RESIDENTIAL BUILDING STOCK ASSESSMENT

Database User Manual

Table of Contents

Introduction

In fall 2017, the Northwest Energy Efficiency Alliance (NEEA) completed its second Residential Building Stock Assessment (RBSA). A broad, regional study, the RBSA divides 2016–2017 building stock into three housing types: single-family homes, manufactured homes, and multifamily buildings. Since its inception, NEEA has conducted research on northwest residential building stock characteristics, conducting its first comprehensive, regionally representative study in 2011–2012.

To serve as a resource for efficiency planners and program designers in the region, a publicly available database contains data gathered through the study. These data were collected using three primary data sources: participant surveys, home and equipment characteristics (gathered during home visits by trained technicians), and historical energy consumption data provided by utilities in the region.

This user manual addresses three areas:

- **Database Overview and Relationships**: These sections provide guidance on how to access data collected during the RBSA
- **Estimation:** This section provides database users with guidance regarding formulas they should use to estimate RBSA metrics for strata or subpopulations of interest
- **Importing into Microsoft Access:** For users who wish to import data into an MS database, these sections provide the script and necessary steps.

Please direct questions and inquiries regarding these data to NEEA’s Market Research team.

Database Overview and Relationships

This section includes several resources that may be helpful for users of the database, including:

- A synopsis of the database structure and relationships
- Definition of the objects within each table in the database
- An entity relationship diagram (ERD) that visually depicts the relationships between the datasets housed within the database

In addition to this document, a data dictionary is incorporated as a standalone table within the database. The data dictionary provides definitions for each field in the database as well as sample data from the database.

Sites, Buildings, and Residences

The RBSA database houses all data gathered during the study, including that for manufactured homes, single-family homes, multifamily residences, and multifamily buildings. Within the database, data are attributed to a Site or a Building. Building data include information collected for nonresidential portions of a multifamily building. Site data include information collected for the entirety of a single-family or manufactured home, and for residences visited at a multifamily building. The database denotes data associated with a Site by a CK_SiteID or PK_SiteID beginning with “SITE”; the database denotes data associated with a Building by a CK_SiteID or PK_SiteID beginning with “BLDG.”

Definition of Common Terms

- **N/A**. Indicates a datapoint is not applicable. For example, “Heating Fuel” is not applicable to central air conditioners. In some cases, the data collection tool only shows fields if a certain value is entered. Where the data collection tool does not show fields to a technician, the database value is “N/A.” For example, if a technician indicated that a wall was not insulated, they would not be asked to enter insulation information. Similarly, if a participant skipped a question in the recruitment survey, subsequent follow-up questions would be “N/A.”
- **Unknown**. Indicates a datapoint is applicable, but its value is unknown. For example, an unknown value for “Wall is Insulated” indicates that the technician could not tell if a wall was insulated. For survey and interview responses, “Unknown” indicates that the respondent didn’t know the answer to the question.

Numeric Fields

The RBSA data collection effort included a large number of numeric fields that are helpful to quantify or describe the observed equipment or characteristic. However, it is not always possible for field staff to record a numeric value, and instead they record a value of “N/A” or “Unknown.” In the RBSA database, these data are separated into two separate columns. One column contains numeric values and NULL values. The second column is a complimentary match to the first column: where the first column has numeric values, the second column has NULL values, and where the first column has NULL values, the second column has a value describing the NULL. An example of this relationship is shown in Table 1.

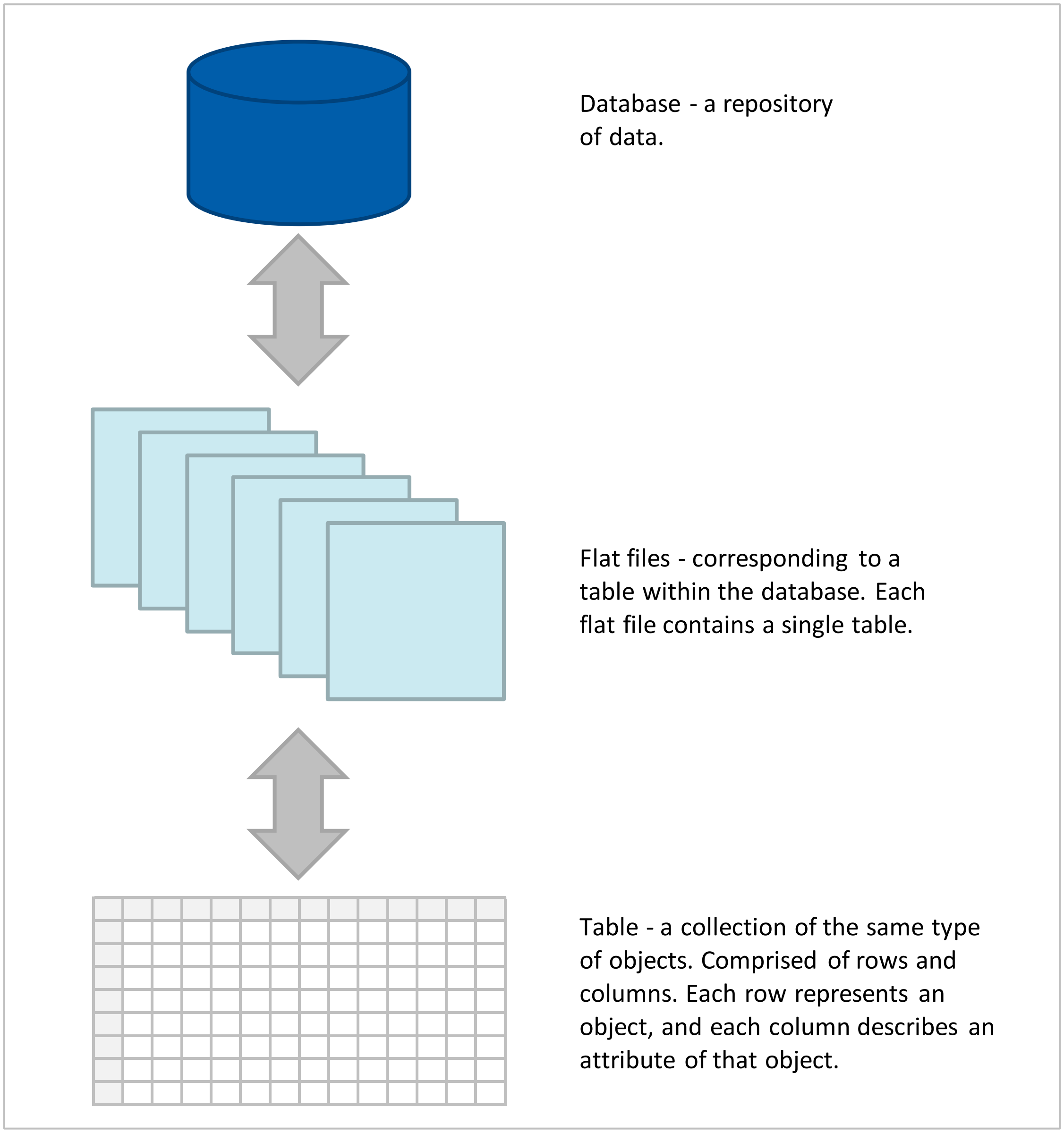
Table 1. Numeric and Complimentary Descriptive Column

Wattage	Wattage_notes
60	
	Three-way wattage: 50-100-150
60	
	Unknown

Database Overview

The RBSA database is a relational database provided as a collection of CSV “flat” files. A relational database presents information in a collection of tables using rows and columns. Each table serves as a collection of objects of the same type, with each file consisting of a single table. Columns within each table represent the attributes for that set of objects. Figure 1 illustrates the database concept.

Figure 1. Relational Database Components



Database - a repository of data.

Flat files - corresponding to a table within the database. Each flat file contains a single table.

Table - a collection of the same type of objects. Comprised of rows and columns. Each row represents an object, and each column describes an attribute of that object.

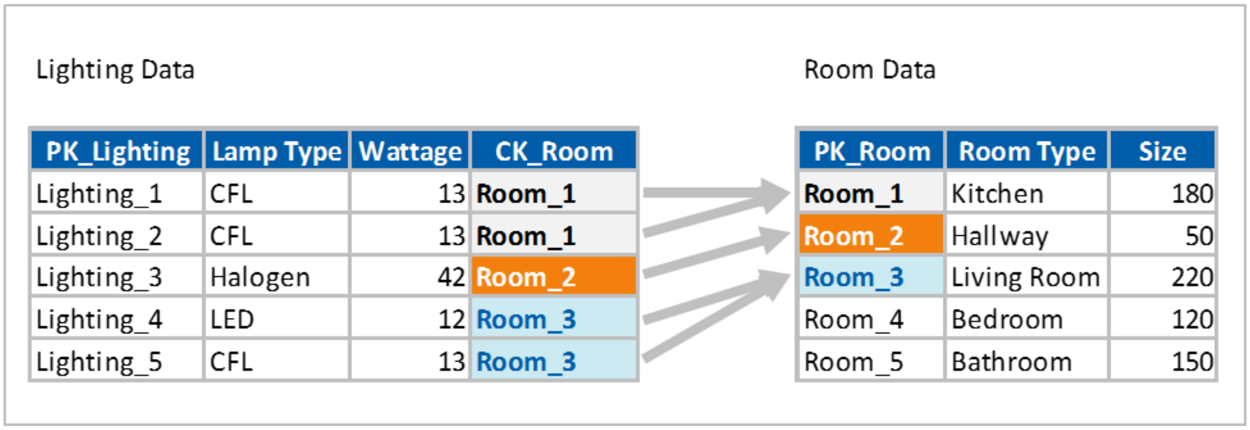
RBSA Database

Tables in the RBSA database relate to one another using keys or unique identifiers. Each row in a table has a unique primary key (PK), making it possible to reference that specific object, and may include other relational keys (CKs) that make it possible to relate to other tables in the database.

For example, each room surveyed during the RBSA provided a variety of data, which included information about the room itself (e.g., room type and size) as well as information about other objects within the room (e.g., lighting, appliances, and windows). Many objects have their own set of attributes independent of the room or space. For example, lighting characteristics (e.g., type of fixture, lamp, lamp wattage) are independent of the space itself. However, knowing the type of room in which the fixture is located allows the user to calculate properties such as lighting power density or the average number of lamps per room. (For this example, it makes sense to store lighting information separately from room data, thus creating a relationship between the tables.

Figure 2 shows an example of two related tables: the Lighting table and the Rooms table.

Figure 2. Example: Room and Lighting Relationships



Lighting Data

PK_Lighting	Lamp Type	Wattage	CK_Room
Lighting_1	CFL	13	**Room_1**
Lighting_2	CFL	13	**Room_1**
Lighting_3	Halogen	42	**Room_2**
Lighting_4	LED	12	**Room_3**
Lighting_5	CFL	13	**Room_3**

Room Data

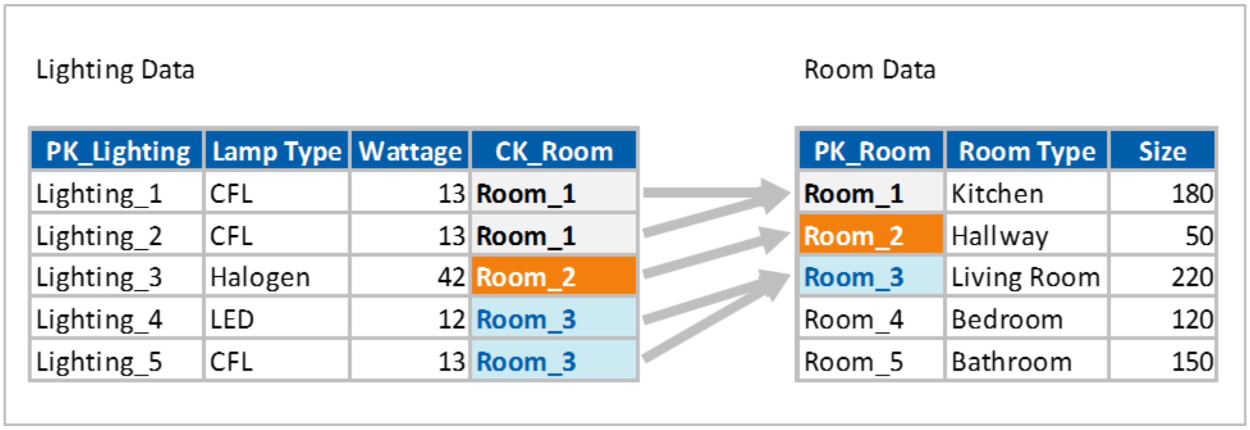
PK_Room	Room Type	Size
Room_1	Kitchen	180
Room_2	Hallway	50
Room_3	Living Room	220
Room_4	Bedroom	120
Room_5	Bathroom	150

Both the Lighting and Rooms tables have their own set of PKs and attribute data, with each row using a unique PK. Additionally, the lighting dataset includes a set of CKs, with values matching the PKs in another table (i.e., each light fixture is assigned to a room).

Types of Relationships

Two primary types of relationships exist within the database: one-to-one relationships; and one-to-many relationships. In a one-to-one relationship, both tables have a single record on either side of the relationship. In a one-to-many relationship, the PK table contains only one record, which may relate to none, one, or multiple records in the related table. Figure 3 and Figure 4 provide examples of these relationships.

Figure 3. Many-to-One Relationship



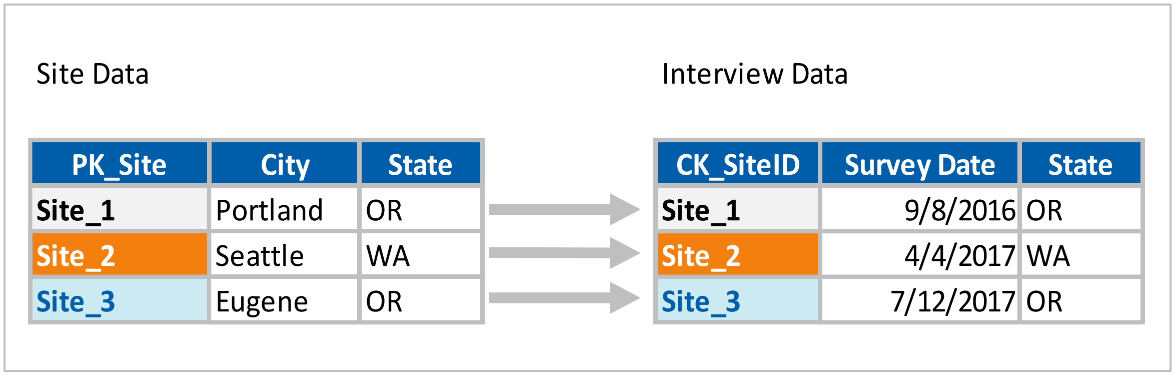
Lighting Data

PK_Lighting	Lamp Type	Wattage	CK_Room
Lighting_1	CFL	13	**Room_1**
Lighting_2	CFL	13	**Room_1**
Lighting_3	Halogen	42	**Room_2**
Lighting_4	LED	12	**Room_3**
Lighting_5	CFL	13	**Room_3**

Room Data

PK_Room	Room Type	Size
Room_1	Kitchen	180
Room_2	Hallway	50
Room_3	Living Room	220
Room_4	Bedroom	120
Room_5	Bathroom	150

Figure 4. One-to-One Relationship



Site Data

PK_Site	City	State
Site_1	Portland	OR
Site_2	Seattle	WA
Site_3	Eugene	OR

Interview Data

CK_SiteID	Survey Date	State
Site_1	9/8/2016	OR
Site_2	4/4/2017	WA
Site_3	7/12/2017	OR

Lighting and Rooms have a one-to-many relationship: one room record may correspond to none, one, or multiple lighting records. Alternatively, Sites and Interviews have a one-to-one relationship: one record exists on either side of the relationship.

Database Contents

Object Definition

As discussed previously, each database table serves as a collection of objects of the same type. Columns within each table represent the attributes for that set of objects. Table 2 defines the type of objects housed within each database table or table group.

Table 2. Object Definition by Table or Table Group

Database Table or Table Group	Definition of Objects (Line Items) within the Table
SiteDetail Table	Each line item corresponds to a single entity: either a residence or a building. This table contains high-level information about the entity such as: State, address, strata, and case weight. This table also defines relationships between multifamily buildings and multifamily residences.
SiteOneLine Table	A single residence (e.g., a single-family home, manufactured home, or residential unit within a multifamily building).
BuildingOneLine Table	A single multifamily building.
Site Interview Table Group	A resident interview.
Building Interview Table Group	A property manager interview.
Envelope Table Group	A unique structural component, with each component recorded during on-site visits displayed as an individual line item.
Mechanical Table Group	A unique piece of mechanical equipment. Each piece of equipment recorded during an on-site visit displayed as an individual line item. In some cases, a quantity field is used to indicate the presence of multiple, identical equipment.
Room Table	A single room.
Water Table	A unique water equipment entry (e.g., Faucet, Showerhead, Bathtub).
Window Table	A unique window or door configuration entry. Each entry may represent one or more individual windows with the same characteristics. A quantity field is provided.
Appliance Table Group	A unique appliance entry, with each piece of equipment recorded during an on-site visit displayed as an individual line item. Each entry may represent one or more individual appliances with the same characteristics. A quantity field is provided for large unusual loads.
Lighting Table	An entry that represents a unique light bulb type/base type/shape/wattage entry. Entries with the same CK_Lighting_ID are on the same fixture. Each entry may represent one or more individual light bulbs with the same characteristics. A quantity field is provided.

Entity Relationship Diagram

An entity relationship diagram (ERD) that visually depicts the relationships between the datasets housed in the database is provided in Figure 5. Relationships between tables are indicated by grey connections between tables. The relational keys and an overview of the relationship types (one-to-one, one-to-many) are also included in the diagram.

Figure 5. Entity Relationship Diagram

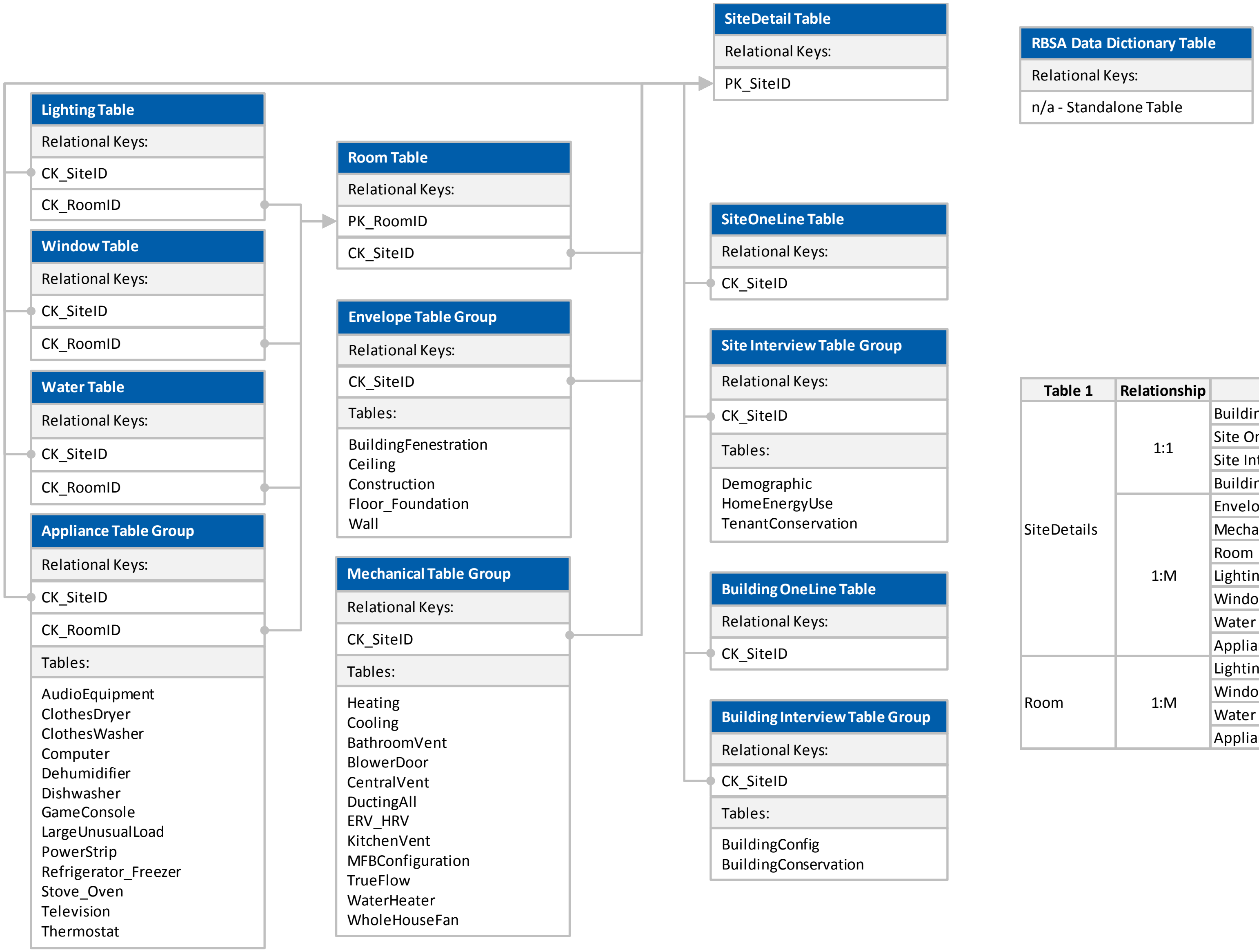


Table 1	Relationship	Table 2
SiteDetails	1:1	Building OneLine
		Site OneLine
		Site Interview Group
		Building Interview Group
	1:M	Envelope Group
		Mechanical Group
		Room
		Lighting
		Window
		Water
		Appliance Group
Room	1:M	Lighting
		Window
		Water
		Appliance Group

Aggregating Data

To aggregate or “roll-up” data, users use relational keys in each table. This section provides an overview of the data aggregation process and illustrates the process through an example.

Note: These instructions are for unweighted data only. Though the aggregation process is similar for weighted data, applying appropriate weighting to each line item requires additional steps. See the Estimation Section, below.

Aggregating Data to the Room Level

The RBSA database captures some data on a room-by-room basis, including information about lighting, appliances, windows, faucets, and showerheads. To understand total quantity and averages for a site, cells must be summed or averaged. For instance, if the user wants to know how many windows each room has, they need to sum all window quantities where the CK_RoomID matches a specified value. If the users wants to determine each room’s total window area, they sum the product of Window Quantity and Window Area for all rows with the same CK_RoomID.

Table 3. Sample Window Data

PK_Window	CK_SiteID	CK_Room	Window Area (sq. ft.)	Window Quantity
Window_1	Site_1	Room_1	9	1
Window_2	Site_1	Room_1	4	2
Window_3	Site_1	Room_1	12	1
Window_4	Site_1	Room_2	6	2
Window_5	Site_1	Room_3	16	1
Window_6	Site_1	Room_3	16	1
Window_7	Site_1	Room_3	3	3

Summing the Window Quantity across each room results in the following totals.

Table 4. Sum of Window Quantity by Room from Sample Window Data

CK_Room	Window Quantity
Room_1	1 + 2 + 1 = 4
Room_2	2 = 2
Room_3	1 + 1 + 3 = 5

Summarizing the window area proves slightly more complex, as that window area represents each window. If the Window Quantity is greater than one, the user must sum the product of the window area, multiplied by the window quantity.

Table 5. Sum of Window Area by Room from Sample Window Data

CK_Room	Window Area (sq. ft.)
Room_1	(9 x 1) + (4 x 2) + (12 x 1) = 29
Room_2	(6 x 2) = 12
Room_3	(16 x 1) + (16 x 1) + (3 x 3) = 41

Aggregating Data to the Site Level

Steps for aggregating data to the site level are roughly the same as those used for aggregating data to the room level. The primary difference between room-level and site-level aggregation is aggregation occurs based on CK_SiteID rather than CK_RoomID.

Revisiting the above example, the total quantity of windows at Site_1 is shown as follows.

Table 6. Sum of Window Quantity by Site from Sample Window Data

CK_Site	Window Quantity
Site_1	1 + 2 + 1 + 2 + 1 + 1 + 3 = 11

And the total window area at Site_1 is show as follows.

Table 7. Sum of Window Area by Site from Sample Window Data

CK_Site	Window Area (sq. ft.)
Site_1	(9 x 1) + (4 x 2) + (12 x 1) + (6 x 2) + (16 x 1) + (16 x 1) + (3 x 3) = 82

Aggregating Data to Different Levels

Users may wish to aggregate data using grouping variables different from CK_SiteID or CK_RoomID. If so, the same concepts discussed in the preceding sections apply, though end-users will need to identify appropriate grouping variables.

Estimation

This section provides database users with guidance regarding formulas they should use to estimate RBSA metrics for strata or subpopulations of interest. Stratification weighting is required to produce accurate estimates that account for the RBSA core sample design as well as the oversamples (see individual housing reports for details on sample design and post-stratification). This user manual assumes that database users will want to calculate metrics similar to those provided in the individual housing type report appendix, but for subpopulations of particular interest rather than for the region as a whole. This user manual is directed towards statisticians and/or assumes that the user will be familiar with mathematical notation for sums and averages that use subscripts to denote when an observation is for a different home, stratum, etc.

Both stratified estimation and domain estimation will be required and this section of the user manual provides descriptions and the corresponding formulas required to calculate means, proportions, and their standard errors for precision estimates.[1]

This manual provides examples and refers to four report tables (Tables 12, 13, 16, and 17) that can be found in Appendix A of the single-family home report.

Stratification Weights and Estimation

Stratified Estimation

Table 8 lists strata used for single family homes. In general, the population was stratified as follows:

- Each state is a separate stratum.
- Oregon is further divided into Eastern and Western regions.
- Washington is further divided into Puget Sound, Eastern, and Western regions.
- Washington's Puget Sound region is further divided into PSE, SCL, and Snohomish PUD utility territories.
- All states and regions are separated into BPA and non-BPA territories.

The result is a total of 21 strata in the single-family population, though sites were only sampled in 19 strata during the RBSA II.[2] The RBSA sample includes observations from homes in each stratum and the database provides the corresponding population and sample sizes. For the purposes of the RBSA, stratified estimation should be applied to metrics that are summarized at the home-level (e.g., home vintage and home square footage but not the square footage per room by room type because there are several rooms and room types per home). Domain estimation should be applied to metrics that are not summarized at the home level and is described below.

[1] All estimates are taken or derived from Cochran. Sampling Techniques. John Wiley & Sons, Inc. 1977.

[2] The additional strata are noted in order to fully characterize the population of the region, but are not included in the database as no sites were sampled. These strata population sizes make up a very small percentage of the region's population.

Table 8. RBSA Single-Family Stratification

State	Region	Territory	Category
ID	-	BPA	Sampled and included in database
ID	-	Non-BPA	
MT	W	BPA	
MT	W	Non-BPA	
OR	E	BPA	
OR	E	Non-BPA	
OR	W	BPA	
OR	W	Non-BPA	
WA	E	BPA	
WA	E	Non-BPA, Non-PSE	
WA	PS	BPA	
WA	PS	PSE	
WA	PS	SCL – Low Income (LI)	
WA	PS	SCL – Elec. Heated (EH)	
WA	PS	SCL – Not LI or EH	
WA	PS	SCL – LI and EH	
WA	PS	SnoPUD	
WA	W	BPA	
WA	W	PSE	
WA	W	Non-BPA, Non-PSE*	Accounted for in stratification, but not sampled in RBSA II
WA	E	PSE*	

Domain Estimation

A “domain” may consist of a particular room type or equipment vintage category—several domains can occur within one home and the summary of each domain includes observations from any or all strata. Domain estimates typically come from observations of more than one unit per home (e.g., from multiple rooms of multiple room types or multiple pieces of equipment that each belong to a vintage category). In contrast to stratified estimates, where one home belongs exclusively to one stratum, in domain estimation, one home can belong to multiple domains (e.g., one home includes square footage of rooms that belong to a number of room type domains such as kitchens, bathrooms, basements, etc.). Domain estimation requires estimating population sizes of units, such as the total number of televisions or pieces of equipment in the population. Domain estimates include stratification weighting.

Formulas

Stratified Estimation

Formulas in this section should be used for estimating whole-home metrics. They use the notation provided in Table 9 for stratified estimation.

Table 9. Notation for Stratified Estimation

Symbol	Description
Y,y	Population (upper-case) or sample (lower-case) observation(s)
l	Stratum defined by a unique combination of state, region, and service territory (Table 8)
i	Home identifier
N,n	Population (upper-case) and sample sizes (lower-case) of homes

Note that all estimates of multifamily common areas, central systems, or building attributes (such as insulation) are building-level metrics and will use stratified estimation formulas. The population sizes, N_l, correspond to building population sizes in stratum *l*, which were estimated, as described in the multifamily final report. The building population sizes N_l are included in the RBSA database. The standard errors do not account for uncertainty in the N_l estimates.

Estimation Within a Single Stratum

Calculating means or proportions within a single stratum does not require stratification weighting. If a stratum listed in Table 8 exactly describes the population of interest, the user should use the formulas for mean and proportion estimates in Equation 1 through Equation 4. Prior to performing calculations, the user should subset RBSA data to the desired stratum.

Means should be calculated using Equation 1 where y_{il} represents the observed metric for home *i* in stratum *l*, n_l represents the number of homes in the sample in stratum *l* with that metric observed, and $\bar{y}_l$ represents the estimated mean of the observed metric in stratum *l*.

Equation 1

$$\bar{y}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} y_{il}$$

Standard errors of the estimated means should be calculated using Equation 2, where the y_{il} and n_l are the same as in Equation 1, N_l represents the population size[3] of homes in stratum *l*, and $SE(\bar{y}_l)$ represents the standard error of the estimated mean in stratum *l*.

Equation 2

$$SE(\bar{y}_l) = \sqrt{\left(1 - \frac{n_l}{N_l}\right)\left(\frac{1}{n_l}\right)\left(\frac{1}{n_l - 1}\sum_{i=1}^{n_l}(y_{il} - \bar{y}_l)^2\right)}$$

Proportions should be calculated using Equation 3, where n_l^* represents the number of homes with the characteristic of interest (e.g., homes of a certain height) and n_l represents the number of homes in the sample in stratum *l*. For example, to estimate the proportion of homes in each building height category (as in Table 6 in the RBSA Appendix), the user should calculate n_l^* by counting the number of homes in stratum *l* with one story, two stories, etc. and then divide each n_l^* by n_l, the total number of homes with building heights observed in the stratum. The result will be one $\hat{p}_l$ value for each characteristic (e.g., building height category).

Equation 3

$$\hat{p}_l = \frac{n_l^*}{n_l}$$

Standard errors of estimated proportions should be calculated using Equation 4, where n_l is the same as above and N_l represents the population size of homes in stratum *l*. The estimate from Equation 3 results in $\hat{p}_l$ and $SE(\hat{p}_l)$ represents the standard error of the estimated proportion in stratum *l*.

Equation 4

$$SE(\hat{p}_l) = \sqrt{\left(1 - \frac{n_l}{N_l}\right)\frac{\hat{p}_l(1 - \hat{p}_l)}{n_l}}$$

[3] The population sizes of homes in each stratum are estimates from the U.S. Census Bureau’s American Community Survey.

Example: Single-Family Tables 12 and 13 within One Stratum

Prior to performing any calculations, the user should subset the RBSA database to homes in the stratum of interest, stratum *l*. Means and error bounds[4] in single-family Table 5 summarize the conditioned floor area by home vintage and should be calculated using Equation 1 and Equation 2 within each vintage category. The user should calculate Equation 1 nine times, once for each of the eight vintage categories and once for the all vintages category which will produce nine mean estimates, or $\bar{y}_l$ values. Similarly, the user should calculate Equation 2 nine times to produce a standard error estimate $SE(\bar{y}_l)$ corresponding to each $\bar{y}_l$ value.

In this example, each y_{il} value represents the observed conditioned floor area for home *i* in stratum *l* and n_l represents the number of sampled homes in each vintage category with conditioned floor area observed. Subscript *i* indexes the homes. The user should sum conditioned floor area in each vintage category and then divide by the n_l value for that category. This will result in the mean conditioned floor area for each vintage category for stratum *l*, or $\bar{y}_l$. The user should calculate the standard error for each estimate using Equation 2. To calculate the standard error, or $SE(\bar{y}_l)$, for the mean conditioned floor area within each vintage category, the user should insert y_{il}, $\bar{y}_l$, and n_l as defined above, and the population size of homes in stratum *l*, N_l, which can be found in the database for each stratum. The result will be nine standard error estimates, one for each vintage category and one for the all vintages category.

Proportions and error bounds in single-family Table 13 summarize the percentage of homes by building height and should be calculated using Equation 3 and Equation 4. Similar to the example for Table 12, users should calculate five proportions and the corresponding standard errors (one for each building height category). In this case though, the total row will be the sum of the proportions from the categories and so the user does not need to use the formulas for the total row.

The user should count the number of homes with the building height in each category to find n_l^*, resulting in five counts, and count the total number of homes with any building height observed to find n_l, resulting in one count. The user should divide each n_l^* value by the n_l value, resulting in five proportion estimates, $\hat{p}_l$ within the stratum. The user should calculate the standard error $SE(\hat{p}_l)$ using Equation 4 and plugging in the $\hat{p}_l$ and n_l values along with the population size of homes in the stratum, N_l.

Estimation Within a Stratum Subpopulation

Suppose the user chooses to produce results similar to those described in the previous example, but based on a subpopulation of one stratum, e.g., the Portland-area subpopulation within the Western Oregon BPA stratum, then the user can use the same formulas as used above. This works when the subpopulation(s) are contained wholly in one stratum and do not belong to any other strata.

Stratified Estimation Combining Multiple Strata

Estimating whole-home means or proportions for populations that include multiple combined strata requires stratification weighting. If the subpopulation of interest can be described by two or more strata listed in Table 8 (e.g., Idaho—consisting of the Idaho BPA stratum and the Idaho Non-BPA stratum), the user should use the formulas for stratified mean and proportion estimates in Equation 5 through Equation 8. These build upon the previous section’s formulas (Equation 1 through Equation 4).

Means should be calculated using Equation 5, where $\bar{y}_l$ is the mean within stratum *l* and N_l is the population size of homes in stratum *l*. The user should sum the product of the population sizes and mean estimates and then divide by the sum of the population sizes. In the summation notation, *L* represents the total number of

[4] The appendix tables report the error bounds. This user manual provides guidance on calculating the standard errors required to calculate error bounds. To calculate the error bound from the standard error, the user should multiply the calculated standard error by the t-statistic at the desired confidence level.

strata. The result is a combined mean estimate, $\bar{y}$. For example, to estimate the mean conditioned floor area in Idaho, the user should assign *l=1* to represent the Idaho BPA stratum and *l*=2 to represent the Idaho Non-BPA stratum; because the total number of strata is two the user should assign *L=2*. The user should sum the population sizes from the two strata to calculate the denominator in Equation 5 and divide the sum of the product of each stratum population size N_l with the stratum mean estimate $\bar{y}_l$ by the summed population sizes.

Equation 5

$$\bar{y} = \frac{\sum_{l=1}^{L} N_l * \bar{y}_l}{\sum_{l=1}^{L} N_l}$$

Standard errors of estimated means should be calculated using Equation 6 with the standard error of the estimated mean in each stratum *l*, $SE(\bar{y}_l)$, defined above in Equation 2. The user should sum the products of the squared population sizes, N_l, and the squared mean standard errors, $SE(\bar{y}_l)$, take the square root of the sum, and then divide by the sum of the population sizes.

Equation 6

$$SE(\bar{y}) = \frac{1}{\sum_{l=1}^{L} N_l} \sqrt{\sum_{l=1}^{L} N_l^2 * SE(\bar{y}_l)^2}$$

Proportions should be calculated using Equation 7, where $\hat{p}_l$ represents the proportion in each stratum *l* and is defined above and in Equation 3. The population size, N_l, represents the number of homes in each stratum *l*. Similar to the combined mean estimate, the user should sum the product of the population sizes and the strata proportions and then divide by the sum of the strata population sizes.

Equation 7

$$\hat{p} = \frac{\sum_{l=1}^{L} N_l * \hat{p}_l}{\sum_{l=1}^{L} N_l}$$

Standard errors of the combined proportion estimates should be calculated using Equation 8, with the standard error of the estimated proportion in each stratum, $\hat{p}_l$, calculated using Equation 4 and the strata population sizes N_l. The user should sum the product of the squared population sizes, N_l, and squared standard errors, $SE(\hat{p}_l)$, in each stratum, take the square root of the sum, and divide by the sum of the population sizes.

Equation 8

$$SE(\hat{p}) = \sqrt{\frac{\sum_{l=1}^{L} N_l^2 \; * SE(\hat{p}_l)^2}{\sum_{l=1}^{L} N_l}}$$

The RBSA database provides population sizes N_l for each stratum *l*, defined in Table 8.

Example: Single-Family Tables 12 and 13 for Combined Strata

Prior to performing calculations, the user should subset the RBSA database to homes in each of the strata of interest (e.g., Idaho BPA and Idaho Non-BPA). As in the previous example, whole-home means and standard errors in single-family Table 5 summarize the mean conditioned floor area by home vintage. Though in this example the user is interested in a population that includes multiple strata combined.

Mean conditioned floor area, $\bar{y}_l$, and the standard error of the mean, $SE(\bar{y}_l)$, in each stratum *l* should be calculated using Equation 1 and Equation 2 from the previous section. As in the example for estimates in one stratum, the user should calculate the mean once for each of the seven vintage categories and once for the all vintage category to produce eight estimates, or $\bar{y}_l$ values (one for each category). Similarly, the user should calculate Equation 2 eight times to produce a standard error estimate, $SE(\bar{y}_l)$, corresponding to each estimated mean value. The user should repeat this process, once for each stratum, resulting in eight mean and eight standard error estimates per stratum. For example, if the user calculates combined results for Idaho, the result will be eight mean estimates $\bar{y}_1$ and eight standard error estimates $SE(\bar{y}_1)$ for Idaho BPA (*l*=1) and eight mean estimates $\bar{y}_2$ and eight standard error estimates $SE(\bar{y}_2)$ for Idaho Non-BPA (*l*=2).

Once the user has calculated the mean and standard error estimates within vintage categories for each stratum, Equation 5 and Equation 6 should be used to combine the estimates. Again, within each home vintage category, the user should apply Equation 5 to calculate the combined mean estimate $\bar{y}$ by summing the product of the population size within Idaho BPA, N_1, and the mean estimate within Idaho BPA, $\bar{y}_1$, with the product of the population size within Idaho Non-BPA, N_2, and the mean in Idaho Non-BPA, $\bar{y}_2$, and then divide by the sum of the population sizes N_1 and N_2.

Similarly, proportions for combined strata and their standard errors in single-family Table 13 should be calculated using Equation 7 and Equation 8, which use estimates within each stratum, calculated using Equation 3 and Equation 4.

Example: Single-Family Tables 12 and 13 for Combined Subpopulations of Strata

Suppose the user wants to produce results similar to those in single-family Tables 12 and 13, but based on subpopulations of two or more strata (e.g., averages that include homes in the Portland-area combined with Oregon Tri-Cities but excluding other portions of Oregon). The user should use the same formulas as for combining multiple strata.

This requires the user to take the following steps:

- Subset the RBSA database to homes in Portland from the Western Oregon BPA stratum and in the Tri-Cities area from the Eastern Oregon BPA stratum, and treat these subpopulations as stratum *l=1* and *l=2*, respectively, in the equations
- Calculate subpopulation means, proportions, and standard errors within each “stratum”
- Combine the subpopulation estimates using the formulas in Equation 5 through Equation 8

Note: the user will be required to supply the population sizes N_l for each subpopulation; they are not included in the RBSA database.

Domain Estimation

Formulas in this section should be used for estimating domain means, domain proportions, and their standard errors. In addition to stratified estimation, users will need to use domain estimation to summarize metrics within subpopulations, or domains. Whereas the previous examples demonstrated how to calculate and then combine estimates from subpopulations of homes that can only exist in one stratum, domain subpopulations refer to subpopulations of characteristics that can occur several times within a home and where a home can belong to more than one domain. For example, a domain may consist of a particular room type or equipment vintage, where several domains can occur within one home (e.g., one home contains a bedroom, bathroom, and kitchen). Domain observations typically come from more than one unit per home (e.g., square footage from each room type or vintage of two or more units of heating equipment). In these cases, domain estimation is required.

Domain estimation requires estimating population sizes of units (e.g., the total number of televisions within a room type and for the region overall). Domain estimates include stratification weighting. This user manual provides formulas and examples below. The formulas for domain estimation use the notation provided in Table 10.

Table 10. Notation for Domain Estimation

Symbol	Description
Y,y	Population (upper-case) or sample (lower-case) observation
k	Domain (e.g., room type, appliance efficiency)
l	Stratum defined by a unique combination of, state, region, and service territory (Table 8)
i	Home identifier
j	Component identifier
M,m	Population (upper-case) and sample sizes (lower-case) of components
N,n	Population (upper-case) and sample sizes (lower-case) of homes

Domain Estimation Within a Single Stratum or Subpopulation of a Stratum

Prior to performing calculations, the user should subset the RBSA data to the desired strata, stratum, or stratum subpopulation as described above in the Stratified Estimation. Note that all estimates of multifamily common areas, central systems, or building attributes (such as insulation) are building-level metrics and will use stratified estimation formulas. The population sizes, N_l, correspond to building population sizes in stratum *l*, which were estimated, as described in the multifamily final report. The building population sizes N_l are included in the RBSA database. The standard errors do not account for uncertainty in the N_l estimates.

Estimation Within a Single Stratum and Estimation Within a Stratum Subpopulation sections. Calculating domain means and proportions within a single stratum requires using formulas for domain estimation that combine multiple strata, provided below.

The only difference will be that the user will not be required to sum over strata because the estimate will apply to one stratum or stratum subpopulation only. This means that in Equation 9 through Equation 13 below, the sum from *l=1 to L* will not be included.

Domain Estimation for Combining Multiple Strata

Calculating domain means and proportions requires using the formulas provided in Equation 9 through Equation 13. These build upon formulas provided in the previous sections.

Domain Mean Estimation

Domain means should be calculated using Equation 9, where y_{jilk} is the observation for unit *j* in home *i*, stratum *l*, and domain *k*. Unit *j* could represent, for example, the *j*th room of room type *k* within home *i*; then, y_{jilk} could represent the square footage of that room. The subscript *j* ranges from one to m_{ilk}, or $j = 1, \dots, m_{ilk}$ and the variables M_{ilk} and m_{ilk} represent the total number of units (e.g., rooms that are bedrooms) and the observed number of units in home *i* in stratum *l*. They can be calculated by counting units $j = 1, \dots, m_{ilk}$ in domain *k* in each home *i* in stratum *l*. In most homes the counts M_{ilk} and m_{ilk} are equal because the RBSA observed all units within a home whenever possible and cases where we were unable to observe a census of metrics within a home rarely occurred.

The user should sum observations within a home to calculate $\sum_{j=1}^{m_{ilk}} y_{jilk}$ and then sum across the n_{lk} homes in stratum *l* that have units in domain *k* to calculate the total observed value, $\sum_{i=1}^{n_{lk}} \sum_{j=1}^{m_{ilk}} y_{jilk}$ in stratum *l*. As in stratified estimation, N_l represents the population size of homes in stratum *l* and n_{lk} represents the sample size, or number of homes in stratum *l* with at least one unit in domain *k*. The ratio of population size to sample size represents the stratification weight in each stratum. The user should sum the product of the stratification weights and stratum total observed values, and then divide by the total number of units in the domain to estimate the domain *k* mean per unit value, or $\bar{y}_k$, in the combined strata population. This calculation requires the estimate $\widehat{M}_k$, which is the estimated population size of units in domain *k* in all homes in the population of combined strata. It should be estimated using Equation 10.

Equation 9

$$\bar{y}_k = \frac{1}{\widehat{M}_k} \sum_{l=1}^{L} \frac{N_l}{n_{lk}} \sum_{i=1}^{n_{lk}} \frac{M_{ilk}}{m_{ilk}} \sum_{j=1}^{m_{ilk}} y_{jilk}$$

In Equation 10, $\widehat{M}_k$ represents the total population size of units in domain *k*. The user should estimate it by summing m_{ilk}, the number of domain *k* units in home *i* in stratum *l,* using stratification weights $\frac{N_l}{n_{lk}}$ in each stratum.

Equation 10

$$\widehat{M}_k = \sum_{l=1}^{L} \frac{N_l}{n_{lk}} \sum_{i=1}^{n_{lk}} m_{ilk}$$

The standard errors of the estimated domain means should be calculated using Equation 11. This formula includes the variables defined above, such as $\bar{y}_k$, the domain *k* mean per unit in the combined strata, N_l and n_l, the population and sample sizes of homes in stratum *l*, n_{lk}, the sample sizes of homes in stratum *l* with units in domain *k*, and $\widehat{M}_k$, the estimated population size of units in domain *k*. Additional terms are required in the standard error calculation, including $\bar{y}_{ilk}$, the estimated mean in domain *k*, home *i*, stratum *l*, and $\bar{y}_{lk}$, the estimated mean in domain *k*, in stratum *l*. The user should calculation these using Equation 12 and Equation 13.

Equation 11

$$SE(\bar{y}_k) = \sqrt{\frac{1}{\widehat{M}_k{}^2} \sum_{l=1}^{L} \frac{N_l{}^2}{n_l(n_l - 1)} \left(1 - \frac{n_l}{N_l}\right) \left[\frac{1}{n_{lk}} \left(1 - \frac{n_l}{N_l}\right) \sum_{i=1}^{n_{lk}} \frac{(\bar{y}_{ilk} - \bar{y}_{lk})^2}{n_{lk} - 1} + n_{lk}(\bar{y}_{lk} - \bar{y}_k)^2 \right]}$$

In Equation 12, the user should calculate $\bar{y}_{ilk}$, the within home mean of observations from units $j = 1, \dots, m_{ilk}$ in domain *k* and stratum *l*, by summing the y_{jilk} observations over units within home *i* and then dividing by m_{ilk}, the number of domain *k* units in home *i*.

Equation 12

$$\bar{y}_{ilk} = \frac{\sum_{j=1}^{m_{ilk}} y_{jilk}}{m_{ilk}}$$

In Equation 13, the user should calculate $\bar{y}_{lk}$**,** the within stratum and domain mean, by summing over the y_{jilk} observations in all n_{lk} homes and then dividing by n_{lk}, the number of homes with units in domain *k* in stratum *l*.

Equation 13

$$\bar{y}_{lk} = \frac{\sum_{i=1}^{n_{lk}} \sum_{j=1}^{m_{ilk}} y_{ilk}}{n_{lk}}$$

Domain Percentage Estimation

Percentages for distribution summaries (e.g., percent of televisions by room type) should be calculated using Equation 14 and Equation 15. Each $\widehat{M}_k$ represents an estimate of the total number of domain *k* units in the population and can be calculated as described above in Equation 10. The user should divide this quantity by the total number of units in the population, $\widehat{M}$, to estimate the domain *k* percentage. As in the formulas above, the estimate $\widehat{M}_k$ represents the population size of domain *k* units in all homes and should be estimated using Equation 10. The estimate $\widehat{M}$ represents the total number of units in all domains, or the sum of units in domains $k = 1, \ldots, K$ and should be calculated by summing $\widehat{M}_k$ values over the domains. The number of domains, or *K*, will depend on the particular table or summary the user produces. For example, in the RBSA there are 12 room type domains—bathroom, bedroom, closet, dining room, etc., so in a summary of the percent of televisions by room type, the user should calculate *K=12* $\widehat{M}_k$ values—$\widehat{M}_1, \widehat{M}_2, \widehat{M}_3, \ldots, \widehat{M}_{12}$, one for each room type and then sum the $\widehat{M}_k$ values to estimate $\widehat{M}$.

Equation 14

$$\hat{p}_k = \frac{1}{\widehat{M}} \widehat{M}_k$$

Equation 15

$$\widehat{M} = \sum_{k=1}^{K} \widehat{M}_k$$

Finally, the user should calculate the standard errors of domain percentages using Equation 16 and Equation 17. The variables are defined as noted above with the addition of $\bar{m}_{lk}$, or the average of m_{ilk} across homes, which can be calculated using Equation 17.

Equation 16

$$SE(\hat{p}_k) = \sqrt{\frac{1}{\widehat{M}_k^{\ 2}} \sum_{l=1}^{L} \frac{N_l^{\ 2}}{n_l} \left(1 - \frac{n_l}{N_l}\right) \sum_{i=1}^{n_l} \frac{(m_{ilk} - \bar{m}_{lk})}{n_l - 1}}$$

Equation 17

$$\bar{m}_{lk} = \frac{1}{n_l} \sum_{i=1}^{n_l} m_{ilk}$$

Example: Single-Family Tables 16 and 17 for Combined Strata

Domain Means in Table 16

Domain means and standard errors in single-family Table 16 summarize mean room areas by room type. For this table, each room type is a domain and each room is a unit of observation. If the user wishes to calculate the average room area by room type in some combination of strata, say in Eastern Oregon (Eastern Oregon BPA and Eastern Oregon Non-BPA strata combined), then the strata should be labeled *l=1* for Eastern Oregon BPA and *l=2* for Eastern Oregon Non-BPA (or vice versa). In this example, the room types represent the domains, $k = 1, \ldots, 12$, and a single room type, say bedrooms, represent one domain *k*. Individual rooms represent the units, $j = 1, \ldots, m_{ilk}$, where the number of rooms of type *k*, m_{ilk}, is equal to one in a one-bedroom house, two in a two-bedroom house, etc. Prior to performing calculations, the user should subset the RBSA database to homes in the two strata.

The user should calculate average floor area by room type using Equation 9 where values y_{jilk} are the room areas for each room *j* of type *k* in home *i* in stratum *l*. The user should sum the room areas over the m_{ilk} rooms of type *k* in each home to calculate $\sum_{j=1}^{m_{ilk}} y_{jilk}$ and then sum over all n_{lk} homes in stratum *l* that have rooms of type *k* to calculate the total square footage for rooms of type *k* in all homes in stratum *l*, $\sum_{i=1}^{n_{lk}} \sum_{j=1}^{m_{ilk}} y_{jilk}$. Next, the user should sum the products of weights, $\frac{N_l}{n_{lk}}$, and stratum total square footages to estimate the total square footage for rooms of type *k* in the combined strata population, $\sum_{l=1}^{L} \frac{N_l}{n_{lk}} \sum_{i=1}^{n_{lk}} \sum_{j=1}^{m_{ilk}} y_{jilk}$. Finally, the user should divide this total by the estimated number of rooms of type *k* in the combined strata population, $\widehat{M}_k$, which can be calculated using Equation 10. The result will be $\bar{y}_k$, the mean square footage per room, of rooms of type *k* in Eastern Oregon.

Domain Percentages in Table 17

Percentages and their standard errors in single-family Table 17 summarize the percent of televisions in each room type. For this table, similar to that in the previous example, room types are the domains. In this example, each television is a unit. Domains are again represented using $k = 1, \dots, 12$, where a single room type represents one domain *k*. Individual televisions represent the units, $j = 1, \dots, m_{ilk}$, where the number of televisions in any room of type k, m_{ilk}, is equal to one in homes with one television in a single room of type *k*, two in homes with one television in each of two rooms of type *k* (or, alternately, two televisions in one of two rooms of type *k* and none in the other room of type *k*), etc. Each home should have one value of m_{ilk}.

The user should calculate the percentage of televisions in each room type using Equation 10, Equation 14, Equation 15, and Equation 16. The user should calculate m_{ilk} as the number of televisions in rooms of type *k* in each home *i* in stratum *l* and then, using Equation 10, calculate $\widehat{M}_k$ by summing the product of the stratum totals and the stratification weights. The user should use Equation 16 to calculate the number of televisions in the combined strata population, $\widehat{M}$, and then divide the $\widehat{M}_k$ by the $\widehat{M}$ to estimate the percentage of televisions in each room of type *k*. The resulting percentage estimate, $\hat{p}_k$, will represent the proportion of televisions that occur in each room type in Eastern Oregon.

Finally, the user should plug these values into Equation 16 and Equation 17 to calculate the standard error of this estimate.

Domain Estimation Combining Subpopulations of Multiple Strata

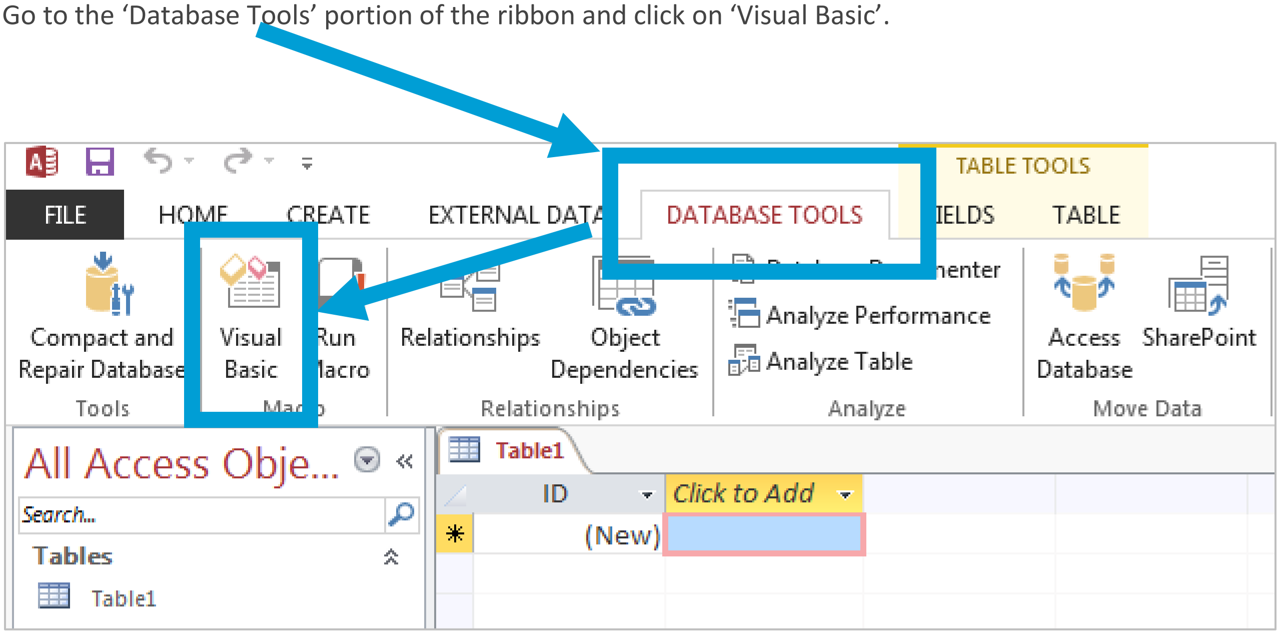
The user should use the formulas in the previous section to calculate domain mean and percentage estimates for a combination of subpopulations of multiple strata. Counts and observations will correspond to each subpopulation in place of the strata (i.e., the index *l* will represent the subpopulations instead of the strata). The user will subset the RBSA database to homes in the subpopulations of interest before counting the number of observations in the sample for each stratum and will be required to supply the N_l values, or population sizes for each subpopulation.

Importing into MS Access

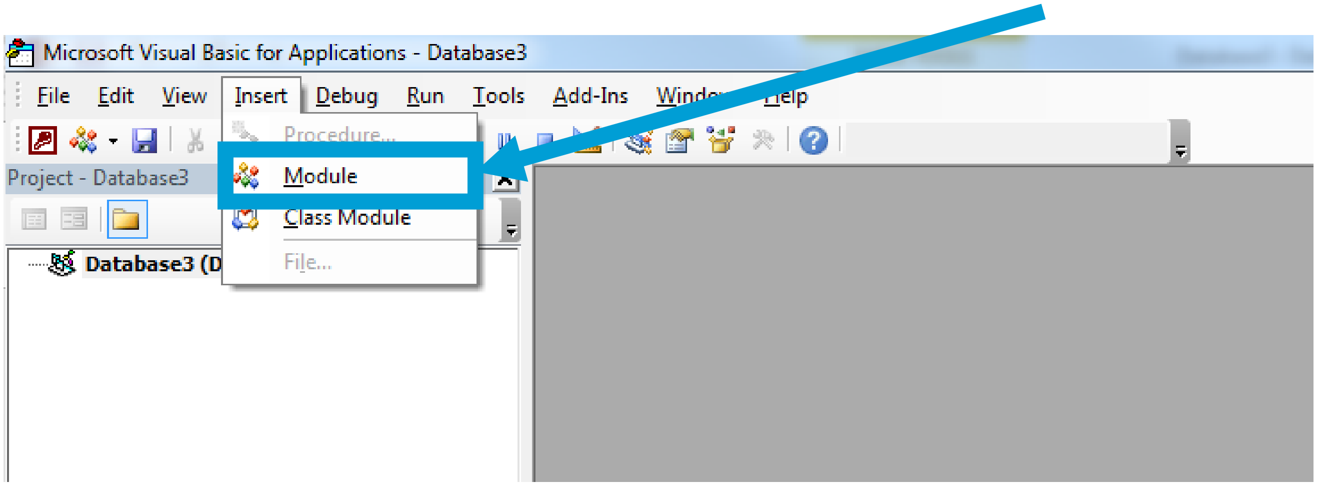
As noted, the RBSA database is made available as a collection of CSV files. Some users may wish to import these data into a MS Access database—the same format used in the previous RBSA database. It is possible to import the entire set of CSV files into MS Access using a Visual Basic for Applications (VBA) script. The script and steps necessary to use the script are outlined below.

The following procedure imports CSVs into MS Access:

1. Open Microsoft Access.
2. Create a 'Blank Desktop Database'.
3. Go to the 'Database Tools' portion of the ribbon and click on 'Visual Basic'.



4. The window that opens looks like the image below. Click on 'Insert' and choose 'New Module'.

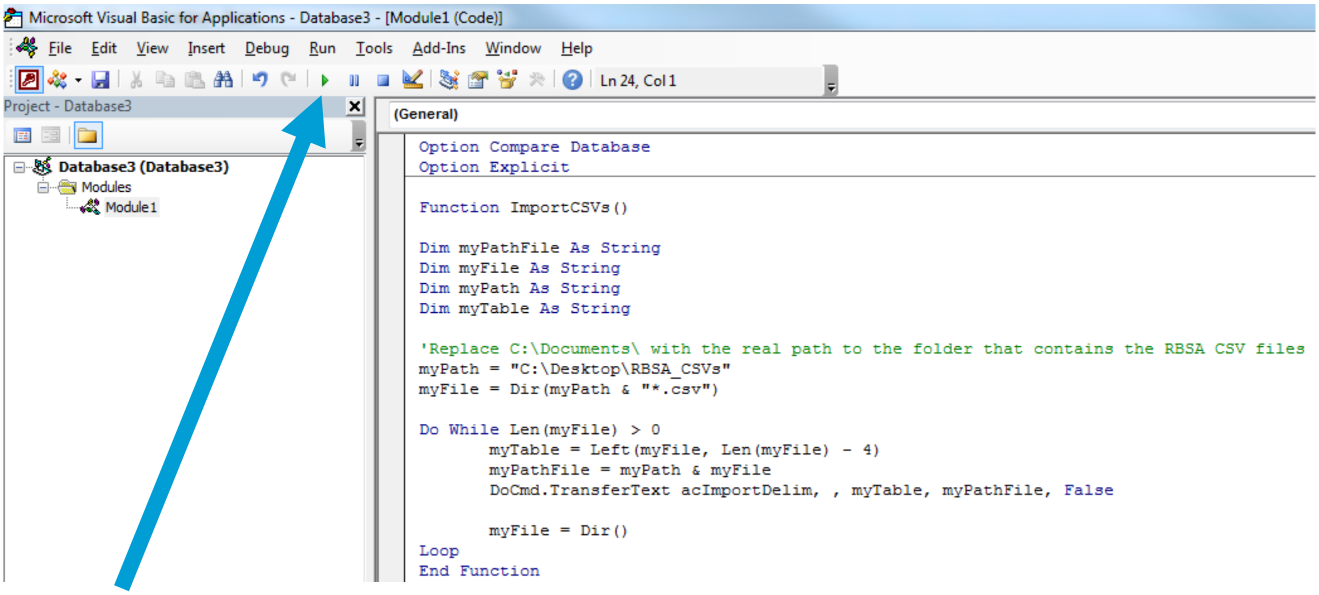


5. Copy the code snippet below into the Visual Basic Module, overwriting anything already existing in the module.

Code Snippet 1. VBA Script to Import CSV files into MS Access

```
Option Compare Database
Option Explicit

Function ImportCSVs()

Dim myPathFile As String
Dim myFile As String
Dim myPath As String
Dim myTable As String

'Replace C:\Documents\ with the real path to the folder that contains the RBSA CSV files
myPath = "C:\Documents\"
myFile = Dir(myPath & "*.csv")

Do While Len(myFile) > 0
    myTable = Left(myFile, Len(myFile) - 4)
    myPathFile = myPath & myFile
    DoCmd.TransferText acImportDelim, , myTable, myPathFile, FALSE

    myFile = Dir()
Loop
End Function
```

6. On the line beginning with ‘myPath = …’, replace the “C:\Documents\” with the location of the downloaded RBSA Database CSV files. In the image below, the location has been updated to “C:\Desktop\RBSA_CSVs\”



```
Option Compare Database
Option Explicit

Function ImportCSVs()

Dim myPathFile As String
Dim myFile As String
Dim myPath As String
Dim myTable As String

'Replace C:\Documents\ with the real path to the folder that contains the RBSA CSV files
myPath = "C:\Desktop\RBSA_CSVs"
myFile = Dir(myPath & "*.csv")

Do While Len(myFile) > 0
       myTable = Left(myFile, Len(myFile) - 4)
       myPathFile = myPath & myFile
       DoCmd.TransferText acImportDelim, , myTable, myPathFile, False

       myFile = Dir()
Loop
End Function
```

7. Click the green arrow to ‘Run Sub / Userform’. The script will execute and import the CSV data into MS Access.
8. Close the Visual Basic Editor.
9. Close and save the database.

Appendix A: Database Strata by Home Type

The following tables summarize the full list of strata by home type, as found in the RBSA database.

Single-Family Strata

Table 11. RBSA Single-Family Stratification

State	Region	Territory	Category
ID	-	BPA	Sampled and included in database.
ID	-	Non-BPA	
MT	W	BPA	
MT	W	Non-BPA	
OR	E	BPA	
OR	E	Non-BPA	
OR	W	BPA	
OR	W	Non-BPA	
WA	E	BPA	
WA	E	Non-BPA, Non-PSE	
WA	PS	BPA	
WA	PS	PSE	
WA	PS	SCL – Low Income (LI)	
WA	PS	SCL – Elec. Heated (EH)	
WA	PS	SCL – Not LI or EH	
WA	PS	SCL – LI and EH	
WA	PS	SnoPUD	
WA	W	BPA	
WA	W	PSE	
WA	W	Non-BPA, Non-PSE*	Accounted for in stratification, but not sampled in RBSA II. Not included in database.
WA	E	PSE*	

* The sample did not include any homes in these strata; their population sizes make up only a very small percentage of the region's population.

Manufactured Home Strata

Table 12. RBSA Manufactured Home Stratification

State	Region	Territory	Category
ID	-	BPA	Sampled and included in database.
ID	-	Non-BPA	
MT	W	BPA	
MT	W	Non-BPA	
OR	E	BPA	
OR	E	Non-BPA	
OR	W	BPA	
OR	W	Non-BPA	
WA	E	BPA	
WA	E	Non-BPA, Non-PSE	
WA	E	PSE	
WA	PS	BPA	
WA	PS	PSE	
WA	PS	SCL	
WA	PS	SnoPUD	
WA	W	BPA	
WA	W	PSE	
WA	W	Non-BPA, Non-PSE*	Accounted for in stratification, but not sampled in RBSA II. Not included in database.

* The sample did not include any homes in these strata; their population sizes make up only a very small percentage of the region's population.

Multifamily Strata

Table 13. RBSA Multifamily Stratification

State	Region	Territory	Category
ID	-	BPA	Sampled and included in database.
ID	-	Non-BPA	
MT	W	BPA	
MT	W	Non-BPA	
OR	E	BPA	
OR	E	Non-BPA	
OR	W	BPA	
OR	W	Non-BPA	
WA	E	BPA	
WA	E	Non-BPA, Non-PSE	
WA	PS	BPA	
WA	PS	PSE King County	
WA	PS	PSE Non-King County	
WA	PS	SCL	
WA	PS	SnoPUD	
WA	W	BPA	
WA	W	PSE Non-King County	
WA	E	PSE King County**	Accounted for in stratification, but not sampled in RBSA II. Not included in database.
WA	E	PSE Non-King County*	
WA	W	Non-BPA, Non-PSE*	
WA	W	PSE King County**	

* The sample did not include any homes in these strata; their population sizes make up only a very small percentage of the region's population.

** The sample did not include any homes in these strata; their population sizes are zero.