Preliminary Draft. Not for circulation.

The G-Decomposition: Estimating Group-Specific Contributions to Life-Table Functions

Eugenio Paglino

Department of Sociology
Population Studies Center
University of Pennsylvania

Introduction

The decomposition of demographic measures is a fundamental tool in a demographer's toolkit. Through decomposition of demographic measures, researchers can gain crucial insights into the importance of different mechanisms in producing demographic change. Most decomposition techniques decompose changes in demographic measures over time or, more generally, differences between demographic rates (Andreev, Shkolnikov, and Begun 2002; Arriaga 1984; Caswell 1989; Das Gupta 1978; Horiuchi, Wilmoth, and Pletcher 2008; Kitagawa 1964; Vaupel and Canudas-Romo 2002). However, fewer techniques are available to researchers needing to estimate the contribution of different subpopulations to a demographic measure for the total population. This paper develops such a decomposition technique and discusses one practical application.

Relationship

Given a population (T) and a partition of the population into mutually exclusive subgroups ($G_1, G_2, \dots, G_N$), how can we compute the contribution of subgroups $G_1, G_2, \dots G_N$ to life table functions ($l_x, {}_nL_x, {}_nT_x, e_x$) starting from the life table of a baseline population (B)? This paper proposes a decomposition method to answer this question and proves some of its properties. It then illustrates one practical application of the proposed decomposition.

The basic form of the decomposition method developed in this paper was originally proposed by Hendi and Ho (2021) to investigate the contribution of the foreign born population to the national life expectancy at age 1 in the United States (Hendi and Ho 2021). In that context, Hendi and Ho considered a total population T including all residents of the United States, a baseline population B including only US-born residents, and a single contributing group G including all foreign-born residents. They proposed to compute the contribution of the foreign-born population to the national life expectancy at age 1 as:

$$C^G = e_1^T - e_1^B$$

Where e_1^T is the life expectancy at age 1 for the total population and e_1^B is the life expectancy at age 1 for US-born residents (the baseline population). The logic behind this formula is clear, the difference between the life expectancy of all US residents and US-born residents must be explained by the contribution of the foreign-born population. However, using Hendi and Ho (2020) example, if x^{USB} is the vector of period age-specific mortality rates for US-born residents, x^{FB} is the vector of period age-specific mortality rates for the foreign-born residents, and x^T is the vector of age-specific mortality rates for the total population, and assuming $e_1^{FB} \geq e_1^{USB}$, there is no guarantee that:

$$e_1^{USB} \leq e_1(x^{FB}w^{FB} + x^{USB}w^{USB}) = e_1^T \leq e_1^{FB}$$

Where w^{FB} and w^{USB} are vectors of age-specific weights proportional to the share of the foreign-born and US-born in the population, respectively, and such that $w_i^{FB} + w_i^{USB} = 1 \; \forall i$. The two inequalities fail to hold because a priori:

$$\max\left(e_1^T(x^{FB}w^{FB} + x^{USB}w^{USB})\right) \; w.r.t. \; (w^{FB}, w^{USB}) = e_1(\min(x^{FB}, x^{USB}))$$

where $\min(x^{FB}, x^{USB})$ is the set of elementwise (age-specific) minimums and:

$$\min\left(e_1^T(x^{FB}w^{FB} + x^{USB}w^{USB})\right) \; w.r.t. \; (w^{FB}, w^{USB}) = e_1(\max(x^{FB}, x^{USB}))$$

where $\max(x^{FB}, x^{USB})$ is the set of elementwise (age-specific) maximums. Equivalently, as long as the two sets of mortality rates x^{FB}, x^{USB} cross at some point (i.e. $\exists i, j \; s.t. \; x_i^{FB} > x_i^{USB} \wedge x_j^{FB} < x_j^{USB}$), there will

be a set of weights such that $e_1^T > \max(e_1^{FB}, e_1^{USB})$ and also a set of weights such that $e_1^T < \min(e_1^{FB}, e_1^{USB})$. The failure of e_1^T to be bounded by (e_1^{FB}, e_1^{USB}) has two consequences that affect how we think of the contribution of a group G to the total life-expectancy given a certain baseline population:

1. It can happen that $C^G = e_1^T - e_1^B > e_1^G - e_1^B$. In words, a group G could contribute more years to the total life expectancy than the difference between its life expectancy and that of the baseline population.
2. Suppose we exchange the role of the contributing group G and that of the baseline population B. We would then define: $C^B = e_1^T - e_1^G$. It is possible that both $C^B > 0$ and $C^G > 0$ even though B and G, by definition, form the total population T. Of course, it is also possible that $C^B < 0$ and $C^G < 0$.

Given these two facts, it is important to understand that the contribution of group G computed with the Hendi and Ho (2020) formula, depends on the mortality rates of the baseline population and thus on the choice of a baseline. More precisely, the contribution C^G does not capture the effect of group G "in isolation" but also its interaction with the mortality conditions of the baseline population. Keeping this is mind, when the choice of a baseline does not have strong theoretical foundations, researchers should test the robustness of the results to the choice of a different baseline and, if mayor differences are found, discuss why it might be the case. While the choice of a baseline may seem intuitive in the case of two populations, it will become less obvious when multiple groups are considered, such as populations living in different Census Divisions, or in counties classified along the urban-rural continuum.

Similar issues of ordering affecting the decomposition results arise in other decomposition techniques as well such as the Step-Wise decomposition (Andreev et al. 2002) and Arriaga's decomposition (Arriaga 1984). These issues have been usually solved by averaging over different orderings (Andreev 1982; Andreev et al.

2002; Pressat 1985) which however becomes computationally more challenging as the possible number of orderings to consider increases.

A Decomposition for $(N = 2)$

Let us consider a population (T) that is composed of a baseline population (B) and two mutually exclusive contributing groups (G_1, G_2). Let us also denote the population combining (G_1, G_2) as G. Finally, let us be as general as we can and leave the life-table function of interest unspecified. While this might seem strange, this is the same approach taken by all major general decomposition tools (Andreev et al. 2002; Caswell 1989; Horiuchi et al. 2008). We will denote with $l(A)$ the life-table function for population A. Hendi and Ho's expression for the total contribution of the two groups is still valid:

$$C^G = l(T) - l(B)$$

Because both the person-years and the death counts involved in the computation of life-tables combine additively, we have that $l(B) = l(T - G)$. Equivalently, we could write $l(T) = l(B + G)$. Together, these two equations allow us to write:

$$C^G = l(T) - l(B) = l(B + G) - l(B) = l(T) - l(T - G)$$

Looking at the equation above it is apparent that the total contribution of population G can equivalently be understood as an addition process starting from the baseline population $C^G = l(B + G) - l(B)$ or as a subtraction process starting from the total population $C^G = l(T) - l(T - G)$. It now becomes useful to introduce another layer of notation. We define $C^{G+} \equiv l(B + G) - l(B)$ and call it the "addition contribution" of G. We instead define $C^{G-} \equiv l(T) - l(T - G)$ and call it the "subtraction contribution" of G. Clearly, when a single contributing population is considered $C^{G+} = C^{G-}$. However, we will see that even in the case of two

contributing subpopulations (G_1, G_2), $C^{1+} \neq C^{1-}$ and $C^{2+} \neq C^{2-}$ so that the notation I just introduced becomes useful. To see why these two inequalities arise, let us consider C^{1+} and C^{1-}. Suppose for the moment that both G_1 and G_2 contribute positively to the life table function we are considering, and more precisely that $C^{1+}, C^{1-}, C^{2+}, C^{2-} > 0$. When we compute C^{1-} we are then starting from a population T which already encapsulates the positive effect of population G_2. On the contrary, when we compute C^{1+} we are starting from the baseline population B that has not yet received the positive effect of population G_2. Furthermore, population B from which are starting in our calculations for C^{1+} is smaller than population $B + G_2$ from which we start in our calculations for C^{1-}. Consequently, the effect of G_1 will be larger when computed as C^{1+} than when computed as C^{1-} and $C^{1+} > C^{1-}$. A specular argument can be made for C^{2+}, C^{2-}. While the inequality $C^{1+} > C^{1-}$ depends on the assumption that both populations have a positive effect, I have illustrated the general principle that will lead to the inequality $C^{1+} \neq C^{1-}$.

The inequality above means that we now have two different measures of the contribution of each group. One way of solving this inconsistency is to introduce a third type of contribution $C^1 \equiv \frac{1}{2}(C^{1+} + C^{1-})$, which we will call the "average contribution". It turns out that the average contributions have a nice property that both the addition and the subtraction contributions generally lack.

$$
\begin{aligned}
C^1 + C^2 &= \frac{1}{2}(C^{1+} + C^{1-}) + \frac{1}{2}(C^{2+} + C^{2-}) \\
&= \frac{1}{2}[(l(B + G_1) - l(B)) + (l(T) - l(T + G_1)) + (l(B + G_2) - l(B)) + (l(T) - l(T + G_2))] \\
&= \frac{1}{2}\big[2\big(l(T) - l(B)\big) + \big(l(B + G_1) - l(T - G_2)\big) + \big(l(B + G_2) - l(T - G_1)\big)\big] \\
&= \big(l(T) - l(B)\big) + \frac{1}{2}\big[\big(l(B + G_1) - l(T - G_2)\big) + \big(l(B + G_2) - l(T - G_1)\big)\big] \\
&= C^G + \frac{1}{2}\big[\big(l(B + G_1) - l(T - G_2)\big) + \big(l(B + G_2) - l(T - G_1)\big)\big]
\end{aligned}
$$

$$= C^G + \frac{1}{2}\left[\left(l(B+G_1) - l(B+G_1)\right) + \left(l(B+G_2) - l(B-G_2)\right)\right]$$

$$= C^G$$

So, the two average contributions sum to the total contribution of G, which establishes C^1, C^2 as a legitimate decomposition of C^G. Note that the substitutions in the second last line hold because of the additivity of person-years and deaths so $l(T - G_2) = l(B + G_1 + G_2 - G_2) = l(B + G_1)$ and similarly $l(T - G_1) = l(B + G_2)$. Notice also that no special properties of the life-table function were used in the proof which establishes that this approach is valid for functions other than life expectancy.

Extending the Decomposition to the Case $(N = 3)$

The case of three groups (G_1, G_2, G_3) introduces a few additional complications which illuminate some properties of this decomposition and will allow us to find a general expression for any number of groups. The main difference with the $(N = 2)$ case is that our notation C^{n+}, C^{n-} for $n = 1,2,3$ is no longer sufficient to describe all possible ways of computing the contribution of group G_n to the life-table function for the total population. Indeed, we now have four ways of computing the contribution of G_1:

1. $l(B + G_n) - l(B)$
2. $l(B + G_2 + G_1) - l(B + G_2)$
3. $l(B + G_3 + G_1) - l(B + G_3)$
4. $l(T) - l(B + G_2 + G_3)$

Number 1 and 4 are equal to C^{1+} and C^{1-}, but number 2 and 3 are outside of what we have seen so far. In the next section, I will introduce a more general notation to capture these cases. However, for the moment, let us focus on how to define the average contribution in this case. It turns out that the average contribution in this

case assigns weights to each term that are inversely proportional to the number of groups $N = 3$ and to the number of contributions involving the same number of subpopulations. Notice that contributions 1 and 4 involve 1 and 3 subpopulations respectively, while contributions 2 and 3 each involve 2 subpopulations. As such, contributions 1 and 4 have have weight $\frac{1}{N}\frac{1}{1} = \frac{1}{3}$, while contributions 2 and 3 have weight $\frac{1}{N}\frac{1}{2} = \frac{1}{3}\frac{1}{\mathbf{2}} = \frac{1}{6}$. The use of these weights is equivalent to first averaging within contributions involving the same number of subpopulations and then averaging between contributions involving different numbers of subpopulations. While this might seem unintuitive, I will now show that it leads to a set of average contributions that sum to the total contribution as in the $(N = 2)$ case.

$$C^1 + C^2 + C^3 = \frac{1}{3}(l(B + G_1) - l(B)) + \frac{1}{6}\big(l(B + G_2 + G_1) - l(B + G_2)\big) + \frac{1}{6}\big(l(B + G_3 + G_1) - l(B + G_3)\big) + \frac{1}{3}\big(l(T) - l(B + G_2 + G_3)\big)$$
$$+\frac{1}{3}(l(B + G_2) - l(B)) + \frac{1}{6}\big(l(B + G_1 + G_2) - l(B + G_1)\big) + \frac{1}{6}\big(l(B + G_3 + G_2) - l(B + G_3)\big) + \frac{1}{3}\big(l(T) - l(B + G_1 + G_3)\big)$$
$$+\frac{1}{3}\big(l(B + G_3) - l(B)\big) + \frac{1}{6}\big(l(B + G_1 + G_3) - l(B + G_1)\big) + \frac{1}{6}\big(l(B + G_2 + G_3) - l(B + G_2)\big) + \frac{1}{3}\big(l(T) - l(B + G_1 + G_2)\big)$$

It is easy to see that one can take out of this complicated expression the target value $l(T) - l(B)$. We can then reorder the remaining terms and write:

$$C^1 + C^2 + C^3 = \big(l(T) - l(B)\big) + \frac{1}{3}(l(B + G_1) + l(B + G_2) + l(B + G_3))$$
$$- \frac{1}{3}\big(l(B + G_1 + G_2) + l(B + G_1 + G_3) + l(B + G_2 + G_3)\big)$$
$$- \frac{1}{6}(2l(B + G_1) + 2l(B + G_2) + 2l(B + G_3))$$
$$+ \frac{1}{6}(2l(B + G_1 + G_2) + 2l(B + G_1 + G_3) + 2l(B + G_2 + G_3))$$

From which one can easily verify that $C^1 + C^2 + C^3 = \big(l(T) - l(B)\big)$ as we wanted to prove. This result establishes (C^1, C^2, C^3) as a legitimate decomposition of C^G.

By comparing the computations involved in the case $(N = 2)$ to those for the case $(N = 3)$, it becomes apparent that the computational requirement for this method increases very fast as the number of groups increases. Luckily, one can show that if we are willing to make some assumptions, the computations can be greatly simplified without loss of precision. Let us assume that:

$$\frac{1}{2}\big(l(B + G_2 + G_1) - l(B + G_2)\big) + \frac{1}{2}\big(l(B + G_3 + G_1) - l(B + G_3)\big) = \frac{1}{2}(l(B + G_1) - l(B)) + \frac{1}{2}\big(l(T) - l(G_2 + G_3)\big)$$

For the four contributions involved in C^1 and that equivalent expressions hold for C^2 and C^3. The expression above is requiring that the average of the two "internal" contributions (2 and 3) is equal to the average of the two "external" contributions (1 and 4). Then, for G_1, we can write:

$$\begin{aligned}
C^1 &= \frac{1}{3}\big(l(B + G_1) - l(B)\big) + \frac{1}{6}\big(l(B + G_2 + G_1) - l(B + G_2)\big) + \frac{1}{6}\big(l(B + G_3 + G_1) - l(B + G_3)\big) + \frac{1}{3}\big(l(T) - l(B + G_2 + G_3)\big) \\
&= \frac{1}{3}\big(l(B + G_1) - l(B)\big) + \frac{1}{6}\big(l(B + G_2 + G_1) - l(B + G_2) + l(B + G_3 + G_1) - l(B + G_3)\big) + \frac{1}{3}\big(l(T) - l(B + G_2 + G_3)\big) \\
&= \frac{1}{3}\big(l(B + G_1) - l(B)\big) + \frac{1}{3}\frac{1}{2}\big(l(B + G_2 + G_1) - l(B + G_2) + l(B + G_3 + G_1) - l(B + G_3)\big) + \frac{1}{3}\big(l(T) - l(B + G_2 + G_3)\big) \\
&= \frac{1}{3}\big(l(B + G_1) - l(B)\big) + \frac{1}{3}\frac{1}{2}\big(l(B + G_1) - l(B) + l(T) - l(B + G_2 + G_3)\big) + \frac{1}{3}\big(l(T) - l(B + G_2 + G_3)\big) \\
&= \frac{1}{3}\big(l(B + G_1) - l(B)\big) + \frac{1}{6}(l(B + G_1) - l(B)) + \frac{1}{6}(l(T) - l(B + G_2 + G_3)) + \frac{1}{3}\big(l(T) - l(B + G_2 + G_3)\big) \\
&= \frac{1}{2}\big(l(B + G_1) - l(B)\big) + \frac{1}{2}\big(l(T) - l(B + G_2 + G_3)\big) \\
&= \frac{1}{2}(C^{1+} + C^{1-})
\end{aligned}$$

The equation $C^1 = \frac{1}{2}(C^{1+} + C^{1-})$ tells us that as long as we are willing to assume that the two internal contributions and the two external contributions have the same average (an assumption that can be verified

empirically), we can ignore the internal terms and simply compute the average contribution of each group as the average the two external contributions as we did for the two group case. With this result we have made the decomposition computationally scalable even when many groups are involved. In the next section I will prove that the same results hold even in the case of N groups.

Generalizing the Decomposition to a Population with N subgroups

To study the general case of N contributing groups, it is useful to introduce a slightly more general notation. We define:

$$C^{n,m,i} = l(B + G_i^m + G_n) - l(B + G_i^m)$$

With $n = \{1,2, \dots, N\}, m = \{0,1, \dots, N-1\}, i = \{1,2, \dots, \binom{N-1}{m}\}$ and where G_i^m is the i^{th} set of subpopulations G_k with $k \neq n$ such that $\#G_i^m = m$. In words, $C^{n,m,i}$ is the contribution of subgroup G_n to the life-table function $l()$ for the total population calculated as the difference between the value of $l()$ for the population obtained by combining the baseline population, the i^{th} possible subset of G_k's not including G_n, and G_n ($l(B + G_i^m + G_n)$), and the value of $l()$ for the population obtained by combining the baseline population and the i^{th} possible subset of G_k's not including G_n ($l(B + G_i^m)$). This definition might seem confusing but it's just a generalization of the types of contributions we have seen in the cases $N = 2, 3$. Indeed, for $N = 2$, our familiar C^{1+} is simply $C^{1,0,1}$ while C^{1-} is $C^{1,1,1}$. For $N = 3$, we have the more interesting "internal" terms for which we had no notation so far. With this new notation we can write:

1. $l(B + G_2 + G_1) - l(B + G_2) = C^{1,1,1}$
2. $l(B + G_3 + G_1) - l(B + G_3) = C^{1,1,2}$

With these expressions, we can see why m should be bounded between 0 and $N-1$ and i between 1 and $\binom{N-1}{m}$. With $m=0$, we recover C^{1+}, when $m=0$, there is no need for an index i which we can conventionally set to 1. For $m=1$, G_i^m contains only one element which we can choose out of the $N-1$ groups that are not G_n. The first of these groups will be denoted as G_1^1, the second as G_2^1, and so on. The order in which we select the groups does not really matter. Notice that, for a general m there will be $\binom{N-1}{m}$ of these groups which explains the limits imposed on i. Finally, when we reach $m=N-1$, we obtain C^{1-} because:

$$l(B+G_1^{N-1}+G_1)-l(B+G_1^{N-1})=l(T)-l(T-G_1)=C^{1-}$$

Clearly, there is only one way of choosing $N-1$ elements from a set of $N-1$ elements.

As for the $N=3$ case, we now need to find appropriate weights for each contribution $C^{n,m,i}$. The general principle is still the same, we need weights that are inversely proportional to the number of contribution types N and inversely proportional to the number of contributions for the specific type m. This consideration leads to a simple expression:

$$w^m=\frac{1}{N\binom{N-1}{m}}$$

Which is just a generalization of the weights we derived for $(N=3)$. In the expression above w^m denotes the weight for all terms $C^{n,m,i}$. These weights are equal for all m-type contributions and ensure that each set of m-contributions is collectively assigned the same weight $\frac{1}{N}$ while each of its members is also assigned the

same weight $\frac{1}{N\binom{N-1}{m}}$. With our new notation and having defined appropriate weights, we can define the average contribution of group G_n as:

$$C^n = \sum_{m=0}^{N-1} \frac{1}{N} \sum_{i=1}^{\binom{N-1}{m}} \frac{1}{\binom{N-1}{m}} C^{n,m,i}$$

To establish that average contributions defined in this way form a valid decomposition of C^G into group-specific contributions, we just need to prove that:

$$\sum_{n=1}^{N} C^n = l(T) - l(B)$$

We already saw that for $N = 2, 3$, the proof involves showing that all terms except $l(T)$ and $l(B)$ have weights summing to 0. This is harder to do directly now that we have many terms. However, we can start by recognizing that the life-table function computed for a given population composed by a set of subgroups plus the baseline population can appear in two ways:

1. As a term of the form $l(B + G_i^m + G_n)$ in a contribution of the type $C^{n,m,i}$
2. As a term of the form $l(B + G_i^{m+1})$ in a contribution of the type $C^{k,m,i}$ where $k \neq n$.

In the first case, the life-table function will have weight w^m and in the second case it will have weight $-w^{m+1}$. To understand what the final weight will be for each term $l(B + G_i^m + G_n)$, we just need to know how many times it will appear with weight w^m and how many times with weight w^{m+1}. For terms of the first type, we have $m + 1$ ways of choosing n while keeping the subgroups involved the same. On the other

hand, terms of the second type can only appear in contributions involving the $N-(m+1)$ excluded groups. With this information, we are now able to compute the weight associated with the life-table function computed for each population:

$$
\begin{aligned}
w(l(B+G_i^m+G_n)) &= (m+1)w^m-(N-m-1)w^{m+1} \\
&= (m+1)\frac{1}{N\binom{N-1}{m}}-(N-m-1)\frac{1}{N\binom{N-1}{m+1}} \\
&= (m+1)\frac{(N-1)!}{m!\,(N-1-m)!}-(N-m-1)\frac{(N-1)!}{(m+1)!\,(N-1-m-1)!} \\
&= (m+1)\frac{m!\,(N-1-m)!}{(N-1)!}-(N-m-1)\frac{(m+1)!\,(N-1-m-1)!}{(N-1)!} \\
&= \frac{(m+1)!\,(N-1-m)!}{(N-1)!}-\frac{(m+1)!\,(N-1-m)!}{(N-1)!} \\
&= 0
\end{aligned}
$$

Which proves that life-table functions for all populations of the type $B+G_i^m+G_n$ cancel out. The only two exceptions are $l(B)$, which appears N times with weight $-\frac{1}{N}$, and $l(T)$, which appears N times with weight $\frac{1}{N}$. Thus, summing the average contributions for our N groups we obtain:

$$\sum_{n=1}^{N} C^n = l(T)-l(B)$$

as desired. This result shows that the average contributions provide a decomposition of the total contribution C^G even in the case of N subgroups. While this is an important result, the number of terms involved in this decomposition grows very fast with N, making computing the full decomposition impractical. In the $N=3$ case, I showed that a convenient assumption allows us to compute the decomposition using just the two

external terms for each group. It turns out that a similar set of assumptions can be formulated even in the general case by referring to the average of the "symmetric" terms with $M = m$ and $M = N - 1 - m$. These two sets of terms are symmetric in the sense that terms with $M = m$ are obtained by adding m subgroups to the baseline population, while terms with $M = N - 1 - m$ are obtained by removing m subgroups from the total population. If some sort of linearity holds, we would expect that the average of the elements of these two sets does not depend on m so that we can write:

$$\frac{1}{2}\sum_{i=1}^{\binom{N-1}{m}} \frac{1}{\binom{N-1}{m}}(C^{n,m,i}) + \frac{1}{2}\sum_{i=1}^{\binom{N-1-m}{m}} \frac{1}{\binom{N-1-m}{m}}(C^{n,N-1-m,i}) = \frac{1}{2}(C^{n+} + C^{n-})$$

Notice that:

$$\binom{N-1}{m} = \frac{(N-1)!}{m!\,(N-1-m)!} = \binom{N-1}{N-1-m}$$

So, we can simplify the statement of our assumption as:

$$\sum_{i=1}^{\binom{N-1}{m}} \frac{1}{2\binom{N-1}{m}}(C^{n,m,i} + C^{n,N-1-m,i}) = \frac{1}{2}(C^{n+} + C^{n-})$$

In the expression above we can recognize a generalized version of the assumption we imposed for $N = 3$. Now, notice that in our actual expression for C^n we have terms that are very similar to the ones in the assumption equation:

$$\sum_{i=1}^{\binom{N-1}{m}} \frac{1}{N\binom{N-1}{m}} (C^{n,m,i} + C^{n,N-1-m,i}) = \frac{2}{N}\frac{1}{2}(C^{n+} + C^{n-})$$

Furthermore, because m ranges from 0 to $N-1$ but the two extremes correspond to C^{n+} and C^{n-} respectively, we have $\frac{N-2}{2}$ sums of the type $\sum_{i=1}^{\binom{N-1}{m}} \frac{1}{N\binom{N-1}{m}} (C^{n,m,i} + C^{n,N-1-m,i})$ in our expression for C^n ($\frac{N-2}{2} - \frac{1}{2}$ if N is odd). Thus, under our assumption we can write:

$$C^n = \frac{N-2}{2}\frac{2}{N}\frac{1}{2}(C^{n+} + C^{n-}) + \frac{1}{N}(C^{n+} + C^{n-})$$

$$= \frac{N-2}{2N}(C^{n+} + C^{n-}) + \frac{1}{N}(C^{n+} + C^{n-})$$

$$= \frac{N-2+2}{2N}(C^{n+} + C^{n-})$$

$$= \frac{1}{2}(C^{n+} + C^{n-})$$

Which proves that under our set of assumptions we can ignore all the internal terms and just compute our contributions as the average of the two external contributions. For completeness, we need to show that this is the case also for odd N. In this case, we have:

$$C^n = \left(\frac{N-2}{2} - \frac{1}{2}\right)\frac{2}{N}\frac{1}{2}(C^{n+} + C^{n-}) + \frac{1}{N}\frac{1}{2}(C^{n+} + C^{n-}) + \frac{1}{N}(C^{n+} + C^{n-})$$

$$= \left(\frac{N-2}{2}\right)\frac{2}{N}\frac{1}{2}(C^{n+} + C^{n-}) + \frac{1}{N}(C^{n+} + C^{n-})$$

$$= \frac{N-2}{2N}(C^{n+} + C^{n-}) + \frac{1}{N}(C^{n+} + C^{n-})$$

$$= \frac{N-2+2}{2N}(C^{n+} + C^{n-})$$

$$= \frac{1}{2}(C^{n+} + C^{n-})$$

Where the difference in the first line comes from the fact that we have $\left(\frac{N-2}{2} - \frac{1}{2}\right)$ symmetric sets plus a single "central" set which equals $\frac{1}{N}\frac{1}{2}(C^{n+} + C^{n-})$. With this final proof I have shown that under relatively mild assumptions (and assumptions one can verify), even in the case of N group we can compute each C^n contribution as the average of just the two external contributions, requiring the computation of just $4N$ life table functions.

Use Cases

The decomposition developed in this paper can be used to investigate the contribution of any set of mutually exclusive groups to the values of a life-table function for the total population. The basic decomposition for the $N = 1$ case was introduced by Hendi and Ho (2021) to study how the foreign-born population contributes to the national life expectancy in the US. However, the contributions subgroups of the foreign-born population, by origin or race/ethnicity for example, can be studied with the decomposition developed in this paper. Other interesting applications include the investigation of how different regions contribute to national life expectancy or how different countries contribute to the regional life expectancy.

Example Application: The Contribution of Counties along the Urban-Rural Classification to National Life Expectancy at Birth

I use death counts and population data from the National Center for Health Statistics obtained through CDC WONDER. Deaths and population are classified by year (2017-2019), sex, five-year age groups (<1,1-4,5-9,…,85+), and one of six urban-rural codes. The six codes schema classifies counties into four metropolitan categories and two non-metropolitan categories. The metropolitan categories are large central metro, large

fringe metro, medium metro, and small metro. The non-metropolitan categories are micropolitan and noncore. Details on the classification criteria are available in the NCHS documentation online (NCHS 2023).

As can be seen in Figure 1, at the national level, large fringe metro counties have the highest life expectancy for males, followed by large central metro counties, medium metro counties, small metro counties, micropolitan counties, and noncore counties. The ordering is essentially the same for females, the only difference being that large central metro counties have the highest life expectancy, with large fringe metro counties coming second. Figure 2 shows the group-specific contributions to national life-expectancy using the G-decomposition. As we would have expected, all groups but large fringe metro counties contribute negatively to the national life expectancy. Micropolitan counties have the largest absolute contribution followed by noncore counties, medium metro counties, small metro counties, and large fringe metro counties. The size of each contribution is determined by the interaction of three factors: the population size of counties belonging to the specific group, the population's age distribution, and its age-specific mortality rates (which life expectancy summarizes).

It turns out that for this example, the external-average approximation leads us to underestimate each contribution as can be seen in Figure 3. The difference is fairly small in absolute terms but nontrivial in relative ones. We can improve on our approximation by introducing an adjustment procedure. We know that the total contribution is:

$$e_0^T - e_0^B$$

And that the bias introduced by our approximation is:

$$Bias = (e_0^T - e_0^B) - \sum_{n=1}^{N} \frac{1}{2}(C^{n+} + C^{n-})$$

To eliminate the aggregate bias, we can thus add the adjustment factor $Bias/N$ to each approximate contribution $\frac{1}{2}(C^{n+} + C^{n-})$. This correction ensures that the sum of the approximate contributions equals the total difference to be decomposed. Figure 3 compares the "adjusted" approximate contributions, the "unadjusted" approximate ones, and the exact ones. We can see that the "adjusted" approximate contributions are barely distinguishable from the exact ones. I've been able to prove that in the case N=3, the adjusted approximate contributions are equal to the exact ones, but this only holds approximately in the general case.

As a further test of the validity of the G-decomposition, I adapted the three general decomposition methods (Andreev et al. 2002; Caswell 1989; Horiuchi et al. 2008)implemented in the `DemoDecomp` package (Riffe 2019) to obtain a decomposition by group similar to the one developed in this paper. The results of the comparison between the results obtained with the G-decomposition and those obtained with each of the three methods are presented in Figure 4 and Figure 5 (in which the stepwise decomposition is removed). They show that the G-decomposition produces result almost identical with the line-integral method and close to the ones of the life table response experiment. The stepwise decomposition fails to produce reasonable results.

Next Steps

I will show that the decomposition can be extended to produce age- and group-specific contributions. I will try to develop a theoretical justification for the adjustment procedure used in the Example Application section. To do so I will first investigate which conditions generate a substantial difference between the approximate and the exact contributions and I will then try obtaining a proof that the bias is approximately the same for all subgroups so that the adjustment factor $Bias/N$ is justified. I have achieved some preliminary success in both steps, but I still have not found a general expression for the bias introduced by

the approximation that links it with the adjustment factor I proposed. I will also discuss in more detail how the three general decomposition methods can be adapted to produce a decomposition by group.

References

Andreev, Evgueni M. 1982. “Metod Komponent v Analize Prodoljitelnosty Zjizni. [The Method of Components in the Analysis of Length of Life].” *Vestnik Statistiki* 9:42–47.

Andreev, Evgueni M., Vladimir M. Shkolnikov, and Alexander Z. Begun. 2002. “Algorithm for Decomposition of Differences between Aggregate Demographic Measures and Its Application to Life Expectancies, Healthy Life Expectancies, Parity-Progression Ratios and Total Fertility Rates.” *Demographic Research* 7:499–522.

Arriaga, Eduardo E. 1984. “Measuring and Explaining the Change in Life Expectancies.” *Demography* 21(1):83–96. doi: 10.2307/2061029.

Caswell, Hal. 1989. “Analysis of Life Table Response Experiments I. Decomposition of Effects on Population Growth Rate.” *Ecological Modelling* 46(3):221–37. doi: 10.1016/0304-3800(89)90019-7.

Das Gupta, Prithwis. 1978. “A General Method of Decomposing a Difference between Two Rates into Several Components.” *Demography* 15(1):99–112. doi: 10.2307/2060493.

Hendi, Arun S., and Jessica Y. Ho. 2021. “Immigration and Improvements in American Life Expectancy.” *SSM - Population Health* 15:100914. doi: 10.1016/j.ssmph.2021.100914.

Horiuchi, Shiro, John R. Wilmoth, and Scott D. Pletcher. 2008. “A Decomposition Method Based on a Model of Continuous Change.” *Demography* 45(4):785–801. doi: 10.1353/dem.0.0033.

Kitagawa, Evelyn M. 1964. “Standardized Comparisons in Population Research.” *Demography* 1(1):296–315. doi: 10.1007/BF03208469.

NCHS. 2023. “NCHS Urban Rural Classification Scheme for Counties.” Retrieved September 21, 2023 (https://www.cdc.gov/nchs/data_access/urban_rural.htm).

Pressat, Roland. 1985. “Contribution Des Écarts de Mortalité Par Âge à La Différence Des Vies Moyennes.” *Population (French Edition)* 40(4/5):766–70. doi: 10.2307/1532986.

Riffe, Tim. 2019. “DemoDecomp: General Demographic Decomposition Methods.”

Vaupel, James W., and Vladimir Canudas-Romo. 2002. “Decomposing Demographic Change into Direct vs. Compositional Components.” *Demographic Research* 7:1–14.

Figure and Tables

Figure 1: Life Expectancy at Birth by Year, Sex, and Urban-Rural Category

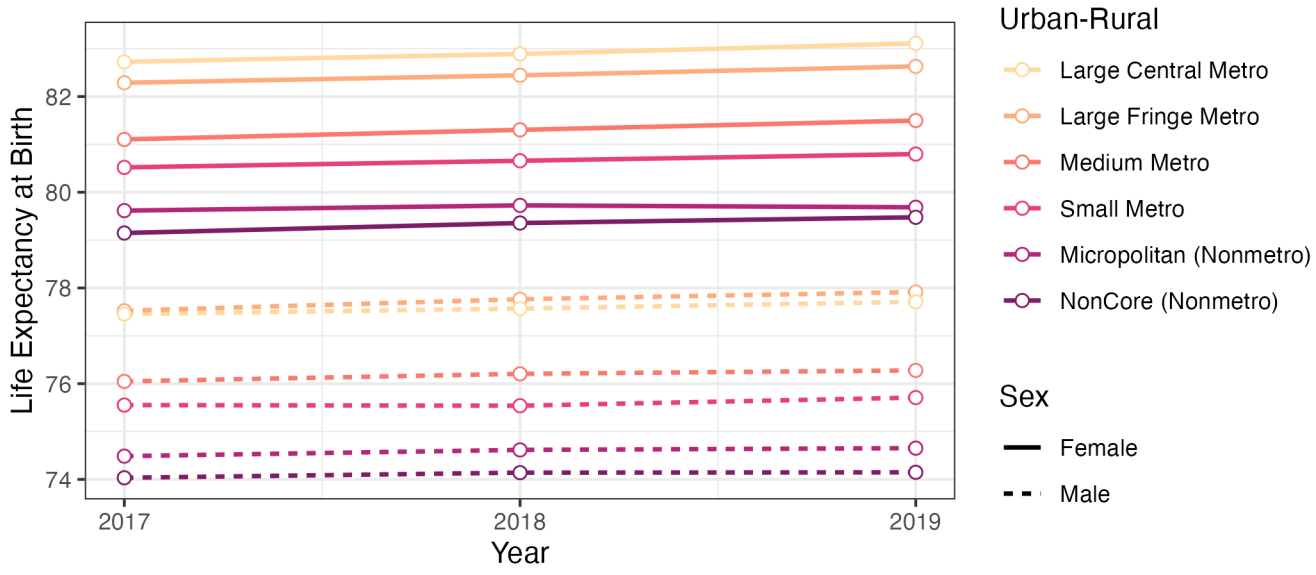

Figure 2: Decomposing the Contribution of Non-Large Central Metro Counties to National Life Expectancy by Urban-Rural Categories.

Figure 3: Comparing Decompositions Results Using the Unadjusted Approximated Contributions, the Adjusted Approximated Contributions, and the Exact Contributions

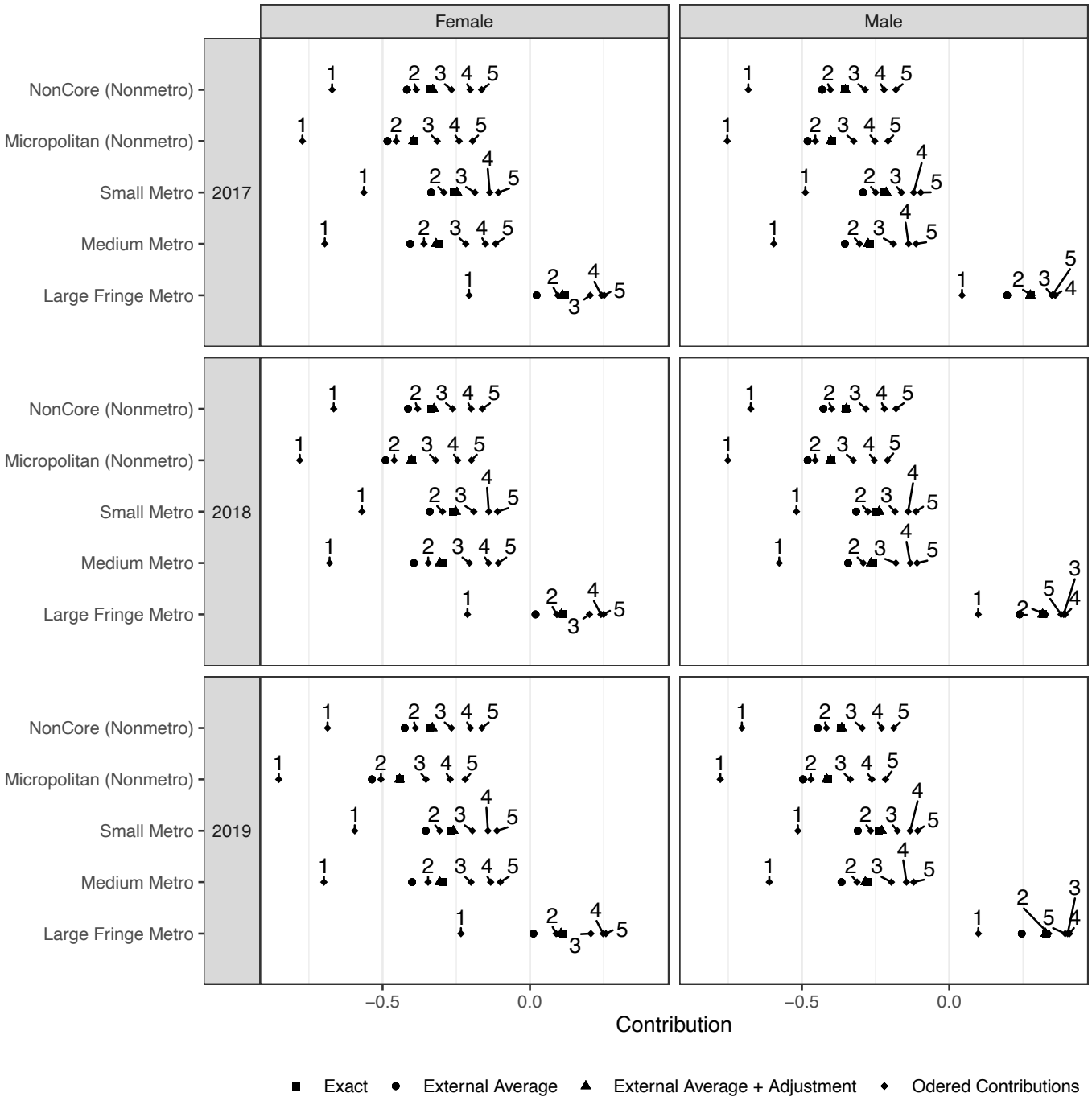

Figure 4: Comparing Decomposition Results from the G-Decomposition with Those Obtained with Other General Decomposition Methods: Horiuchi, LTRE, and Step-Wise.

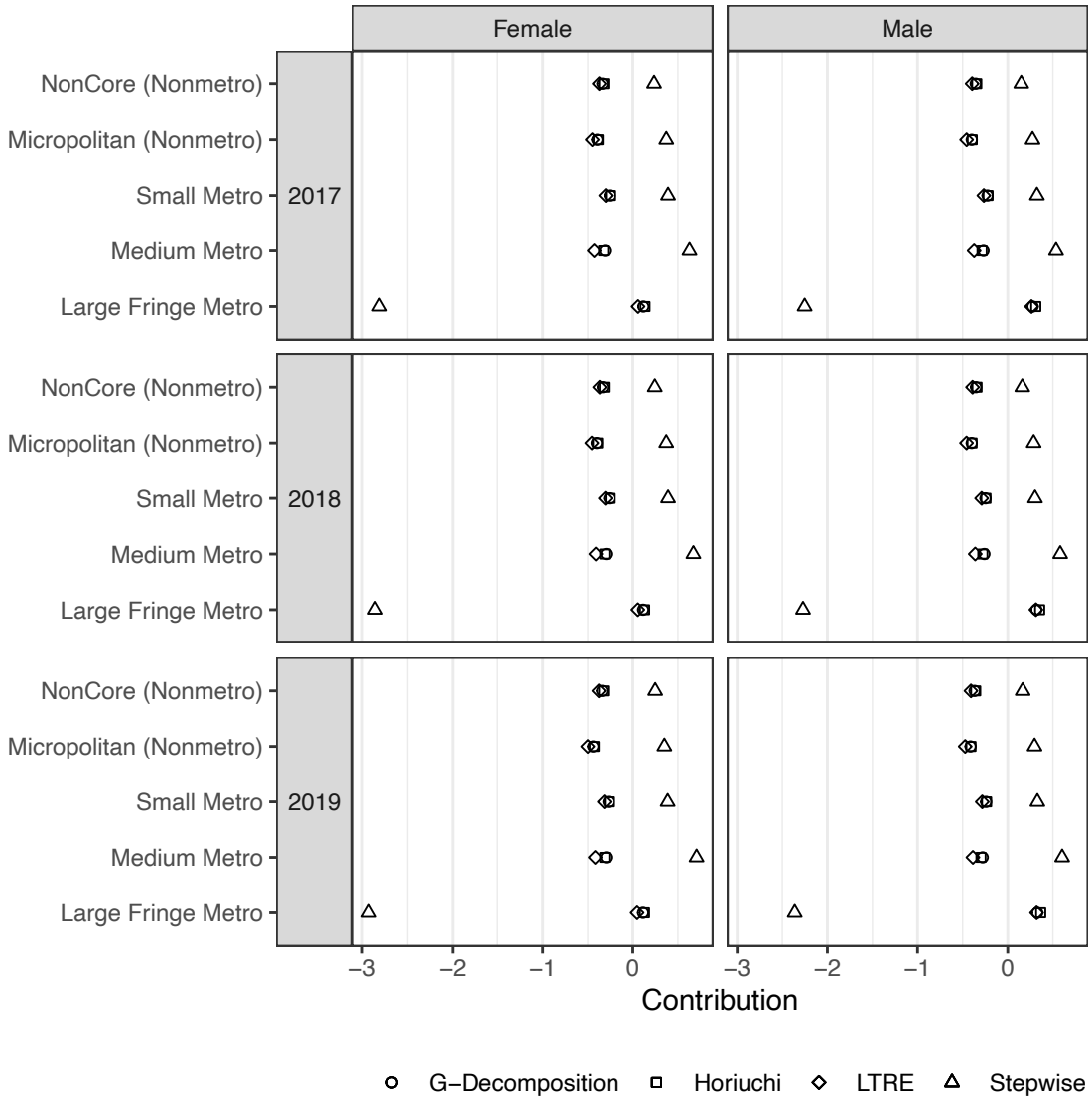

Figure 5: Comparing Decomposition Results from the G-Decomposition with Those Obtained with the Horiuchi and LTRE General Decomposition Methods.