revengc: Reverse Engineering Censored, Decoupled Residential Data for Population Density Estimation

Samantha Duchscherer and Robert Stewart
Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831

Abstract

A wealth of open source information is available that points to building usage including floor area of building size and likely occupancy. In the case of residential structures, census data provides a number of attributes including two values important to interior density estimation: household size (hhs) and area. If a national census revealed the raw data or provided a full uncensored contingency table (hhs x area), computing interior density as people/area would be straightforward. However, agencies rarely report this contingency table. Rather hhs and area are often decoupled and reported as separate univariate frequency tables, average values, or a combination of the two. In addition, the decoupled or contingency tables provided are typically left ($<$, $\leq$), right ($>$, $\geq$), and interval (-) censored. This type of information becomes problematic in estimating interior residential occupancy for numerous reasons. How can the people/area ratio be calculated when no affiliation between the variables exist? If a census reports an hhs average of 5.3, then how many houses are there with 1 person, 2 people,.., 10 people? If a census reports that there are 100 houses in an area of 26-50 square meters, then how many houses are in 26, 27,..., 50 square meters? The challenge therefore is to infer the people/area ratio when given decoupled and summarized data. The statistical package **revengc** was designed to reverse engineer censored, decoupled census data into a likely hhs x area uncensored contingency table for estimating interior residential occupancy.

Keywords: census, residential, building, occupancy, bivariate Poisson distribution, censored, decoupled, contingency table

1 Introduction

The interest in open source population density data continues to grow. Modeled population densities can be implemented to associate buildings with sociocultural activities, to distinguish between daytime and nighttime occupancy levels, to approximate natural hazard loss analytics, and to minimize building energy consumption [15]. The main desire of most population studies is to determine the ratio of people/area for a facility (museums, hospitals, schools, etc.). This statistical approach focuses on residential information containing specific types of uncertainty. The concentrate on residential data was chosen primarily because it is the most dominant population data in the open source environment. It encompasses a wide spectrum including urban, rural, multi-family, refugee Settlements, tents, and etc. [15]. Furthermore, residential data is recognized as being notorious for reporting key variables used in interior density estimation. A majority of the time household size (hhs) and area are found in any national census.

The issue occurs when national census authors may have privy too but normally do not reveal the clear information of people/area. Reoccurring problems are decoupled variables (e.g. separate averages and frequency tables) and numeric censoring (e.g. area is between 50-100 m^2 or hhs is listed as $>$ 8 people). Decoupled variables provides no availability for a cross tabulation between hhs and area (people/area) while censoring obscures the true underlying values. If a census reports an hhs average of 5.3, then we know that average was calculated based off some houses being above and below 5.3. However, we don't know the exact amount of houses that are above or below 5.3. Censoring distorts data the same way. If a census reports that there are 100 houses in an area of 26-50 m^2, then we don't know how many houses are in 26m^2, 27m^2,.., 50m^2. This decoupled and summarized information is misleading, thus this model reverse engineers censored, decoupled census data into a likely hhs x area contingency table that more accurately describes interior residential occupancy. The following summarizes the four scenarios that can be reverse engineered by the **revengc** package.

Case I: Averages for both hhs and area
Case II: Decoupled hhs and area frequency tables
Case III: Combination of hhs (area) average and area (hhs) frequency table
Case IV: hhs x area contingency table (censored)

2 Method

The reverse engineering techniques relies on the Poisson and bivariate Poisson distribution. For the decoupled information (averages, tables, or combination of the two), separate Poisson distributions are calculated for hhs and area. These distributions are generated with a λ that is found with two possibilities. A given average will represent the λ value while a frequency table can use a custom maximum likelihood function to estimate the needed λ value. The separate Poisson distributions will then be used to represent the marginals in Iterative Proportional Fitting (IPF). IPF is a generalized method that estimates cross tabulations between decoupled variables. For the case of the censored contingency table, another customized maximum likelihood function will be implemented to approximate λ_3, which subsequently provides the necessary λ_1 and λ_2 parameters. The estimated λ_1, λ_2, and λ_3 values can then be used to generate cross tabulations with a bivariate Poisson distribution. For all cases, the model outputs an uncensored cross tabulation of probabilities (contingency table) relating hhs and area. The final contingency table is also subject to right shifts and right truncation for representation of lower and upper bounds existing in both variables. Using actual census reports, we present a methodology workflow of all cases in section 4.6.

2.1 Poisson

The Poisson distribution has interesting and convenient properties that allow a common solution across all case studies considered here. Jennings [5] showed that the hhs size follows a univariate Poisson distribution across a number of countries. However, the area variable is more problematic, and no formal peer reviewed area studies available seem equivalent to the hhs analysis by Jennings. A review of open source information suggests that area varies. For example, Klaiber [8] reports a lognormal "appearance" to area data. Some even show data that come close to a normal distribution, such as Shuman [13]. Also, Ohnishi [10] unusually fit house size to an exponential distribution in Tokyo in an unpublished work. Within this context, there is a practical engineering solution to the problem. Specifically it was decided to model area as a Poisson distribution while fully aware that area is a continuous property. It is convenient to model it as a Poisson distribution of "counted" units of measure for several reasons. First, census data rarely reports area with precision greater than simple integer values for any unit of measure. Considering this a practical matter, it is therefore not unreasonable to think of these data as unit of measure "counts." It is also known that for large values of the Poisson parameter (e.g. $\lambda > 20$) the Poisson and the Normal distribution are roughly equivalent. If the underlying continuous distribution for area is Gaussian, the Poisson solution here will still provide a very close approximation. Furthermore, the Poisson distribution has the convenient property that it's only parameter (λ) is equal to the mean of the count data, a commonly encountered value in the open source. This allows availability to reverse engineer Cases I, II, III under a common and consistent statistical model. Finally, by treating square meters as integer counts, the results given by Kawamura [7] can be utilized. Kawamura showed that fitting the three parameters of the bivariate Poisson distribution (BP) can be reduced to fitting only the 3$^{\text{rd}}$ parameter when the univariate Poisson marginals (hhs, area) are known (or estimated). This provides a practical and consistent approach to reverse engineering Case IV that continues to rely on the Poisson as a common underlying distribution.

Let's now recall the probability mass functions for Poisson and bivariate Poisson. Equation (1) shows the formula for the Poisson probability mass function

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad (x = 0, 1, 2, ...) \tag{1}$$

A way to calculate a bivariate Poisson distribution (X,Y) is to let $\mathrm{Z_i} \sim$ Poission (λ_i) where i = 1, 2, 3 and set

$$\mathrm{X} = \mathrm{Z_1} + \mathrm{Z_3}$$
$$\mathrm{Y} = \mathrm{Z_2} + \mathrm{Z_3}$$
$$(\mathrm{X,Y}) \sim \mathrm{BP}(\lambda_1,\, \lambda_2,\, \lambda_3)$$

The probability function of the bivariate Poisson distribution (X, Y) then can be generated by Equation (2).

$$P(X = x, Y = y) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^x}{x!} \frac{\lambda_2^y}{y!} \sum_{i=0}^{min(x,y)} i! \binom{x}{i} \binom{y}{i} \left(\frac{\lambda_3}{\lambda_1 \lambda_2}\right)^i \tag{2}$$

2.2 Lower and Upper Bounds

It can be concluded that a certain amount of people in a small area (e.g. 1 person per 0 m^2) is impossible. It can also be assumed that a given number of people in a large residential area (e.g. 2 people per 1,000,000 m^2) would be a minuscule probability worth refining to zero. Thus, this model incorporates bounds to narrow infinite ranges. The final contingency table will have rows ranging from a selected lower and upper hhs bound along with columns ranging from a lower to upper area bound. Ideally, the four bounds should be chosen based off prior knowledge and expert elicitation, but they can also be selected intuitively with a brute force method. If the reverse engineering tool outputs a final contingency table with higher probabilities near the edge(s) of the table, then it would make sense to increase the range of the bound(s). For both the hhs and area variables, this would just involve making the lower bound less, making the upper bound more, or doing a combination of the two. The opposite holds true as well. If the final contingency table has very low probabilities near the edge(s) of the table, then a person should decrease the range of the particular bound(s).

The use of these bounds in the censored contingency tables (Case IV) is simple. Bivariate Poisson probabilities are only calculated for the ranges provided by the lower and upper bounds for hhs and area. For the other three cases, decoupled right shifted and right truncated Poisson distributions have to be calculated. A right shift indicates the probabilities below a lower bound do not exist. A right-truncation differs because the values above an upper bound do exist but are omitted in the final contingency table. Focusing on the implantation of a right shift first, it should be remembered that for a univariate Poisson distribution the λ parameter is just an average. Below shows how a right shift affects a λ value for a particular data set.

Proof. Consider the data set $X = x_1, x_2, ..., x_n$ with an average $\bar{x}$. For $1 \leq i \leq n$ set $y_i = x_i - l$ (where l is a representation of a lower bound). Below shows show $\mu_\mathrm{y} = \mu_\mathrm{x}$ - l.

$$\mu_y = \frac{\sum y_i}{n} = \frac{\sum x_i - l}{n} = \frac{1}{n}\sum x_i - \frac{1}{n}\sum l = \mu_x - \frac{1}{n}(l * n) = \mu_x - l$$

□

Therefore, a right shift is executed when a lower bound (l) is subtracted from a corresponding average (λ value). This computation is simple when we are given an average, but now consider the cases with the frequency table(s) and censored contingency table. Here we conducted a right shift by easily adjusting the categories of the tables. The count values will stay while each category value ($\mathrm{c_i}$) in a given table is altered by $\mathrm{c_i}$ - l. Giving the altered frequency table(s) to a customized likelihood function will then provide the desired μ_x - l. Lastly, to complete the right shifts for all cases, the model respectfully adds back the hhs and area lower bound values to the row and column category values of the final contingency table. This last step signifies the lower bounds representing the new zero values, and will be completed after Iterative Proportional Fitting.

Now let's bring in the upper bounds and right-truncation Poisson probabilities. In Equation (3), we show how Saffari [12] displayed a Poisson distribution for a right truncated variable Y_i. Note t_i is the truncation point for y_i.

$$P_T(Y_i = y_i) = \frac{e^{-\lambda_i} {\lambda_i}^{y_i}}{(y_i)!\Big(1 - \sum_{y_i=t_i+1}^{\infty} P(Y_i = y_i|x_i)\Big)} \tag{3}$$

Arbitrarily using Equation (3) for hhs, λ_{i} would be the right shifted λ (calculated by μ_{x}-l) while the right truncated value (t_{i}) is the hhs upper bound. The area variable will have the same format. Hence, the right shifted and right truncated Poisson distributions for hhs and area can easily be calculated. For hhs (rows) the original probabilities of *i = 0,1,...,*t_i change to *i = hhs lower bound, hhs lower bound + 1,..., (hhs upper bound - hhs lower bound)* , while for area (columns) the original probabilities change to *i = area lower bound, area lower bound + 1,..., (area upper bound - area lowerbound).*

2.3 Iterative Proportional Fitting (IPF)

No affiliation of hhs to area can be developed with decoupled information. So for Cases I, II, and III, the right shifted and right truncated Poisson distributions of hhs and area represent the marginals in the Iterative Proportional Fitting (IPF) algorithm. IPF, introduced by Deming and Stephan [3], is an iterative least square adjustment method that can use known marginals to estimate joint probabilities. So in our case, the joint probabilities are the hhs x area cross tabulations. We will implement the original IPF algorithm because it is still the most common approach and standard population synthesizer and the Alaska Department of Labor [1] developed an IPF program in R that could simply be add to this package.

Fienberg [4] gave an easy to follow summarization involving the mathematics for the IPF procedure. The goal now is to replicate Fienberg's IPF explanation with our parameters. Let $\mathrm{hhs}^{\mathrm{r}}$ represent the last numeric row value (hhs upper bound - hhs lower bound) and $\mathrm{area}^{\mathrm{c}}$ be the last numeric column value (area upper bound - area lower bound). Then in the following equations, we can denote this model's fixed row (Equation (4)) and column (Equation (5)) marginal totals for IPF.

$$p_{i}\bullet = \sum_{j=0}^{area^c} p_{ij} \qquad (i = 0, 1, ..., hhs^r) \tag{4}$$

$$p\bullet_{j} = \sum_{i=0}^{hhs^r} p_{ij} \qquad (j = 0, 1, ..., area^c) \tag{5}$$

IPF acts as a weighting system. Each row in a fabricated matrix, known as a seed matrix, is proportionally adjusted to fit its corresponding row marginal. After the rows have converged to their row marginal (within a specified closure amount), the process is repeated for the columns. If assuming n_{ij}>0, IPF will eventually estimate individual cell probabilities in the hhs x area contingency table by minimizing Equation (6).

$$\sum_{i=0}^{hhs^r} \sum_{j=0}^{area^c} \frac{(n_{ij} - np_{ij})^2}{n_{ij}} \qquad (subject\,to\,equation\,4\,and\,5) \tag{6}$$

2.4 Censored Frequency Tables

Cases where a direct average is given easily represent the λ parameter needed for any further calculation in the model (right shift, right truncation poisson probabilities and then IPF). However, censored frequency tables are not as straightforward. Censored frequency tables have to be fit to a truncated Poisson distribution using a maximum likelihood function customized to handle left ($<$, $\leq$), right ($>$, $\geq$), and interval (-) censored data. The primary interested is to find the value for λ that maximizes the log likelihood of a univariate truncated Poisson distribution. To see an example function, Equation (7) represents the log-likelihood for uncensored (y), left censored ($y < c$), interval censored ($a \leq y \leq b$), and right censored ($y > d$) where *a, b, c,* and *d* represent censoring limits. The right shift is represented by letting *x = y - lower bound.* Also note, l signifies the lower bound quantity and u is the notation for an established upper bound value.

$$
\begin{aligned}
L(\lambda|x)_{log} = & \\
& \sum_{x} ln\Big(P(x|\lambda)\Big) \\
& + \sum_{x<c} ln\Big(P(x<c|\lambda)\Big) \\
& + \sum_{a\leq x\leq b} ln\Big(P(a\leq x\leq |\lambda)\Big) \\
& + \left(\sum_{x>d} ln\Big(P(x>d|\lambda)\Big) - \sum_{x>(u-l)} ln\Big(P(x>(u-l)|\lambda)\Big)\right)
\end{aligned}
\tag{7}
$$

A validation test to check the customized formula was also performed. First an arbitrary λ parameter computed a large randomized sample (10,000) of Poisson deviates. An uncensored univariate table was then produced by calculating the frequency of each unique deviate. The next step took this frequency table, and made adjustments to represent all types of censoring. The newly assembled censored frequency table could then be given to this log likelihood formula. The output returned the approximate λ value that was selected to make the table, thus validating the function. If the arbitrary $\lambda = 4$, then the method behind this test involves the R code of

```
table(rpois(10000, lambda = 4))
```

2.5 Censored Contingency Tables

With decoupled data (Case I, II, and III), IPF made estimation from completely unknown cross tabulations. The censored contingency case follows a different approach. There is already a cross tabulation provided in the table, but the values are distorted from censoring. The goal now becomes taking concealed values and estimating what these values would have been before censoring. To complete this reverse engineering technique, Kawamura's method [7] of reducing three parameters to a single parameter optimization problem can be implemented. Kawamura showed that for BP(X, Y) with corresponding Poisson marginals P(X|y) and P(Y|x) that the following relationships hold

$$
\begin{gathered}
\lambda_x = E(X) = \lambda_1 + \lambda_3 \\
\lambda_y = E(Y) = \lambda_2 + \lambda_3
\end{gathered}
$$

Where by definition of univariate Poisson, the marginal parameters λ_{x} and λ_{y} are given by the E(X|y) and E(Y|x) and the parameter selection problem is reduced to

$$
\begin{gathered}
\lambda_1 = E(X) \text{ - } \lambda_3 \\
\lambda_2 = E(Y) \text{ - } \lambda_3
\end{gathered}
$$

Or equivalently

$$
\begin{gathered}
\lambda_1 = \lambda_x \text{ - } \lambda_3 \\
\lambda_2 = \lambda_y \text{ - } \lambda_3
\end{gathered}
$$

Kawamura then applied a simple numerical approach to find λ_3 that maximizes the log likelihood function along the line

$$
0 \leq \lambda_3 \leq \min\,(\lambda_{\mathrm{x}},\lambda_{\mathrm{y}})
$$

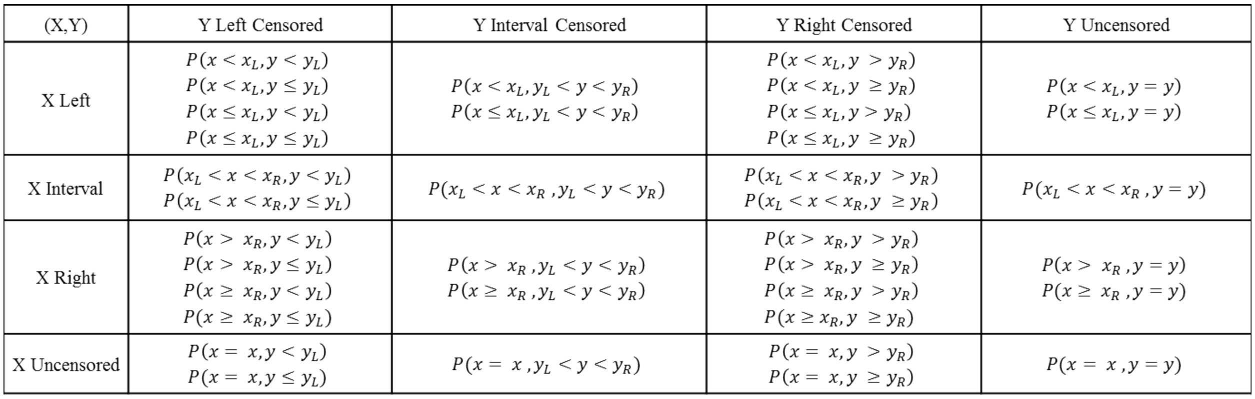
By building on this result, to find λ_3 let

$$
L\Bigg(\lambda_3|(X,Y),\lambda_x,\lambda_y\Bigg) = \prod_{i,j} L_{ij}\Bigg(\lambda_3|(X,Y),\lambda_x,\lambda_y\Bigg)
$$

where L_{ij} represents the likelihood component of the *(i, j)* type of censoring and is defined as

$$L_{ij}\Bigg(\lambda_3|(X,Y),\lambda_x,\lambda_y\Bigg) = \prod_{k=1}^{M} L_k\Bigg(\lambda_3|(X,Y),\lambda_x,\lambda_y\Bigg)$$

and $L_k\Big(\lambda_3|(X,Y),\lambda_x,\lambda_y\Big)$, k=1, M represents each likelihood components sharing a common censoring type *(i, j)*. All possibilities of the *(i, j)* censoring types are given in Table 1. Resembling the univariate case, notice the two forms of left and right censoring.



(X,Y)	Y Left Censored	Y Interval Censored	Y Right Censored	Y Uncensored
X Left	$P(x < x_L, y < y_L)$ $P(x < x_L, y \leq y_L)$ $P(x \leq x_L, y < y_L)$ $P(x \leq x_L, y \leq y_L)$	$P(x < x_L, y_L < y < y_R)$ $P(x \leq x_L, y_L < y < y_R)$	$P(x < x_L, y > y_R)$ $P(x < x_L, y \geq y_R)$ $P(x \leq x_L, y > y_R)$ $P(x \leq x_L, y \geq y_R)$	$P(x < x_L, y = y)$ $P(x \leq x_L, y = y)$
X Interval	$P(x_L < x < x_R, y < y_L)$ $P(x_L < x < x_R, y \leq y_L)$	$P(x_L < x < x_R, y_L < y < y_R)$	$P(x_L < x < x_R, y > y_R)$ $P(x_L < x < x_R, y \geq y_R)$	$P(x_L < x < x_R, y = y)$
X Right	$P(x > x_R, y < y_L)$ $P(x > x_R, y \leq y_L)$ $P(x \geq x_R, y < y_L)$ $P(x \geq x_R, y \leq y_L)$	$P(x > x_R, y_L < y < y_R)$ $P(x \geq x_R, y_L < y < y_R)$	$P(x > x_R, y > y_R)$ $P(x > x_R, y \geq y_R)$ $P(x \geq x_R, y > y_R)$ $P(x \geq x_R, y \geq y_R)$	$P(x > x_R, y = y)$ $P(x \geq x_R, y = y)$
X Uncensored	$P(x = x, y < y_L)$ $P(x = x, y \leq y_L)$	$P(x = x, y_L < y < y_R)$	$P(x = x, y > y_R)$ $P(x = x, y \geq y_R)$	$P(x = x, y = y)$

Table 1. Censoring types and corresponding likelihood forms. Here X will represent household size variable while Y will be area.

Again by assuming hhs and area follow a Poisson distribution, this implies a bivariate Poisson distribution with corresponding Poisson marginal exist. Continuing with Kawamura methodology, the maximum likelihood function for censored frequency tables allows the estimation of the needed λ_x and λ_y. These two parameters are then used to approximate an optimal λ_3. Subsequently, λ_1 and λ_2 values are calculated by simple arithmetic. Using λ_1, λ_2, and λ_3 a Bivariate Poisson distribution can be used to model the censor-uncensored data mixtures and produce an uncensored contingency table. The bivariate Poisson probability distribution R functions of **pbivpois** and **bivpois.table** [6] were implemented for this methodology.

Again, just as in the univariate case, another test to validate the λ bivariate parameters was performed. Random λ_1, λ_2, and λ_3 values were selected to compute probabilities for a bivariate Poisson distribution. These joint probabilities were put in matrix format to produce an uncensored contingency table. The contingency table was then subjected to censoring by adjusting the category values with their corresponding probabilities for both rows and columns. When given to our formula, the new censored bivariate table did return the approximate λ values that were originally selected. If λ_1 = 1, λ_2 =2, and λ_3=3 values were selected values, then the method behind this test involves the R code [6] of

```
bivpois.table(100, 100, c(1,2,3))
```

3 Data Entry

It is impossible to satisfy all formats found in census information. Trying to be as robust as possible, this revengc package still has specific requirements. First, there has to be hhs and area bounds meaning the input for the four cases will be the following:

Case I: hhs average, area average, hhs lower bound, hhs upper bound, area lower bound, and area upper bound
Case II: hhs frequency table, area frequency table, hhs lower bound, hhs upper bound, area lower bound, and area upper bound

Case III: hhs average or frequency table, area average or frequency table, hhs lower bound, hhs upper bound, area lower bound and area upper bound
Case IV: contingency table (hhs, area) or (area, hhs), hhs lower bound, hhs upper bound, area lower bound, and area upper bound

The tables also have to be formatted. The univariate frequency table found in Case I, Case II, and Case III must be formatted where there are two columns with n number of rows. The categories must be in the first column and the frequencies must be in the second column. Row names should never be placed in this table, the default name should always be 1:n where n is number of rows in the table. Both columns should not have a header (header=FALSE) and no words are allowed for censoring. The only censoring symbols accepted are < and ≤(left censoring), - (interval censoring), > and ≥ and + (right censoring). Table 2 shows a formatted table example for these cases.

≤ 6	11800
7-12	57100
13-19	14800
20+	3900

Table 2. Example of a formatted table for Case I, II, and III.

The table for Case IV also has restrictions. Again, no words are allowed for censoring. Only the censored values of < and ≤(left censoring), - (interval censoring), > and ≥ and + (right censoring) are permitted. This table works when there is a column header present or absent. However, the only column header that is allowed has to be the hhs or area category values. Row names should never be placed in this table, the default name should always be 1:n where n is number of rows in the table. The inside of this table is the cross tabulation of hhs x area which are either positive frequency values or percentages. The row and column total marginals have to placed in this table. The top left, top right, and bottom left corners of this table have to be NA or blank, but the bottom right corner can be a total sum value, NA, or blank. This code will transpose a contingency table if given a table with area as the rows and hhs as the columns, but the output will always be hhs as the rows and area as the columns. This transpose will only occur under the assumption that the sum of area category value is greater than the sum of household size category value. Table 3 is a formatted example with percentages as the cross-tabulations, the bottom right corner as a total sum, and the column header as the area category values.

NA	<20	20-30	>30	NA
<5	0.18	0.19	0.08	0.45
5-9	0.13	0.08	0.12	0.33
≥10	0.06	0.05	0.10	0.21
NA	0.38	0.32	0.31	1.00

Table 3. Example of a formatted table for Case IV.

3.1 Sample Datasets

Since the format for the tables is strict, we will now show how to format these tables properly using actual census data. If a user wants to read in a file, the format must look like the following sample datasets: *nepal_hhs, hongkong_hhs, hongkong_area, iran_hhs,* and *indonesia_contingency*. Creating tables with R code is possible too. The following code shows how these sample datasets can be created in R.

```
hhsdata_nepal<-cbind(as.character(c("1-2", "3-4", "5-6", "7-8", ">=9")),
  c(16.2, 41.7, 29.0, 9.0, 4.1))
```

```
hhsdata_hongkong<-cbind(as.character(c("1", "2", "3", ">3")),
  c(27600,25600,20900,13500))

areadata_hongkong<-cbind(as.character(c("<7", "7-12", "13-19", ">19")),
  c(11800,57100,14800,3900))

hhsdata_iran<-cbind(as.character(c("1", "2", "3", "4", ">=5")),
  c(7.08,18.29,29.64,27.95,17.04))

contingencytable<-matrix(c(6185,9797,16809,11126,6156,3637,908,147,69,4,
  5408,12748,26506,21486,14018,9165,2658,567,196,78,
  7403,20444,44370,36285,23576,15750,4715,994,364,136,
  4793,17376,44065,40751,28900,20404,6557,1296,555,228,
  2354,11143,32837,33910,26203,19301,6835,1438,618,245,
  1060,6038,19256,21298,17774,13864,4656,1039,430,178,
  273,2521,9110,11188,9626,7433,2608,578,196,112,
  119,1130,4183,5566,5053,3938,1367,318,119,66,
  33,388,1707,2367,2328,1972,719,171,68,37,
  38,178,1047,1672,1740,1666,757,193,158,164),
  nrow=10,ncol=10, byrow=TRUE)
rowmarginal<-apply(contingencytable,1,sum)
contingencytable<-cbind(contingencytable, rowmarginal)
colmarginal<-apply(contingencytable,2,sum)
contingencytable<-rbind(contingencytable, colmarginal)
row.names(contingencytable)[row.names(contingencytable)=="colmarginal"]<-""
contingencytable<-data.frame(c("1","2","3","4","5","6", "7", "8","9","10+", NA),
  contingencytable)
colnames(contingencytable)<-c(NA,"<20","20-29","30-39","40-49","50-69","70-99",
  "100-149","150-199","200-299","300+", NA)
```

4 Examples of Applying revengc to Census Data

4.1 Usage

First note that the main function in **revengc** is called **rec**. **rec** has the following format with a description of each argument directly below

```
rec(hhsdata, areadata, hhslowerbound, hhsupperbound, arealowerbound, areaupperbound)
```

hhsdata: This household size value can be a univariate frequency table or numeric value that represents an average. This input could also be a contingency table, but only if the areadata = 0.

areadata: This area (size of house) value can be a univariate frequency table or numeric value that represents an average. This input could also be a contingency table, but only if the hhsdata = 0. The areadata can be any unit of measure.

hhslowerbound: This is a numeric value to represent the household size lower bound. This lower bound variable needs to be numeric value ≥ 0.

hhsupperbound: This is a numeric value to represent the household size upper bound. This upper bound variable cannot be less than the highest category value (e.g. if a table has '>100' then the upper bound cannot be 90).

arealowerbound: This is a numeric value to represent the area lower bound. This lower bound variable needs to be numeric values ≥ 0.

areaupperbound: This is a numeric value to represent the area upper bound. This upper bound variable cannot be less than the highest category value (e.g. if a table has '>100' then the upper bound cannot be 90).

4.2 Nepal

The Nepal Living Standards Survey [9] provides averages and censored tables for household size and also averages for area of dwelling. This census data provides an example for Case I and Case III. To produce a final hhs x area contingency table (rows ranging from 1 to 20 people and columns ranging from 520 to 620 square feet) for urban Nepal you would run

```
#Case I
rec(4.4,571.3,1,20,520,620)
#Case II
rec(nepal_hhs,571.3,1,20,520,620)
```

4.3 Hong Kong

The Census and Statistics Department of Hong Kong [2] provides censored frequency tables for hhs and area as well as medians for both variables. This census data provides an example for Case I, Case II, and Case III. To produce a final hhs x area contingency table (rows ranging from 1 to 15 people and columns ranging from 1 to 30 square meters) for sub-divided units in Hong Kong you would run

```
#Case I
rec(2.0,10.3,1,15,1,30)
#Case II
rec(hongkong_hhs,hongkong_area,1,15,1,30)
#Case III
rec(2.0,hongkong_area,1,15,1,30)
rec(hongkong_hhs,10.3,1,15,1,30)
```

4.4 Iran

For different provinces, The Statistical Centre of Iran [14] reports averages and censored tables for household size as well as averages for floor area. This census data provides an example for Case I and Case III. To produce a final hhs x area contingency table (rows ranging from 1 to 10 people and columns ranging from 80 to 130 square meters) for East Azerbayejan (Azerbaijan), Iran you would run

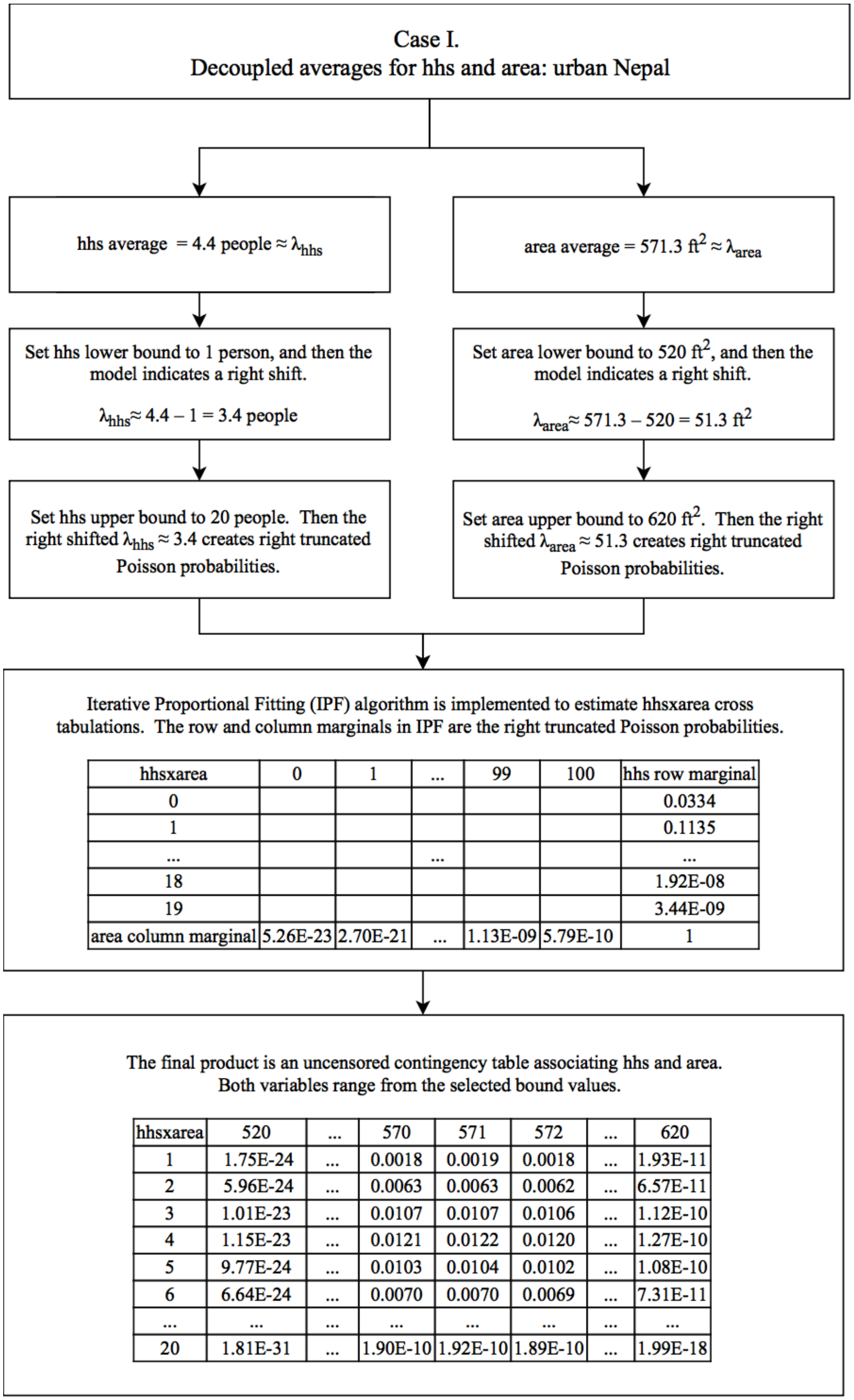
```
#Case I
rec(3.4,100.5,1,10,80,130)
#Case III
rec(iran_hhs,100.5,1,10,80,130)
```

4.5 Indonesia

The Population Census Data - Statistics Indonesia [11] provides over 60 censored contingency tables of Floor Area of Dwelling Unit (m2) x Household Member Size separated by province, urban, and rural. This census data provides a Case IV example. To produce a final hhs x area contingency table (rows ranging from 1 to 15 people and columns ranging from 10 to 310 square meters) for Indonesia 's rural Aceh Province you would run

```
#Case IV
rec(indonesia_contingency,0,1,15,10,310)
rec(0,indonesia_contingency,1,15,10,310)
```

4.6 Methodology Workflows Behind the rec Function



Case I.
Decoupled averages for hhs and area: urban Nepal

hhs average = 4.4 people ≈ λ_{hhs}

area average = 571.3 ft^2 ≈ λ_{area}

Set hhs lower bound to 1 person, and then the model indicates a right shift.

λ_{hhs}≈ 4.4 – 1 = 3.4 people

Set area lower bound to 520 ft^2, and then the model indicates a right shift.

λ_{area}≈ 571.3 – 520 = 51.3 ft^2

Set hhs upper bound to 20 people. Then the right shifted λ_{hhs} ≈ 3.4 creates right truncated Poisson probabilities.

Set area upper bound to 620 ft^2. Then the right shifted λ_{area} ≈ 51.3 creates right truncated Poisson probabilities.

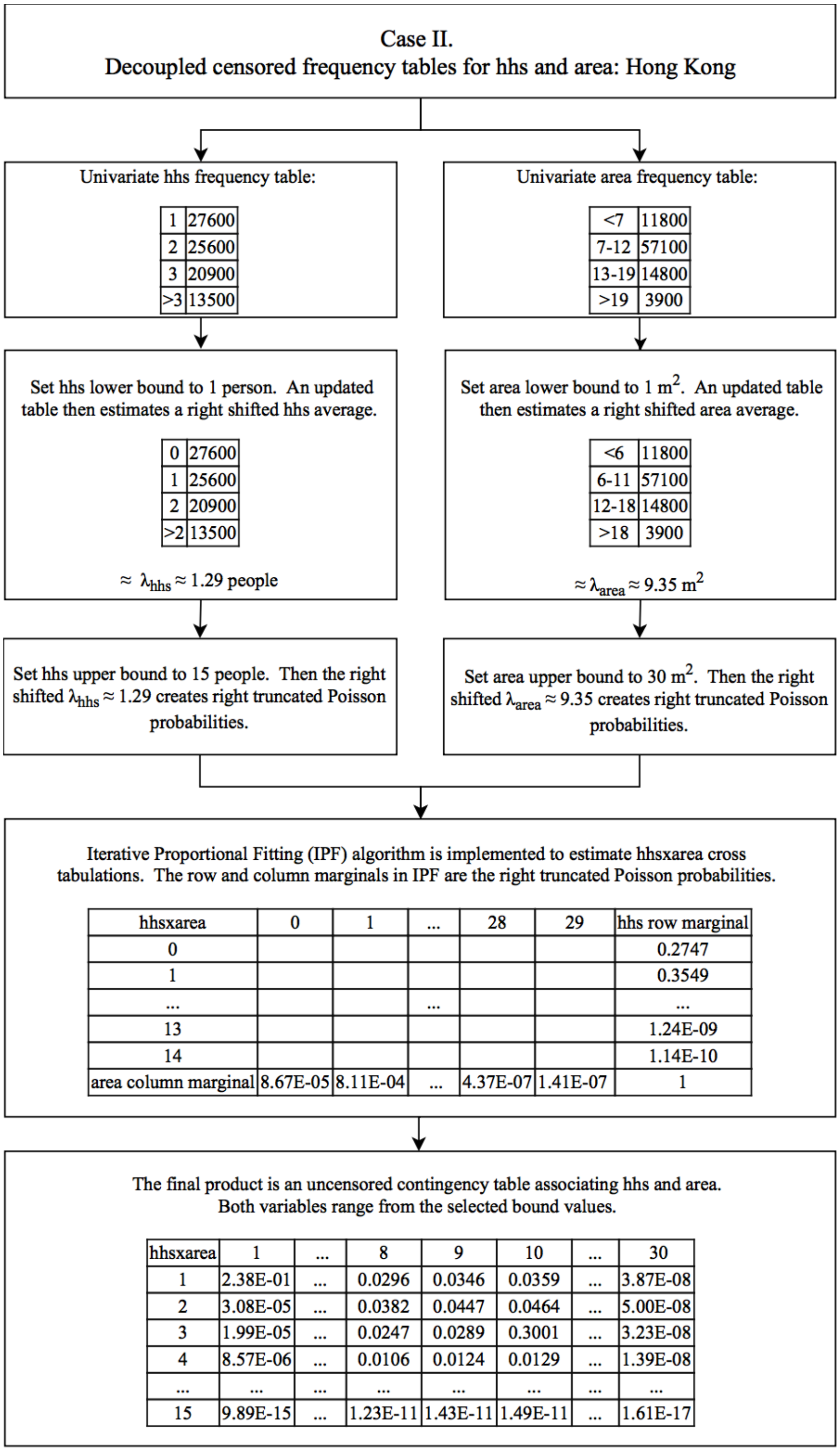
Iterative Proportional Fitting (IPF) algorithm is implemented to estimate hhsxarea cross tabulations. The row and column marginals in IPF are the right truncated Poisson probabilities.

hhsxarea	0	1	...	99	100	hhs row marginal
0						0.0334
1						0.1135
...			...			...
18						1.92E-08
19						3.44E-09
area column marginal	5.26E-23	2.70E-21	...	1.13E-09	5.79E-10	1

The final product is an uncensored contingency table associating hhs and area.
Both variables range from the selected bound values.

hhsxarea	520	...	570	571	572	...	620
1	1.75E-24	...	0.0018	0.0019	0.0018	...	1.93E-11
2	5.96E-24	...	0.0063	0.0063	0.0062	...	6.57E-11
3	1.01E-23	...	0.0107	0.0107	0.0106	...	1.12E-10
4	1.15E-23	...	0.0121	0.0122	0.0120	...	1.27E-10
5	9.77E-24	...	0.0103	0.0104	0.0102	...	1.08E-10
6	6.64E-24	...	0.0070	0.0070	0.0069	...	7.31E-11
...	...	...	...	...	...	...	...
20	1.81E-31	...	1.90E-10	1.92E-10	1.89E-10	...	1.99E-18

Workflow of Case I. We are only focus on household size and size of dwelling averages for urban Nepal [9] to illustrate a Case I methodology workflow. However, note that when the censored univariate table provided for household size (*nepal_hhs*) is adjusted to represent a right shift (selecting a lower bound of 1 person) this produces a comparable right shifted hhs average of 3.41 people.



Case II.
Decoupled censored frequency tables for hhs and area: Hong Kong

Univariate hhs frequency table:

1	27600
2	25600
3	20900
>3	13500

Univariate area frequency table:

<7	11800
7-12	57100
13-19	14800
>19	3900

Set hhs lower bound to 1 person. An updated table then estimates a right shifted hhs average.

0	27600
1	25600
2	20900
>2	13500

$\approx \lambda_{\text{hhs}} \approx 1.29$ people

Set area lower bound to 1 m^2. An updated table then estimates a right shifted area average.

<6	11800
6-11	57100
12-18	14800
>18	3900

$\approx \lambda_{\text{area}} \approx 9.35\ \text{m}^2$

Set hhs upper bound to 15 people. Then the right shifted $\lambda_{\text{hhs}} \approx 1.29$ creates right truncated Poisson probabilities.

Set area upper bound to 30 m^2. Then the right shifted $\lambda_{\text{area}} \approx 9.35$ creates right truncated Poisson probabilities.

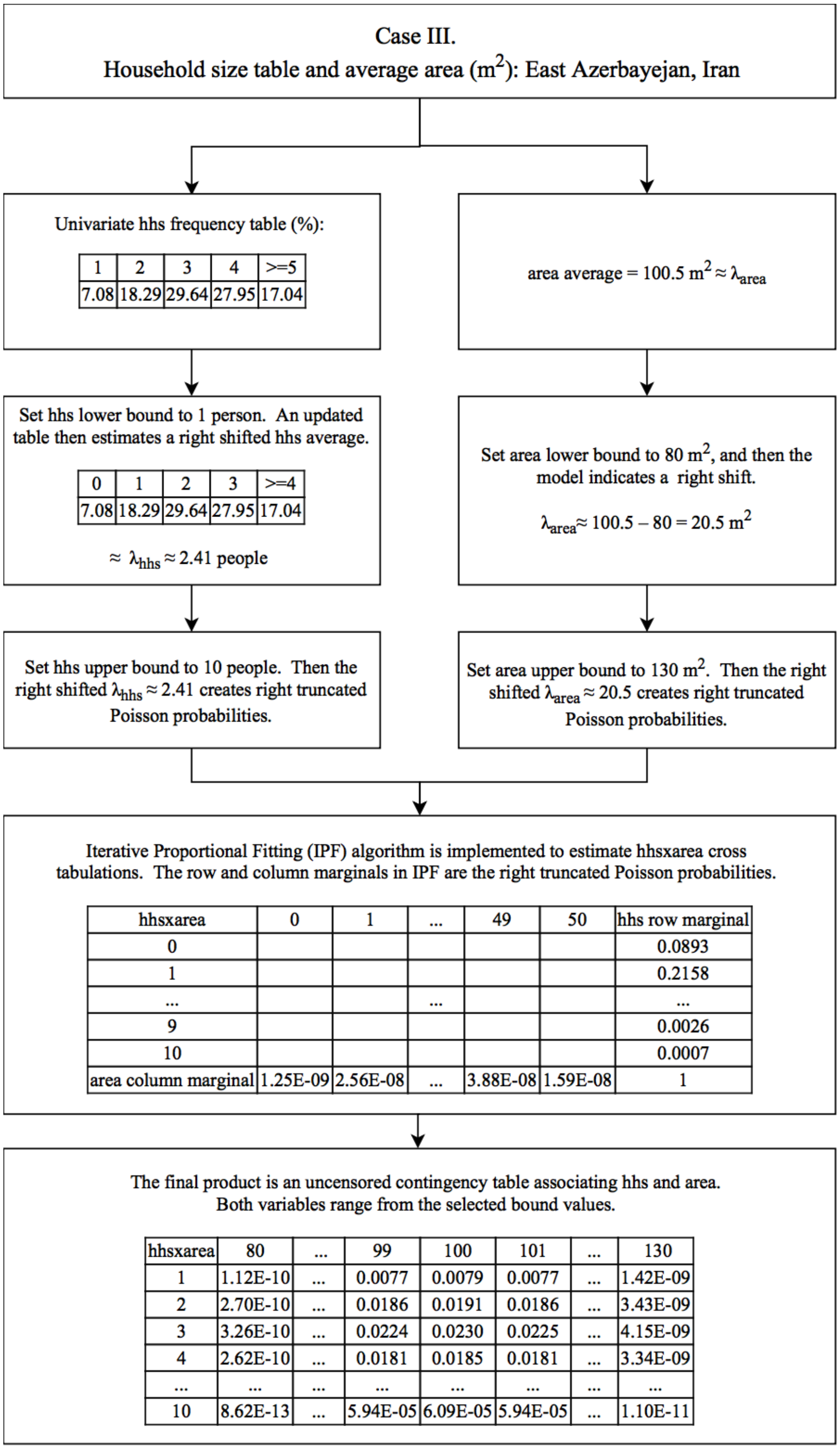
Iterative Proportional Fitting (IPF) algorithm is implemented to estimate hhsxarea cross tabulations. The row and column marginals in IPF are the right truncated Poisson probabilities.

hhsxarea	0	1	...	28	29	hhs row marginal
0						0.2747
1						0.3549
...			...			...
13						1.24E-09
14						1.14E-10
area column marginal	8.67E-05	8.11E-04	...	4.37E-07	1.41E-07	1

The final product is an uncensored contingency table associating hhs and area.
Both variables range from the selected bound values.

hhsxarea	1	...	8	9	10	...	30
1	2.38E-01	...	0.0296	0.0346	0.0359	...	3.87E-08
2	3.08E-05	...	0.0382	0.0447	0.0464	...	5.00E-08
3	1.99E-05	...	0.0247	0.0289	0.3001	...	3.23E-08
4	8.57E-06	...	0.0106	0.0124	0.0129	...	1.39E-08
...	...	...	...	...	...	...	...
15	9.89E-15	...	1.23E-11	1.43E-11	1.49E-11	...	1.61E-17

Workflow of Case II. For the provided 87,600 Hong Kong households in sub-divided units [2], we will focus on the frequency tables for household size and area (m^2) to illustrate a Case II methodology workflow. Keeping the same lower bound values of 1, notice how the estimated right shifted averages ($\lambda_{\text{hhs}} \approx 1.29$ people and $\lambda_{\text{area}} \approx 9.35$ m^2) are comparable to the right shifted averages provided in the census (median household size = 2-1=1 and median area = 10 -1= 9.3 m^2).



Case III.
Household size table and average area (m^2): East Azerbayejan, Iran

Univariate hhs frequency table (%):

1	2	3	4	>=5
7.08	18.29	29.64	27.95	17.04

area average = 100.5 $m^2 \approx \lambda_{area}$

Set hhs lower bound to 1 person. An updated table then estimates a right shifted hhs average.

0	1	2	3	>=4
7.08	18.29	29.64	27.95	17.04

$\approx \lambda_{hhs} \approx 2.41$ people

Set area lower bound to 80 m^2, and then the model indicates a right shift.

$\lambda_{area} \approx 100.5 - 80 = 20.5\ m^2$

Set hhs upper bound to 10 people. Then the right shifted $\lambda_{hhs} \approx 2.41$ creates right truncated Poisson probabilities.

Set area upper bound to 130 m^2. Then the right shifted $\lambda_{area} \approx 20.5$ creates right truncated Poisson probabilities.

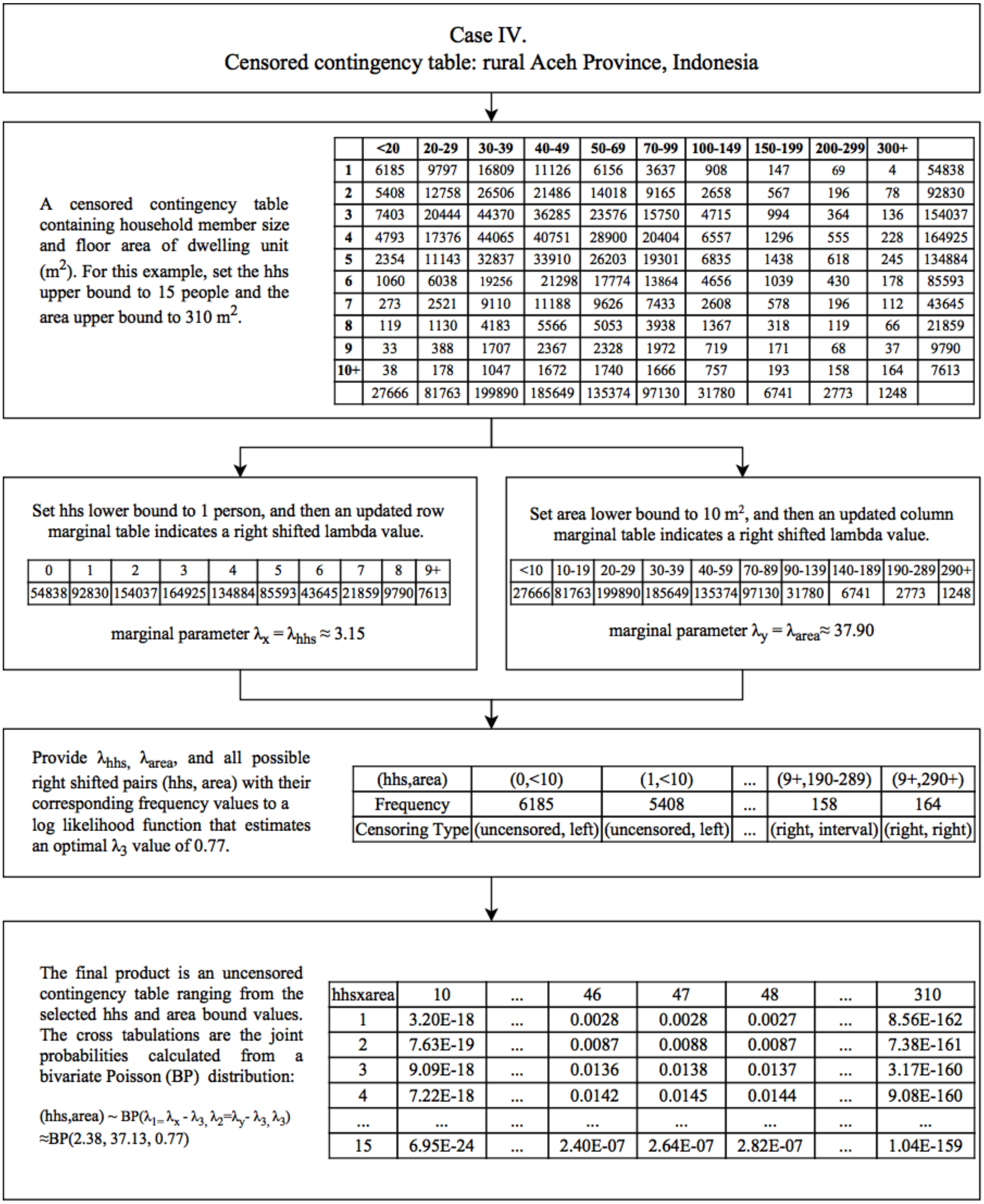
Iterative Proportional Fitting (IPF) algorithm is implemented to estimate hhsxarea cross tabulations. The row and column marginals in IPF are the right truncated Poisson probabilities.

hhsxarea	0	1	...	49	50	hhs row marginal
0						0.0893
1						0.2158
...			...			...
9						0.0026
10						0.0007
area column marginal	1.25E-09	2.56E-08	...	3.88E-08	1.59E-08	1

The final product is an uncensored contingency table associating hhs and area.
Both variables range from the selected bound values.

hhsxarea	80	...	99	100	101	...	130
1	1.12E-10	...	0.0077	0.0079	0.0077	...	1.42E-09
2	2.70E-10	...	0.0186	0.0191	0.0186	...	3.43E-09
3	3.26E-10	...	0.0224	0.0230	0.0225	...	4.15E-09
4	2.62E-10	...	0.0181	0.0185	0.0181	...	3.34E-09
...	...	...	...	...	...	...	...
10	8.62E-13	...	5.94E-05	6.09E-05	5.94E-05	...	1.10E-11

Workflow of Case III. For East Azerbayejan, Iran [14] we will use the household size table and the average floor area to illustrate a Case III worked example. With the same lower bound hhs value of 1 person, notice that the model's estimated right shifted hhs λ (2.41 people) is comparable to the right shifted hhs average value provided in the documentation (3.4 - 1 = 2.4 people).



Case IV.
Censored contingency table: rural Aceh Province, Indonesia

A censored contingency table containing household member size and floor area of dwelling unit (m^2). For this example, set the hhs upper bound to 15 people and the area upper bound to 310 m^2.

	<20	20-29	30-39	40-49	50-69	70-99	100-149	150-199	200-299	300+	
1	6185	9797	16809	11126	6156	3637	908	147	69	4	54838
2	5408	12758	26506	21486	14018	9165	2658	567	196	78	92830
3	7403	20444	44370	36285	23576	15750	4715	994	364	136	154037
4	4793	17376	44065	40751	28900	20404	6557	1296	555	228	164925
5	2354	11143	32837	33910	26203	19301	6835	1438	618	245	134884
6	1060	6038	19256	21298	17774	13864	4656	1039	430	178	85593
7	273	2521	9110	11188	9626	7433	2608	578	196	112	43645
8	119	1130	4183	5566	5053	3938	1367	318	119	66	21859
9	33	388	1707	2367	2328	1972	719	171	68	37	9790
10+	38	178	1047	1672	1740	1666	757	193	158	164	7613
	27666	81763	199890	185649	135374	97130	31780	6741	2773	1248	

Set hhs lower bound to 1 person, and then an updated row marginal table indicates a right shifted lambda value.

0	1	2	3	4	5	6	7	8	9+
54838	92830	154037	164925	134884	85593	43645	21859	9790	7613

marginal parameter $\lambda_x = \lambda_{hhs} \approx 3.15$

Set area lower bound to 10 m^2, and then an updated column marginal table indicates a right shifted lambda value.

<10	10-19	20-29	30-39	40-59	70-89	90-139	140-189	190-289	290+
27666	81763	199890	185649	135374	97130	31780	6741	2773	1248

marginal parameter $\lambda_y = \lambda_{area} \approx 37.90$

Provide $\lambda_{hhs,}$ λ_{area}, and all possible right shifted pairs (hhs, area) with their corresponding frequency values to a log likelihood function that estimates an optimal λ_3 value of 0.77.

(hhs,area)	(0,<10)	(1,<10)	...	(9+,190-289)	(9+,290+)
Frequency	6185	5408	...	158	164
Censoring Type	(uncensored, left)	(uncensored, left)	...	(right, interval)	(right, right)

The final product is an uncensored contingency table ranging from the selected hhs and area bound values. The cross tabulations are the joint probabilities calculated from a bivariate Poisson (BP) distribution:

$(hhs,area) \sim BP(\lambda_{1=} \lambda_x - \lambda_{3,} \lambda_2 = \lambda_y - \lambda_{3,} \lambda_3)$
$\approx BP(2.38, 37.13, 0.77)$

hhsxarea	10	...	46	47	48	...	310
1	3.20E-18	...	0.0028	0.0028	0.0027	...	8.56E-162
2	7.63E-19	...	0.0087	0.0088	0.0087	...	7.38E-161
3	9.09E-18	...	0.0136	0.0138	0.0137	...	3.17E-160
4	7.22E-18	...	0.0142	0.0145	0.0144	...	9.08E-160
...	...	...	...	...	...	...	...
15	6.95E-24	...	2.40E-07	2.64E-07	2.82E-07	...	1.04E-159

Workflow of Case IV. Indonesia's rural Aceh Province [11] is used to illustrate a Case IV worked example.

5 Conclusion

Census data frequently provides residential household size and area values. These are key attributes when estimating interior density, but the problem is when these variables are reported as summarized information. Averages, decoupled variables, and censored tables do not provide clear results and overall obscures the original data. The statistical **revengc** package was designed to reverse engineer this censored, decoupled census data. The final result is an uncensored hhs x area contingency table that can be used more accurately for estimating interior residential occupancy. All efforts go towards improving residential occupancy found in census data.

References

[1] Alaska Department of Labor and Workforce Development (2009). Iterative Proportional Fitting Information and Code. Retrieved from: `http://u.demog.berkeley.edu/~eddieh/datafitting.html`

[2] Census and Statistics Department of Hong Kong Special Administrative Region. (2016). *Thematic Household Survey Report - Report No. 60 - Housing conditions of sub-divided units in Hong Kong.* Retrieved from: http://www.censtatd.gov.hk/hkstat/sub/sp100.jsp?productCode=C0000091

[3] Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when expected marginal totals are known. *Annals of Mathematical Statistics*, 11 (4), 427-444.

[4] Fienberg, S. E. (1970). An Iterative Procedure for Estimation in Contingency Tables. *Annals of Mathematical Statistics*, 41 (3), 907-917.

[5] Jennings V, Lloyd-Smith B, Ironmonger D (1999). Household size and the poisson distribution. *Journal of the Australian Population Association*, 16, 65–84.

[6] Karlis, D., and Ntzoufras, I. (2005). Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. *Journal of Statistical Software*, 14 (10), 1-36.

[7] Kawamura, K. (1984). Direct calculation of maximum likelihood estimator for the bivariate Poisson distribution. *Kodai Math. J.,* 7 (2), 211-221.

[8] Klaiber Jr, H. A. (2008). *Valuing open space in a locational equilibrium model of the Twin Cities.* North Carolina State University.

[9] National Planning Commissions Secretariat, Government of Nepal. (2011). *Nepal Living Standards Survey.* Retrieved from: `http://siteresources.worldbank.org/INTLSMS/Resources/3358986-1181743055198/3877319-1329489437402/Statistical_Report_Vol1.pdf`

[10] Ohnishi, T., Mizuno, T., Shimizu, C., and Watanabe, T. (2010). On the evolution of the house price distribution.

[11] Population Census Data - Statistics Indonesia. (2010). *Household by Floor Area of Dwelling Unit and Households Member Size.* Retrieved from: `http://sp2010.bps.go.id/index.php/site/tabel?wid=1100000000&tid=334&fi1=586&fi2=`

[12] Saffari, S. E., Adnan, R., and Greene, W. (2011). Handling of over-dispersion of count data via truncation using Poisson regression model. *Journal of Computer Science and Computational Mathematics*, 1(1), 1-4.

[13] Shuman, A. (2011, January 10). Behrens Ranch Round Rock TX Market Report 2009 – 2010. Retrieved July 10, 2014, from `http://theappraisaliq.com/behrens-ranch-round-rock-tx-market-report-2009-2010/`

[14] The Statistical Centre of Iran. (2011). *Selected Findings of National Population and Housing Census.* Retrieved from: `https://www.amar.org.ir/Portals/1/Iran/90.pdf`

[15] Stewart, R., Urban, M., Duchscherer, S., Kaufman, J., Morton, A., Thakur, G., ..., and Moehl, J. (2016). A Bayesian machine learning model for estimating building occupancy from open source data. *Natural Hazards*, 81 (3), 1929-1956.