to conduct collection in rural areas. Additionally, head office personnel that are currently tasked with monitoring and managing the collection efforts can be redirected, at least partially, to perform other tasks. With respect to increased data availability, typically an administrative source includes data for units that feel obligated to report to the administrative source but not to the NSO, thus increasing the data that is available for estimation activities while reducing costs.

It is recognized that depending upon the situation, some activities could be an advantage for some offices, but a disadvantage for others. For example, administrative data could initially be more expensive if a special tool is required to capture paper versions of administrative data, separate data quality checks are required for reviewing the administrative data, or if the administrative data need to be transformed to fit the survey definitions. However, once these new tools and approaches are developed, the ongoing costs associated with the administrative data collection and processing should become much less expensive than conducting surveys. Additionally, once more data providers move toward electronic storage of information, data capture costs should disappear.

Timeliness is another aspect that could be an advantage or a disadvantage associated with administrative data. Depending on how the information is stored (paper documents that need to be

captured or electronic files) and when the information can be made available, the administrative data could be more or less timely than survey responses.

However, there are key aspects of administrative data that are always an advantage. The reduced burden on businesses to report the same information multiple times is an important advantage. Also, the administrative data, in addition to containing data to replace survey responses will often contain data about the businesses that can be used to update survey frames, such as whether it has ceased operation, as well as information about new names or addresses. Making use of administrative data also opens the door to working with the MDAs to improve the long-term quality of the data. Even though most NSOs are not allowed to feedback record level information, they can feedback generalized results of studies such as units within a particular industry code (ISIC) that tend to be systematically miscoded. Similarly, joint workshops can be developed in order to train both MDA and NSO personnel on key concepts such as ISIC coding. Additionally, in countries where the National Statistical System (NSS) is decentralized, working with the NSS partners can help to raise awareness of their role within the NSS, as well as the importance of the NSO's role in coordinating the nation's statistics (where it is the case).

Nevertheless, administrative data will also come with some drawbacks. One aspect that could be of concern is the quality of the incoming data. That is, if the NSO requires data items that are less relevant to the MDA, the quality could be lower. However, the primary concern is how well the concepts being collected for administrative reasons match what is needed for producing the survey estimates. When the two sets of concepts do not align well, the ideal solution would be to negotiate with the MDA to see if it they can modify what they are collecting to align with the data requirements or international standards. However, this is not always possible and it may be possible to process the incoming data to make it align with data requirements. This modification could be a simple linear adjustment such as increasing a sales figure to have it represent revenue (such rates should be based on analysis of previous data and different rates could be applied for different industries). Similarly, some values could be reduced or even combined to represent data items while others could be split to represent multiple data items. When the values do not align well but are highly correlated with survey requirements, it is also possible that such values can be used in imputation models, to model totals for groups of very small records that were intentionally excluded from sampling (called take-none strata) or as auxiliary information in a calibration strategy. When applying such adjustments to administrative data, assumptions are being made about the relationships between the administrative data and the survey requirements. As such, it is important to periodically review the adjustments to ensure that the assumptions still hold true. Otherwise, systematic biases can be introduced into the survey estimates.

One aspect that is often overlooked when integrating administrative data into a survey process is the need to have a correspondence between the administrative data and the frame. For example, when administrative data is received, it is important to know which frame units have been received and which units are new to the frame. Similarly, when requesting data for particular records it is important to be able to use the data provider's unique identifier when requesting data in order to ensure consistency over time. This is typically achieved by developing a record linkage strategy to create and maintain a correspondence table between the Business Register (BR) identifier and each data provider's unique identifier. It is recognized that record linkage can be quite cumbersome at the beginning, but becomes less once the linkage procedures are refined and the correspondence table is established such that the links are only being maintained. The ideal situation is to have a unique business identifier across all MDAs but this is rarely the case in practice.

Note:
When a correspondence table is used to identify new units from an administrative data source to be added to a BR, the NSO needs to check the new units very carefully against the existing BR units to ensure that these are truly new and not due to changes in unique identifier and / or business name. If frame update procedures create duplicates on the frame, the estimates could quickly become biased in an upward fashion; just as an under coverage of the population will generate an underestimation.

When working with administrative data, one needs to keep in mind that there could be systematic coverage issues created by the reasons for reporting (or not) a certain type of data. For example, there are often situations where certain types of businesses, such as small ones that are below a certain threshold, will not be required to report. On the other hand, depending upon the source being used to populate the frame, there could be duplicates due to accounting practices that are designed to minimize the amounts required to remit. Similarly, in some situations, groups of businesses could be very well (or even over) represented due to government incentives that can only be obtained through reporting to a certain MDA. In such cases, certain types of businesses receive an incentive to report, and other businesses may take creative latitude in reporting their activities in order to become eligible to collect a government incentive.

Occasionally, there could also be situations where a MDA updates the concepts that they are collecting or experiences issues in receiving their information. Such cases should be rare, but one needs to keep the possibility in mind and be prepared to react quickly to compensate for any possible negative impact. Having regular contact with the MDAs will provide the NSO with quick notification of any possible upcoming issues, thus maximizing the time to develop plans to mitigate the impact.

Most disadvantages associated with the use of administrative data can be minimized through the ongoing collection and review of the metadata associated with the administrative data. For example, the forms that are used for reporting the administrative data are typically a rich source of information about the coverage of the universe and definitions of the data items. Record layouts of the files being maintained by the MDAs could also provide insights into additional variables that are maintained by the MDAs, but not listed on the collection forms (a key example is that many MDAs maintain links of their records to the unique identifiers for other MDAs, which would greatly reduce record linkage efforts while increasing the linkage rates). Documentation describing the processing of the administrative data could be obtained from the MDA or written by the NSO in order to have a record of which variables have been reviewed for quality, updated in some fashion or are likely of poor quality. A key aspect of maintaining metadata is that it provides a concrete reference point for all present and future users of the data.

The following link provides another summary of the advantages and disadvantages associated with the use of administrative data as well as some considerations that should be taken into account. One key area that it covers is the various privacy concerns that an NSO should consider when initiating a record linkage project. As this document is an easy read, it is strongly recommended that it be referenced in order to obtain a further understanding of some of the aspects that should be considered when using administrative data:
http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm

The document in the above link is part of a larger document that covers quality issues related to the entire survey process, in the same easy-to-read format (5th edition of Statistics Canada's Quality Guidelines). As such, the NSO may want to reference it when redeveloping any aspect of a survey in order to determine additional improvements that could be implemented:
http://www5.statcan.gc.ca/olc-cel/olc.action?objId=12-539-X&objType=2&lang=en&limit=0

It is noted that even if there are some disadvantages to incorporating administrative data into a survey process, the advantages typically outweigh them, especially in an environment where survey response rates tend to be low and resources are scarce. Additionally, through time, most disadvantages can either be minimized or studied and accounted for when compiling the estimates.

3. Work Required to Gain Access

Gaining access to administrative data can sometimes take a bit of time and effort by the NSO. For example, it is often necessary to negotiate with MDAs in order to convince them that the NSO's Statistics Act grants access to the data, and to determine if there are any other regulations that might supersede the Statistics Act.

Once it has been determined that the NSO has permission to access the data, a Memorandum of Understanding (MOU) should be developed in order to specify how and when data will be delivered, and the contents of the deliveries. In terms of delivery, the frequency (monthly, quarterly, annually) and format (electronic or NSO personnel will visit the MDA to capture the pertinent information) will need to be negotiated. It will also be necessary for the NSO to discuss the available data in order to determine the list of information that will be requested in the MOU. The measures that will be implemented by the NSO to ensure the security of the data and how the NSO will be allowed to use the data should also be specified. It may also be desirable to include details of what will be returned to the data provider in return for receiving their data (special analysis reports, training, etc.).

A document focused on the various tasks associated with an enhanced access, use and maintenance of administrative data can be found at the following link. It provides information about setting up the proper infrastructure to deal with various aspects of receiving and using administrative data as well as templates and examples of documents that would be helpful when working with MDAs and their administrative data.
http://www.caricomstats.org/helpdesk/documentarycentre/Dom4/PRASC_Roadmap_to_improve_the_use_of_administrative_data_final_PDF.pdf

A document focused on how to assess quality, from both a data user perspective and a data producer perspective can be found at the following link:
http://www.statcan.gc.ca/eng/data-quality-toolkit

An assessment checklist to be completed by anyone contemplating the use of data produced by another organization is found at the following link:
http://www.statcan.gc.ca/eng/data-quality-toolkit/data-user

It is noted that data from more than one data provider can be used for a statistical program, and equally each data source could be used in more than one statistical program. For example, multiple sources could be used for BR development and maintenance as well as in compiling National Accounts products.

4. Partnerships with MDAs

Developing partnerships with the MDAs is important, especially in countries where the compilation of statistics is not centralized with the NSO. By developing a partnership, it helps each MDA to better understand their role within the NSS. If such relationships do not currently exist with certain offices, then developing a MOU is often a good place to start.

Drafting the MOU and discussing results of analysis are good reasons to initiate periodic meetings with the data providers. Depending upon the items to be discussed, such meetings could be monthly, quarterly or annually. The focus of such meetings could start with the development of the MOU and evolve through time to be more focused on the quality of the data.

For example, as the NSO receives administrative data, quality checks should be performed, and any general results and trends (anomalies) could be documented for discussion at the regular meetings. This information can then be used by the data providers to improve their processes and result in better data for their own programs.

Another way that offices can work together to improve the overall quality of the administrative data is to share training opportunities. Some areas where training could be beneficial include industry (ISIC) and product (CPC) classification coding, definitions associated with the administrative and survey data, determining what constitutes a valid address, general accounting practices to better understand the financial data, etc. By sharing the training opportunities, it helps to ensure that both offices have the same understanding of the concepts and can therefore work towards the same quality goals.

Developing partnerships with MDAs can go a long way in improving the overall quality of the estimates being produced by the NSS. The easiest way to develop such partnerships is to have regular meetings with the suppliers in order to establish a relationship of openness in discussing upcoming changes in the data and working together to improve the quality of the information for both the NSO and the MDAs.

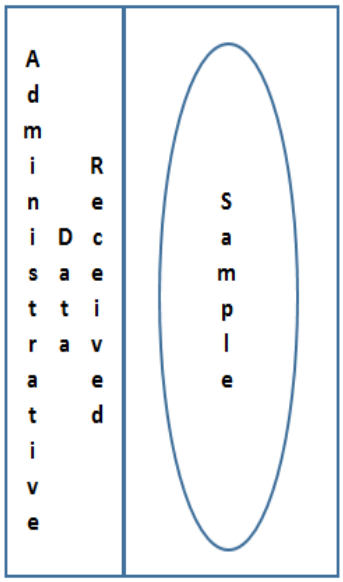
5. Using Administrative Data in Place of Survey Data

Depending upon when the administrative data become available, there are a few different schemes that can be used to incorporate it into the sampling and estimation procedures. The following sections outline some possibilities and the appendices provide details on how each type of record contributing to the estimates could be weighted.

> Caution:
> When using administrative data to replace survey data, one needs to keep in mind that the concepts from the administrative data need to align with those required by the survey. The alignment of concepts could be achieved by methods such as changing the survey definitions to be the same as the administrative data, changing what is collected by administrative data to match the survey needs or processing can be applied to the administrative data in order to have it align with the survey concepts. Having misaligned concepts will create a bias in the project's estimates. As more administrative data are used to replace survey data, this bias will grow over time thus creating false trends within the estimates.

a. **Detailed Administrative Data are Available Prior to Sample Selection**

In this situation, the administrative data are available prior to the sample being selected, which will allow the population to be split into two portions. The first portion already has data for all of the units and therefore can be measured directly without the use of a survey. The second portion does not have data and therefore needs to be surveyed in order to estimate this portion of the population. In small sectors and sectors with a fairly high coverage from the administrative data, it may be necessary to select all units into the sample (especially once the expected non-response rate is accounted for).



In sectors where the administrative data are known to cover a large portion of the total revenue (e.g. At least 80%), it could be possible to omit the sampling strategy and apply a macro adjustment to the administrative data estimates. In such cases, care will need to be taken to ensure that the macro adjustment is an accurate one; especially if the estimates will be released as level estimates, and are not only for input into macroeconomic tables (such as the Supply and Use tables).

This is the preferred strategy for using administrative data. In fact, it may be desirable to delay the release dates of the estimates in order to maximize the receipt of administrative data, and reduce the portion of the population that will be subjected to the sampling. As previously mentioned, if some sectors have enough current administrative data, it may be possible to forgo sampling and data collection activities, and make a minor adjustment to the administrative totals in order to produce the estimates.

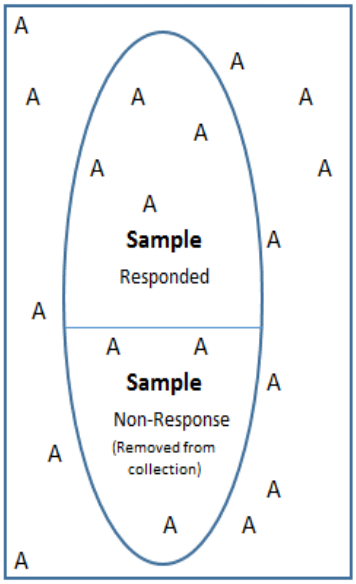
Advantages:

- This strategy makes use of all of the available administrative data and has the highest potential for reducing response burden as well as survey costs.
- Since the sample is redesigned each year, increases in the availability of administrative data could reduce the portion of the population that needs to be surveyed, thus possibly reducing sample sizes.

Disadvantages:

- If the units reporting the administrative data change a lot from one year to the next, volatility could be introduced into the survey estimates (due to a reduced overlap of units contributing to the estimates from one year to the next). However, this issue could be minimized by maximizing overlap of the sample with the previous year's responses (administrative and survey).
- Every year, all steps of the sample design will need to be conducted in order to take into account the different set of records being reported in the administrative data each year.

b. <u>Detailed Administrative Data become Available After Sample Selection</u>



When the administrative data become available after the sample has been selected but before data collection has started, the units that have been sampled and for which sufficient quality administrative data are available, can be removed from collection. Similarly for units which have administrative data become available during data collection, it is possible to cease collection attempts for such units. In such cases, it will be necessary to adjust the sampling weights for the survey responses in order to have them correctly represent the population. Details for doing this can be found in Appendix A, with the theory behind the adjustment presented in Appendix B.

To allow units to be removed from data collection efforts as quickly as possible, it may be desirable to have the MDAs send information to the NSO on an ongoing basis. For example, during the collection period, it may be desirable to receive weekly reports of units that have been received. For information that is received electronically, the incoming data can be assessed in a timely manner and units with complete and accurate administrative data can be removed from collection. If the administrative data is only available in paper format, it may be desirable to have the MDA send listings of received records so that the NSO can quickly identify which are in-sample and make arrangements to review the information for completeness and accuracy. By removing units from data collection as soon as possible, the interviewers and their supervisors will have more time to follow up the more difficult cases.

Even with a well-managed collection strategy, there will be some situations where both administrative and survey data are available for some units. For this reason, it will be necessary to decide which source is the most reliable and to make full use of all data available from that source.

Additionally, when there are both survey and administrative data for some units, it is possible to compare the two sets of data in order to obtain insight into possible quality issues of the two sources. For example, differences in values could be due to misaligned concepts, respondents misinterpreting the information being requested, data capture errors, reporting bias, etc.

<u>Advantages:</u>

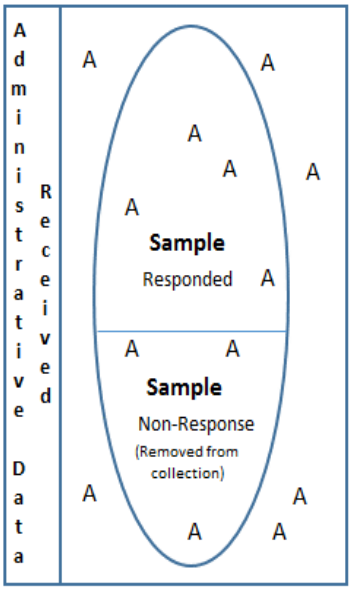
- All administrative data can be used.
- Collection costs can be significantly reduced while response rates and the timeliness of the survey results are increased.
- Some sampled units will have both survey and administrative data that can be compared to gain insight into the quality of the two sources.

<u>Disadvantages:</u>

- The interviewer assignments should be updated on a regular basis to remove units from collection.
- For the units where both survey and administrative data are available, it will be necessary to decide which values will be used.

c. **Detailed Administrative Data is Reported Monthly**

Given that businesses tend to have varied fiscal year ends, the administrative data sources likely receive filings on an ongoing basis. As such, NSOs receiving monthly (or quarterly) administrative files will be able to use a combination of the previously described strategies. Additionally, the work of capturing and reviewing the data could be spread out over a period of time so that it is not all being conducted under the pressure of an upcoming deadline.



To summarize the strategy, the population would be split into two portions (those with administrative data and those without). To produce an estimate for the units without available administrative data, a sample will need to be designed and drawn. When a file is received after selecting the sample, but before collection, some units can be removed prior to setting the interviewer assignments. Similarly, if monthly files are received during collection, additional units can be removed from collection.

Advantages:

- All administrative data can be used in an optimal manner.
- This strategy maximizes the potential for reducing response burden as well as survey costs.
- Since the sample is redesigned each year, increases in the availability of administrative data could reduce the portion of the population that needs to be surveyed, thus possibly reducing sample sizes.

Disadvantages:

- If the units reporting the administrative data source change significantly from one year to the next, volatility could be introduced into the survey estimates (due to a reduced overlap of units contributing to the estimates from one year to the next). However, this issue could be minimized by maximizing overlap of the sample with the previous year's responses (administrative and survey).
- Every year, all steps of the sample design will need to be conducted in order to take into account the different set of records being reported by the administrative data each year.
- The interviewer assignments should be updated on a regular basis to reflect the units that can be removed from collection.

6. Integration Strategy

It is recognized that a significant shift in the way survey estimates are produced can create new tasks that can be time consuming. As previously mentioned, new tools and strategies to capture and process the data may be required. Given that such tasks could differ for different groups of records, it could be desirable to phase in the use of administrative data. For example, when dealing with business surveys, it may be helpful to retain the current survey strategy in the industries where the administrative data coverage is low and would not offer much of an advantage. This would allow NSO personnel to focus on the areas that will see the highest benefits; perhaps there are domains that have enough coverage that data collection activities can be eliminated. Once the tools and

methods for domains with the highest impact are completed, work can then shift to the other survey domains. With experience, the transition will become quicker and easier, especially since fewer resources will be consumed by data collection activities.

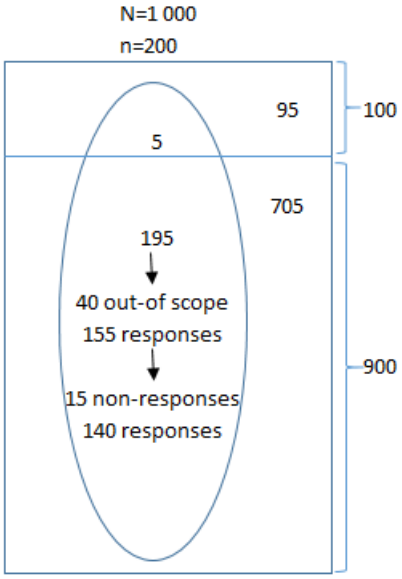
7. Concluding Remarks

Change is often difficult and met with resistance. However, the benefits of moving towards increasing the use of administrative data comes with many benefits and is quickly becoming the way of the future as many NSOs throughout the world move in this direction. The sooner that members of the NSS and other sources of data work together, the greater the savings will be within the statistical system, and the sooner that the team will be able to start working together to improve the administrative data and to release estimates that are comparable across all sources. Increased use of administrative data also reduces the burden on respondents, thus increasing goodwill and support of the statistical system generally. It is recognized that an initial investment in new tools, methods and concepts may be necessary. However, the long-term benefits for a country certainly have the potential to outweigh this investment.

Appendix A
Weighting Scheme

This appendix provides an unbiased weighting scheme that could be used when integrating administrative data with survey data, along with a practical example. Additionally, Appendix B provides the theory that underpins this approach.

As summarized in the graphic, this example considers a case where a stratum has a population of 1, 000 units and a sample size of 200 units where 100 units have associated administrative data and only 5 fall within the sample. This leaves 195 sampled units without administrative data, of which 40 units were out of scope (no longer operating, were found to be merged with another BR unit, etc.). Of the remaining 155 in-scope units, 140 have completed survey responses and 15 do not.



In this situation, there are **3** types of records that should have different weights. The **administrative data** represent themselves and therefore do not need to be weighted. Typically, the **out of scope** records have all been identified and therefore do not need to be adjusted for non-response, while the **survey responses** should be adjusted for non-response. The weighting for each of the 3 types of records is below.

All **administrative records** have a weight of 1.

The sampled **out of scope records** within this stratum have a weight of:
(derivation can be found in Appendix B)

$$w_{h\ i\ (out\ of\ scope)} = \frac{N_{h\ No\ admin}}{n_{h\ No\ admin}} = \frac{900}{195}$$

Where
$N_{h\ No\ admin}$ is the total number of units in the stratum that do not have administrative data, and
$n_{h\ No\ admin}$ is the total number of **sampled** units in the stratum that do not have administrative data.

The **survey responses** within this stratum have a weight of:

$$w_{h\ i\ (in-scope)} = \frac{N_{h\ No\ admin}}{n_{h\ No\ admin}} * \frac{n_{h\ In-scope\ sample\ No\ admin}}{n_{h\ Responding}} = \frac{900}{195} * \frac{155}{140}$$

Where
$n_{h\ In-scope\ sample\ No\ admin}$ is the total number of sampled units within the stratum that do not have administrative data and are in-scope for the survey, and
$n_{h\ Responding}$ is the total number of **responding** units within the stratum that will contribute to the estimates.

One check that this will produce an unbiased estimate is to ensure that all of the weights will sum to the population total:

$$w_{h\ i\ (out\ of\ scope)} * Number\ of\ out\ of\ scope\ records + w_{h\ i\ (in-scope)} * Number\ of\ responses$$

$$\frac{900}{195} * 40 + \frac{900}{195} * \frac{155}{140} * 140 = \frac{900}{195} * (40 + 155) = 900$$

(The remaining 100 units are accounted for by the administrative data.)

Appendix B
Derivation of the Weighting Scheme

This appendix provides the theory behind the weighting scheme presented in Appendix A.

The derivation of the weight starts by determining what needs to be weighted. In this case, it is the sampled units that are not in an administrative data file. Using the fact that the weight of a given unit is the inverse of its probability of being in that group, this derivation will start by deriving this probability:

Probability that the unit is in the sample given that it is NOT in the administrative data file, or

$$\mathrm{P}(i \in s \mid i \notin F)$$

Bayes' Theorem:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

where A and B are events and $P(B) \neq 0$.

- $P(A \mid B)$ is a conditional probability: the likelihood of event A occurring given that B is true.
- $P(B \mid A)$ is also a conditional probability: the likelihood of event B occurring given that A is true.
- $P(A)$ and $P(B)$ are the probabilities of observing A and B independently of each other; this is known as the marginal probability.

(Copied from https://en.wikipedia.org/wiki/Bayes%27_theorem)

Using Bayes' Theorem, the probability can be stated as:

$$\mathrm{P}(i \in s \mid i \notin F) = \frac{\mathrm{P}(i \notin F \mid i \in s)\, P(i \in s)}{P(i \notin F)}$$

Where:
$\mathrm{P}(i \notin F \mid i \in s)$ is the probability that the unit is not in the administrative data given that it is in the sample,
$P(i \in s)$ is the probability that the unit is in the sample, and
$P(i \notin F)$ is the probability that the unit is not in the administrative data.

In the example from Appendix A, the above values for the stratum are:

$\mathrm{P}(i \notin F \mid i \in s) = 195/200$

$P(i \in s) = 200/1000$

$P(i \notin F) = 900\ /1000$

Recalling that the weight for the units within this group is the inverse of the above probability, gives:

$$w_{hi} = \frac{P(i \notin F)}{\mathrm{P}(i \notin F \mid i \in s)\, P(i \in s)} = \frac{\frac{900}{1000}}{\frac{195}{200}\,\frac{200}{1000}} = \frac{900}{1000}\,\frac{200}{195}\,\frac{1000}{200} = \frac{900}{195}$$