(IJISE) 2018, Vol. No. 8, Jul-Dec e-ISSN: 2454-6402, p-ISSN: 2454-812X

# EXPLORATION OF REBUILDING LOOSELY STRUCTURED DATABASES TO INVESTIGATE INFORMATION FOR GENERATING LOCAL STATISTICS BY EMPLOYING R

**Saumya Gupta**

## ABSTRACT

*Local-level statistics are created from various information hotspots for arranging and essential leadership. A procedure is expected to rebuild existing databases to build an incorporated database for producing insights. To accomplish this goal, information structures of existing across the nation studies and enumeration were contemplated, and necessary information warehousing procedures in consolidating these information hotspots for developing an incorporated database were recognized. The means embraced to construct an incorporated database were united to figure the required system. To represent the veracity of the technique, across the Philippine- nation overview and evaluation databases were consolidated, and neighborhood-level measurements were produced for exploratory information investigation. The examination had the option to recognize highlights of existing databases that were utilized for the database rebuilding and working of an incorporated database. Similarly, the examination had the option to distinguish the most reduced level unit of perception to which information could be disaggregated which will likewise fill in as the reason for joining various information sources. From the getting the hang of utilizing a portion of the Philippine across the country overviews and registration, the examination has built up a system to access and join distinctive across the national databases that are truly isolated yet are identified with one another as far as their information engineering. This prompted building and incorporated database which could be a hotspot for the age of measurements that are helpful in exploratory information examination. The consequences of this examination are helpful for arranging and basic leadership. Besides, the investigation had the option to recognize approaches to distribution center the created measurements across existence for future information examination. Such ways were likewise joined in the created procedure. The detailed system of consolidating various databases into an incorporated database as a hotspot for producing nearby measurements is versatile to more databases for a more extravagant arrangement of created insights.*

## 1. INTRODUCTION

Information is gathered, examined, and afterward transformed into insights to get valuable in arranging and basic leadership forms. Information could be gathered through studies, censuses, vaults, managerial records, or even through web assets. Also, information is gathered for explicit purposes. When gathered, information is changed into an electronic configuration and put away in a database. Due to the various approaches to gathering just as their various purposes, the structure of their databases additionally will, in general, be unique. Censuses are led to gather information for every individual or unit of perception under investigation. In the Philippines, populace censuses are directed by the Philippine Statistics Authority (PSA) at regular intervals to acquire qualities of each Filipino occupant of the nation. The PSA likewise directs across the nation overviews for explicit purposes. For instance, Labor Force Survey (LFS) is being directed by PSA for business insights while Family Income and Expenditure Survey (FIES) is led by clockwork for money and consumption information to be utilized for the age of destitution measurements. The unit of perception in the LFS is an individual while it is a family unit in FIES. Libraries, for the most part, have a person as the unit of perception like a Cancer Patient Registry yet there are different vaults that have a gathering of people as the unit of perception like the Registry of Cooperatives where an association is the unit of perception. An informational index gathered from a specific source is put away in a database with a structure, adjusting to the unit of perception. In this manner, with this situation, informational indexes from various sources do have various structures. Be that as it may, these informational collections should be consolidated to produce progressively point by point insights which are required by the neighborhood government in their arranging and basic leadership forms. For instance, the FIES informational index could be joined with LFS informational collection and enumeration information to produce neediness insights at neighborhood government levels like district and city level. At present, informational indexes are consolidated when there is a need to create civil and city-level insights, and the way toward joining informational indexes is rehashed for another measurement to be produced. Thus, the undertaking gets repetitive as the coordinated database is not effectively accessible. A solution for this issue is to have a coordinated database within reach that is anything but difficult to keep up and update from databases with various structures.

Likewise, there is a need for a data warehouse that could store and maintain a big compiled data set that comes from several major database systems. A data warehouse is different from a data set that was developed and created for a specific purpose, and such a data set follows the concept of operational data. Nationwide data sets usually follow the concept of operation data. But what is needed is an integrated database is a data warehouse. This is supported in the discussion given in the Data Warehouse Concepts by the Office of Institutional Research and Academic Planning at Rutgers University.1 The comparison between operational data and a data warehouse suggests that the latter is more appropriate for an integrated database. From the same literature, it was stated that there are two main processes in designing a data warehouse. The first process is to determine the required information of the system and its metrics, while the second process is the actual development process, which is iterative in nature.

In this process, a better version of the data warehouse system is produced in every iteration. The iteration continues until the final version of the warehouse is acceptable to the users of the system. Once the design of the data warehouse was determined, the next step is to populate it with the required information. There are three steps to do this and these steps include extraction, transformation, and loading data. Extraction includes taking all the data out of the system before going to the next step of transformation. In this step, the idea is to change any information that needs updating and consequently fix any and all anomalies that might have or may occur during the transformation of data. Finally, the transformed information will then be placed back into the data warehouse which can be labelled as loading back the data to the warehouse. With a data warehouse, mining it to generate statistics for planning and decision making becomes much easier than working on different individual data sets. The process of data mining according to Ref.2 depends on the type of relationship that exists among the data in the system. The different types are labelled as a class, cluster, association, or in sequential pattern. A relationship that is classified as a class is when the information extracted is based on the previously stored data. An example is using the information in the customer's purchases in a restaurant chain to formulate some daily specials for a bigger profit. The second type of relationship is a cluster. Clusters are groups that are determined by their logical relationships or consumer preference, like when the data are organized in such a way that market segments or consumer affinities can be determined from them. The third type of relationship that can be considered when arranging information is an association. Association, as the name pertains to, collects and arranges data in such a way that the associations among the different data values are readily seen. An example is arranging different products in a store and having products that are normally sold together closer to each other for the convenience of the customer. The last type of relationship is the presence of sequential patterns. An example of this is using information that is mined to predict what the customer will be purchasing next based on the purchases that he or she has already made. Once the relationship of the data is determined, information and statistics processed and then presented for visualization to obtain the desired outputs.

Using these concepts in data warehousing and data mining, this study aims to present a strategy that could result in an efficient way of combining different database systems for an integrated database that could be warehoused and use the information in the database to generate local level statistics.

## 2. STRUCTURES OF SOME EXISTING DATABASE SYSTEMS IN THE COUNTRY

### 2.1. Population Census Data Set

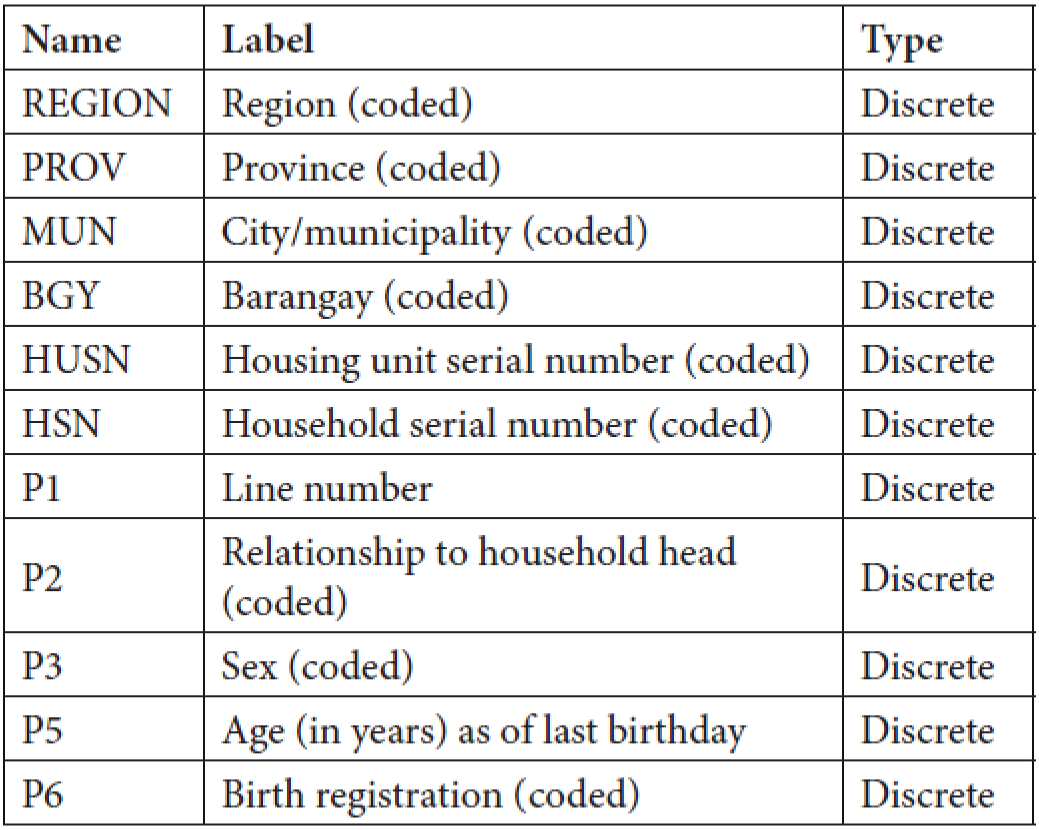
A populace statistics informational index has an individual inhabitant of the nation as the unit of perception. Segment attributes of the individual fill in as the factors in the informational collection. These segment qualities incorporate the topographical area of the home of the person just as the essential data like sex, age in years starting last birthday, relationship to the family unit head, incapacity, and birth enlistment. A large portion of the information is subjective in nature, yet all information is put away in numeric configuration since numeric codes are utilized rather than the subjective qualities. In the Philippines, the decennial evaluation covers the individuals, however, it additionally covers the lodging units in the nation that is the reason it is alluded to as Census of Population and Housing or CPH. Along these lines, the CPH of the nation can be considered to have two information frameworks dependent on their units of perceptions. One framework is with a unique individual as a unit of perception while the other framework has a lodging unit as the unit of perception. Mid-Decade Censuses are additionally directed in the Philippines however, this time it covers just populace enumeration, which is alluded to as the Census of Population or CP by the PSA.

In 2015, PSA directed the latest CP while in 2010; PSA led the latest CPH. A portion of the information things on every person as the unit of perception in the informational evaluation collections are recognized in the accompanying Table 1 with relating structure.

2.2. Nationwide Survey Data Set

Nationwide surveys are conducted for specific purposes. It takes only representative sample units of observations. The unit of observation may be different from one survey to another. Some surveys have households as the unit of observation while others may have a person or an establishment as the unit of observation. PSA is also the government agency mandated to conduct nationwide surveys. PSA uses a master sample for its integrated household surveys. Two of the surveys being conducted by PSA, which were considered in this study, are described below.

**Table 1.** Structure for some variables in the 2010 CPH data file



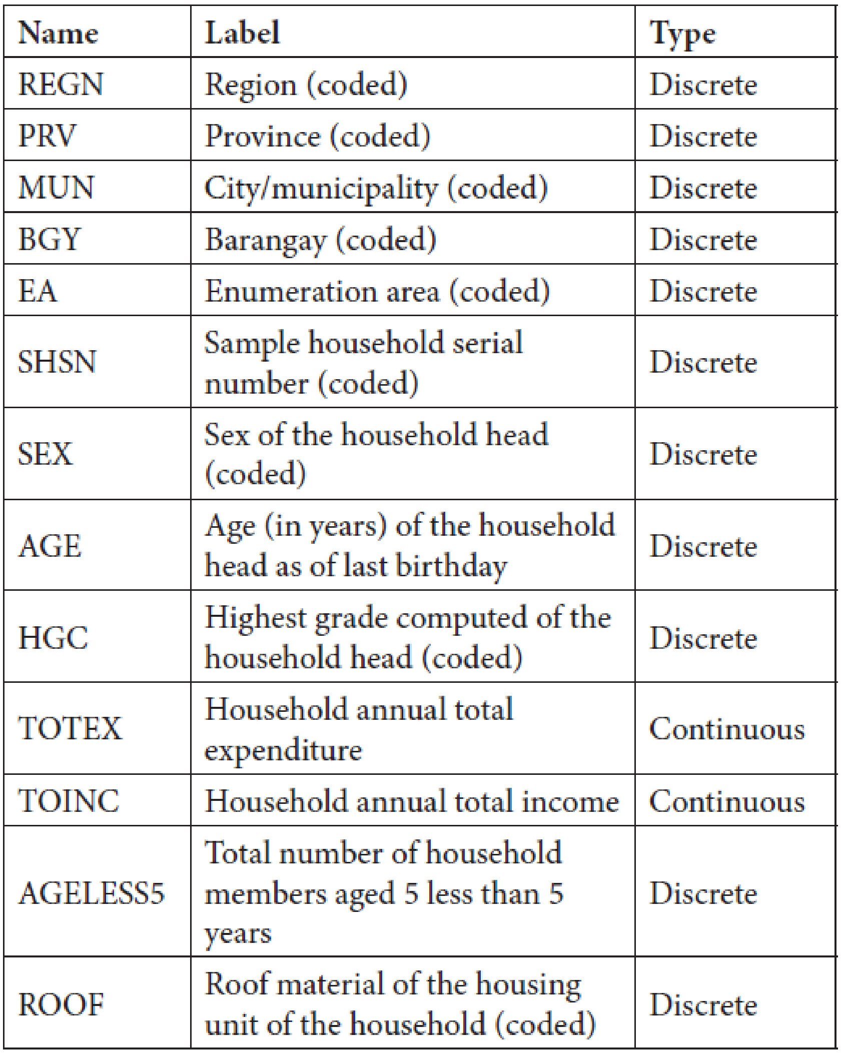
| Name | Label | Type |
|---|---|---|
| REGION | Region (coded) | Discrete |
| PROV | Province (coded) | Discrete |
| MUN | City/municipality (coded) | Discrete |
| BGY | Barangay (coded) | Discrete |
| HUSN | Housing unit serial number (coded) | Discrete |
| HSN | Household serial number (coded) | Discrete |
| P1 | Line number | Discrete |
| P2 | Relationship to household head (coded) | Discrete |
| P3 | Sex (coded) | Discrete |
| P5 | Age (in years) as of last birthday | Discrete |
| P6 | Birth registration (coded) | Discrete |

1) The Family Income and Expenditure Survey (FIES) is an across the country overview of families embraced at regular intervals by the National Statistics Office (NSO). It is the principle wellspring of information on family pay and use, which incorporate among others, levels of utilization by a thing of use just as wellsprings of pay in real money and in kind. As NSO (presently part of PSA) depicts it, the consequences of FIES give data on the degrees of living

and aberrations in the pay of Filipino families, just as their spending designs. The 2015 FIES utilized the ace example intended to give salary and use information that is illustrative of the nation and its 17 districts. A portion of the factors remembered for the FIES informational index is data on the qualities of the family unit head like his/her age, marital status, sex, most high instructive fulfillment, and other segment attributes. Since this is a family unit review, the attributes of the family as far as its synthesis were additionally watched. Similarly, the pay and use factors were seen at the family level, that is, full family unit pay and complete family unit use on explicit wares. A few qualities of the lodging unit of the family unit were additionally remembered for the arrangement of information things of FIES. Table 2 shows a portion of the information things and its structure in the FIES informational index. There are in excess of 300 factors being seen in FIES.3

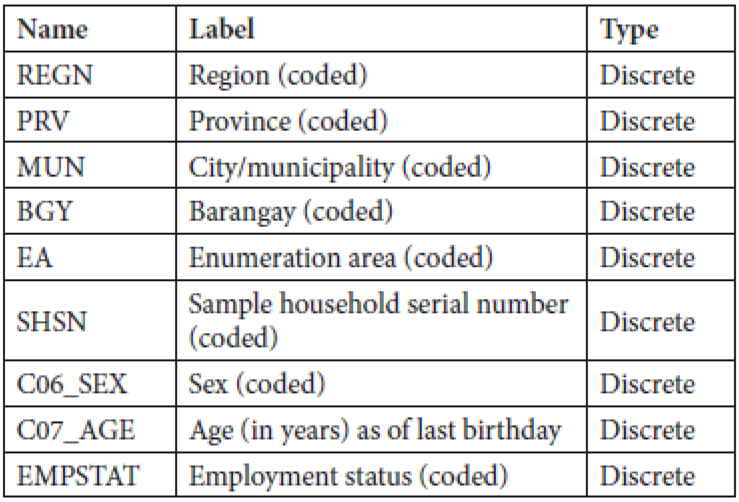
2) The Labor Force Survey (LFS) is another review which is being led by PSA each quarter of the year. LFS intends to give a quantitative structure to the arrangement of plans and detailing strategies influencing the work to advertise. As per PSA, the review is intended to give insights on levels and patterns of work, joblessness and underemployment for the nation, in general, and for every one of the regulatory locales, including areas and key urban communities. This overview incorporates factors for a family unit part matured 15 years or more, similar to relationship to the family unit head, age in years starting last birthday, conjugal status and most noteworthy evaluation finished. Among these family individuals, the individuals who were utilized were gotten some information about their principle movement/common occupation during the reference time frame, essential occupation, sort of business, class of specialist, nature of work, ordinary working hours out of every day during the previous week, all out hours worked during the previous week and whether he/she needs more long stretches of work. Then again, the individuals who had no activity/business were gotten some information about their pursuit of employment strategy and a number of weeks searching for work.4 See Table 3 for the portion of the information things and its structure in the LFS informational index.

**Table 2.** Structure for some variables in a FIES data file



| Name | Label | Type |
|---|---|---|
| REGN | Region (coded) | Discrete |
| PRV | Province (coded) | Discrete |
| MUN | City/municipality (coded) | Discrete |
| BGY | Barangay (coded) | Discrete |
| EA | Enumeration area (coded) | Discrete |
| SHSN | Sample household serial number (coded) | Discrete |
| SEX | Sex of the household head (coded) | Discrete |
| AGE | Age (in years) of the household head as of last birthday | Discrete |
| HGC | Highest grade computed of the household head (coded) | Discrete |
| TOTEX | Household annual total expenditure | Continuous |
| TOINC | Household annual total income | Continuous |
| AGELESS5 | Total number of household members aged 5 less than 5 years | Discrete |
| ROOF | Roof material of the housing unit of the household (coded) | Discrete |

**Table 3.** Structure for some variables in an LFS data file



| Name | Label | Type |
|---|---|---|
| REGN | Region (coded) | Discrete |
| PRV | Province (coded) | Discrete |
| MUN | City/municipality (coded) | Discrete |
| BGY | Barangay (coded) | Discrete |
| EA | Enumeration area (coded) | Discrete |
| SHSN | Sample household serial number (coded) | Discrete |
| C06_SEX | Sex (coded) | Discrete |
| C07_AGE | Age (in years) as of last birthday | Discrete |
| EMPSTAT | Employment status (coded) | Discrete |

3) Administrative Records

There are data sets which were collected by some institutions like a government agency. These data are usually based on reports provided by local department level to the national offices. The data sets are used in planning purposes as well as for monitoring of the performances of their constituent units. A good example is the data system of the Department of Education (DepEd) referred to as Basic Education Information System or BEIS. In their data system, the unit of observation is a school under the DepEd. The data items in this system are the characteristics of the school like location or address, district, division, total enrolment per grade or year level, total count of teachers, chairs, and books. This is the source of data in generating the basic education statistics of the country.

Another administrative data is coming from the barangay health centers and government hospitals and summarized by the Department of Health usually at the national, regional, and provincial level. The system called Field Health Services Information System or FHSIS provides basic health indicators like total count of hospitals, health workers, and recipients of vaccines as well as incidences of diseases. With its recent development, the electronic version of the FHSIS with barangays as unit of observation is being developed nationwide.

4) Registry

A registry can also be considered as a database which contains information about the observational unit. There are registries of individuals and there are also registries of cooperatives or organisations. Registries are also created for a specific purpose. An example is a registry of cancer patients which contain the basic information about a patient related to his/her disease. This kind of registry is usually confidential in nature and is used by medical practitioners and researchers for the purpose of finding a cure for the disease or at least finding ways to prevent the disease.

There are also other registries created to monitor the status and conditions of the members listed in the registry. This kind of registry like the Registry of Farmers is being used by concerned agencies in planning and decision making for programs envisioned to the betterment of the farmers. This registry has also an individual as the unit of observation. Registry of cooperatives and/or associations is also available like the registry of coffee growers' cooperatives and registry of associations of weavers in the country. But this kind of registry does not have individual as the unit of observation, but rather has an association or a cooperative as the unit of observation.

5) Discussion of the proposed strategy for the integrated database

The creation and maintenance of the integrated database is a process that extracts the different information from the different databases that may be useful in the study and consequently combine and condense them so that access to the data values would be fast and easy. Things that were considered in the creation of this integrated database include the compilation of crosschecked information and computation of frequently used values. In the compilation of crosschecked information, one has to get the information from the different databases and crosscheck them with existing entries in the other databases and copying them as is into the integrated database.

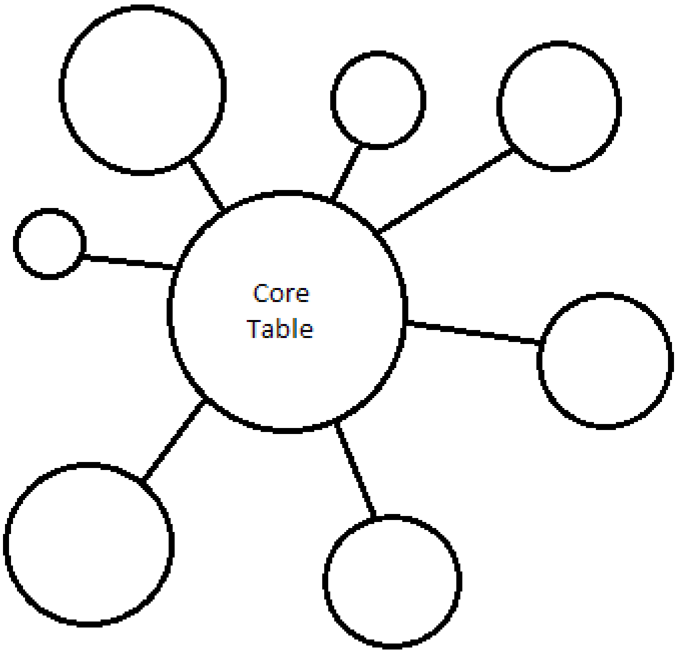


**Figure 1.** Star-shaped architecture.

In order to do this integration, it is required to first understand how the information is arranged in the source databases. As illustration, three source databases were used in the study. All of these databases have the same structure known as star-shaped architecture which basically means that each of them has a center table where most, if not all of the tables are related to. This architecture is illustrated in the following Figure 1. Because of this common star-shaped architecture, the different databases could be combined into one big data warehouse. The data warehouse has three main parts, namely: data tables, reference tables, and summary tables. The data tables consist of information that was extracted from the database sources. It basically holds a part of the information from the source database as it was laid out in the source database. Secondly, the integrated database has the reference tables which are used to create a uniform and clear coding scheme by using some of the available key features of the source databases. In this way, a standard coding scheme will be provided for the integrated database. Lastly, the summary tables include values that are usually needed by statisticians for data analysis, hence it should be able to handle and maintain the mined data from the source databases. The foreseen data analysis includes the generation of descriptive statistics to be placed in the summary table, the determination of relationship or association among the indicators found in the summary table and visualisation of the statistics generated.

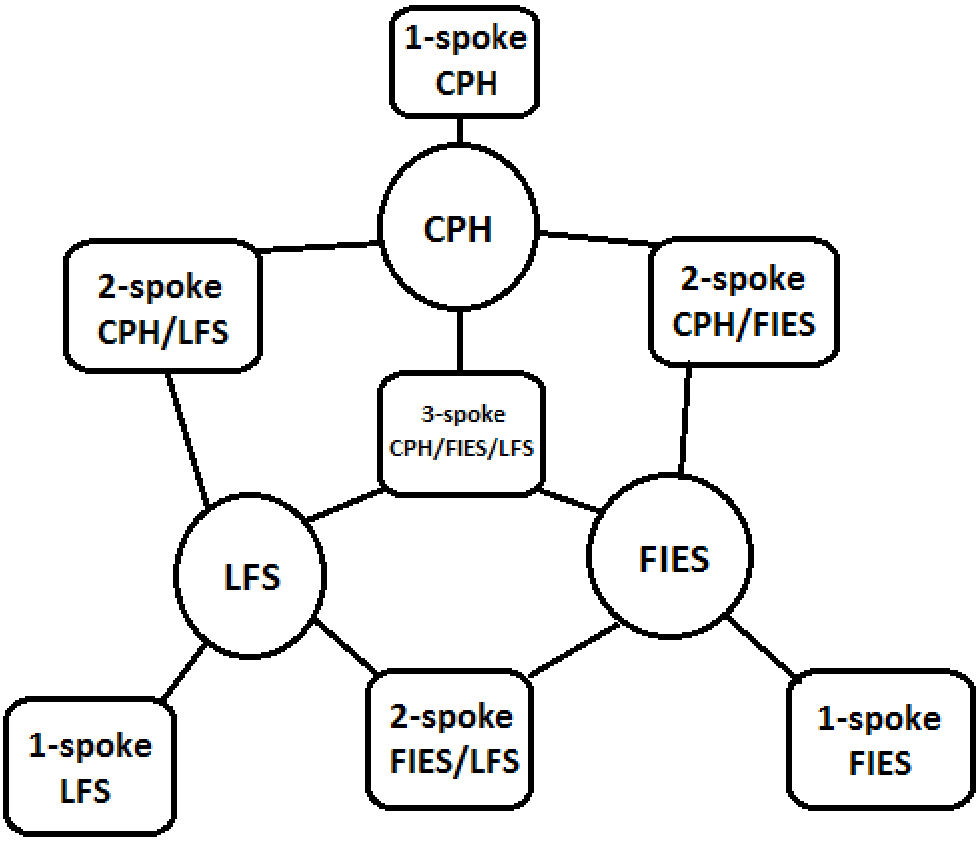


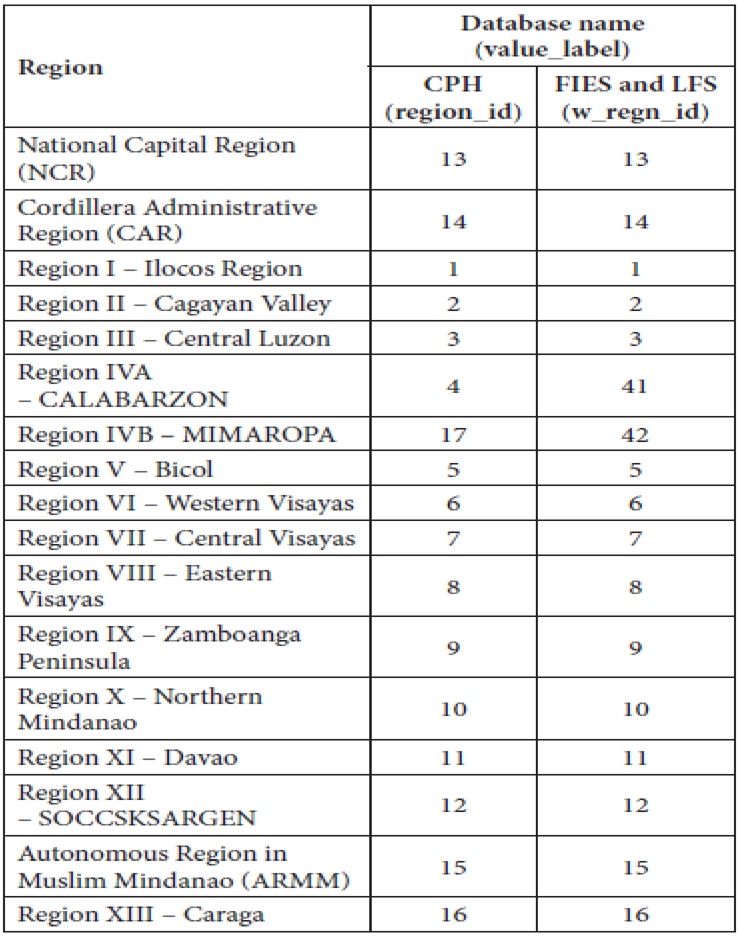
**Figure 2.** Integrated database architecture.

In the construction of the reference table of the integrated database, the core tables of the source databases should be considered. In this study, the three different source databases used to illustrate the strategy include the CPH, FIES, and LFS. To form an integrated database from these three source databases, the center of the architecture should have the core table for CPH, the core table for FIES, and the core table for LFS. The reference tables need to revolve around these three main databases and consequently cater to all of the three core tables. Thus, the main architecture of the integrated database will be something like the one shown in Figure 2. In the proposed data architecture of the integrated database, 1-spoke pertains to reference tables that are only used or only cater to one of the three core tables. This implies that in all of the available data in the integrated database, only one of the core databases have this information and therefore, are the only ones that will be using the reference table in the integrated database.

Likewise, a 2-spoke pertains to the reference tables that cater to two of the three core tables in the integrated database. This means that if there is a discrepancy between the two databases, these discrepancies will have to be resolved first before it will be able to proceed with the analysis of the information within them. Lastly, 3-spoke pertains to the reference tables that cater to all of the three core tables in the database. This means that discrepancies among all of the three databases will have to be resolved before analysis can proceed using the variables that use these types of reference tables. By unifying the reference tables in the three source databases, most of this information will be available for the three databases to use and all of the three will not be looking at redundant tables in the database.

In creating uniform reference tables, the common columns and information that are available in the three sources are determined and a uniform set of reference tables for all of them is to be created. With this, all of the data from the three different databases are then placed into one data warehouse with the differences edited so that it will now be able to take the information from the unified reference tables. Examples of the tables that are common in the different databases include the variables labelled as region, province, and sex, relationship to household-head, FIES indicator, and urban rural classification among others. As for the tables that needed a transformation before transferring, one of the best examples is found in the Region Table that is available in all of the databases. The CPH had a different reference table from the other two databases. This implies that it needed to be transformed before it has to be transferred to the integrated database. The two reference tables were both transformed to be uniform, taking the form of the reference table for the CPH. These two reference tables are found in the following Table 4.

**Table 4.** Region table of the CPH compared to that of FIES and LFS



| Region | Database name (value_label) | |
|---|---|---|
| | CPH (region_id) | FIES and LFS (w_regn_id) |
| National Capital Region (NCR) | 13 | 13 |
| Cordillera Administrative Region (CAR) | 14 | 14 |
| Region I – Ilocos Region | 1 | 1 |
| Region II – Cagayan Valley | 2 | 2 |
| Region III – Central Luzon | 3 | 3 |
| Region IVA – CALABARZON | 4 | 41 |
| Region IVB – MIMAROPA | 17 | 42 |
| Region V – Bicol | 5 | 5 |
| Region VI – Western Visayas | 6 | 6 |
| Region VII – Central Visayas | 7 | 7 |
| Region VIII – Eastern Visayas | 8 | 8 |
| Region IX – Zamboanga Peninsula | 9 | 9 |
| Region X – Northern Mindanao | 10 | 10 |
| Region XI – Davao | 11 | 11 |
| Region XII – SOCCSKSARGEN | 12 | 12 |
| Autonomous Region in Muslim Mindanao (ARMM) | 15 | 15 |
| Region XIII – Caraga | 16 | 16 |

With the reference tables transformed, the next to transform is the data table which contains the entries for the three databases. Three tables were created which indicated the source of each information. These three tables were extracted from the three databases and were transformed to be placed into the database that is going to be used for the analysis. The transformation includes having all of the entries edited so that the tables will have the correct information within their rows like changing the values of the regions per row in the FIES and the LFS tables. The inclusion of the raw information of the three databases in the data warehouse represents the first part of the two-step process. This first part of the process created a local copy of all the information that are available in the three databases which then gives the ability for the data warehouse to do the computations independent of the database sources.

As for the last part of the integrated database, the focus is the set of computed values from the three databases. This mainly includes simple common derived values such as total number of entries that have a specific value recorded in the tables. Some examples of these computed values are the total number of males and females in the FIES database, the total number of people who are married in one region from the CPH database, and other simple tally of values. With this kind of derived information available in the integrated database, further computations could be done that might be needed for the analysis part. This part of the integrated database creation represents the second part of the two-step process. The result of the two-step process is the completed integrated database system which is an input to data warehousing. It contains not only a unified version of the three database sources, but also some derived values and attributes of the three databases. The database in the warehouse is now available for data mining. Mining the system means traversing the integrated database to create and maintain summary tables in the database. The summary table contains statistics at the smallest unit of observation or lowest level of disaggregation that is common for all source databases. The recommended statistics is in the form of total count or sum of the values. From the sum and/or total count other statistics like average and proportion could easily be computed. For a measure of dispersion or variability, the sum of the square of the values is also recommended to be part of the summary tables. For the source databases in the illustration, a barangay is the recommended smallest unit of observation or lowest level of disaggregation to use. Total counts are to be computed for discrete variables while the sum of the values and sum of the square of the values are the statistics to be computed at the barangay level. These statistics are then placed in the summary tables which could be used for data analysis. Table 5 contains variables in the summary table, their description and from which database they were derived from. Using the information in the summary table, descriptive statistics can then be computed. Likewise, the association of the variables could also be explored graphically or even measuring the degree of the association. This information can help planners and/or decision makers in their course of actions to take. Scatter plots like the plots shown in Figures 3 and 4 illustrate the degree of association between two variables of interest could be generated from data found in the summary tables. For example, based on these scatter plots, the associations shown have direct relationships. Because of this, some claims can be further looked into. For this particular example, one certain claim that may warrant a closer look is the association of having college graduates choosing to go abroad for work and college undergraduates choosing to stay either because they lack the skills to be qualified for work abroad or for some other reasons.

Similar to this kind of analysis is what could be done when one has the summary tables obtained from an integrated database. A closer look at the numbers could help planners or decision makers or even government officials in their decision-making process. Having these statistics to base the plan or the decision is an advantage and a well-designed integrated database could support this process are Table 5.

**Table 5.** Variable name and description in the sample summary table

| **Variable name** | **Variable description** | **Source database** |
|---|---|---|
| Regn | Region | CPH/FIES/LFS |
| Prov | Province | CPH/FIES/LFS |
| Mun | Municipality | CPH/FIES/LFS |
| Bgy | Barangay | CPH/FIES/LFS |
| countNG | Total count of persons who did not have formal education | CPH |
| countEU | Total count of persons who started elementary education but did not graduate | CPH |
| countEG | Total count of persons who completed elementary education | CPH |
| countHU | Total count of persons who started secondary education but did not graduate | CPH |
| countHG | Total count of persons who completed secondary education | CPH |
| countCU | Total count of persons who started tertiary education but did not graduate | CPH |
| countCG | Total count of persons who completed tertiary education | CPH |
| countYes | Total count of persons who are overseas workers | CPH |
| countNo | Total count of persons who are not overseas workers | CPH |
| texpendSum | Total household expenditure | FIES |
| texpendSum2 | Sum of squares of household expenditure | FIES |
| tincomeSum | Total household income | FIES |
| tincomeSum2 | Sum of squares of household income | FIES |
| countNES1 | Total count of persons who are employed | LFS |
| countNES2 | Total count of persons who are unemployed | LFS |
| countNES3 | Total count of persons who are underemployed | LFS |

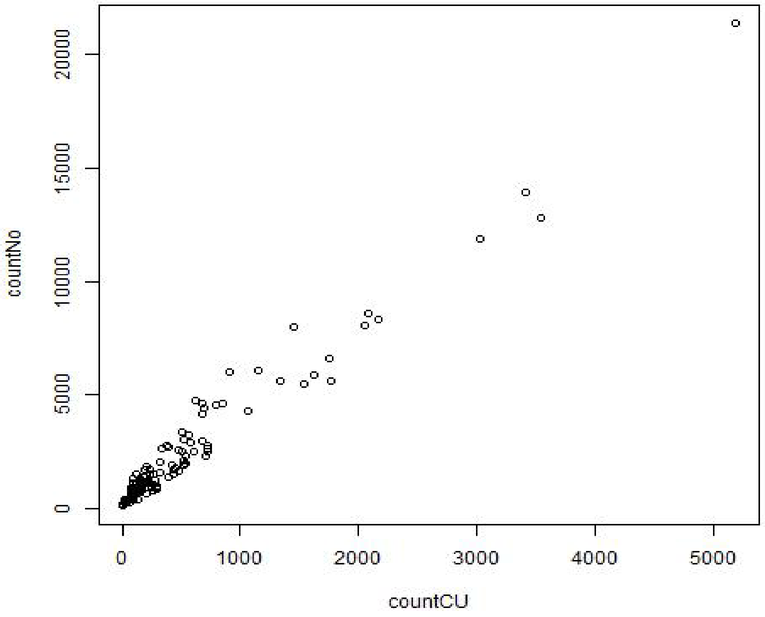


**Figure 3**. Scatter plot of the total number of college undergraduate and total count of overseas work.

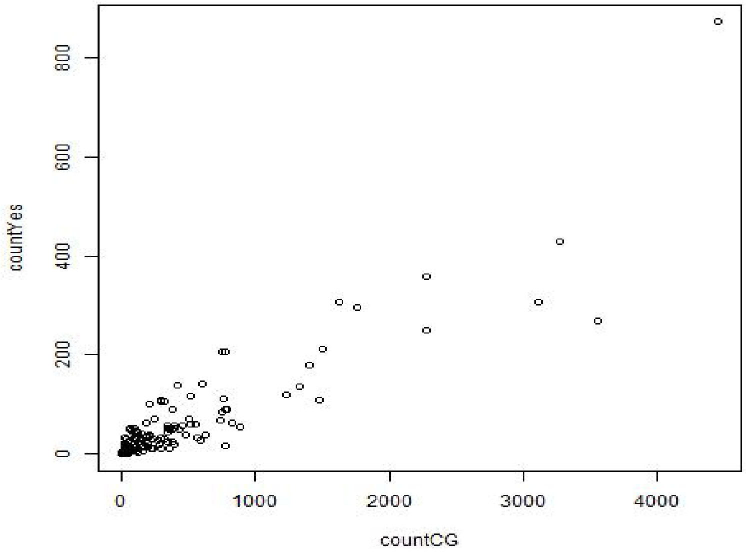


**Figure 4.** Scatter plot of the total number of college graduate and total count of overseas work.

## 4. CONCLUSION

With all of the systems designed and implemented, it is safe to say that there is a possible way to integrate and use the information from the different data sources. Using technology that is both free and accessible, the creation and maintenance of the integrated database is plausible and much needed. The benefits it brings are many and the reasons for it to exist support this. The integrated database serves as a tool for organizing data to generate statistics. Furthermore, an integrated archive will be useful in the future when it comes to studying what has happened in the past to determine solutions for problems or even predicting trends in the future.

The database itself is a sustainable and maintainable archive of information. Improvements can still be done on different parts of the system. Although, on its own, it is a working prototype that can do some of the more tedious tasks to help in the work of different professionals. The main improvement can be placed in either the choice of a database management system or the implementation of the management of information. As for the setting up of the combined data sets for analysis, the summary tables have been created and all of them are catering to the general needs of professionals to explore and use in different types of analysis with regard to the three source databases. The summary tables have paved a way for the professionals to not only create and explore more information but to also give them a more detailed way of data analysis by setting up the variables using different statistical techniques. In addition, with the compilation and creation of the visualizations of the data, it is very easy to see that there is a lot of potential in looking at the trends they may show. With the creation of the visualizations, some of the more difficult to see trends can be more visible to the users. With these trends being visible, users will now be able to formulate their own observations and analysis of the data.

## 5. REFERENCE

1) Data warehouse concepts. https://www.tutorialspoint.com/dwh/dwh_data_warehousing.htm.
2) Data mining: what is data mining?, https:// ucarecdn.com/1bc035bc-0f0e-454e-a9b0-fe72beacd960/
3) Family income and expenditure survey (FIES), https://psa.gov.ph/tags/family-income-and-expendituresurvey.
4) Technical notes on labor force survey (LFS), http://www.psa.gov.ph/content/technical-notes-laborforce- survey-lfs.