Austrian Journal of Statistics
April 2014, Volume 43, 77–88.
http://www.ajs.or.at/

Family Generation Process from Administrative Data Sources and the Austrian Register-Based Census 2011

Henrik Rechta, Eliane Schwerer, Christoph Waldner
Statistics Austria, Vienna

Abstract

For the Austrian register-based census techniques to generate family statistics from administrative data sources were developed. The approach is based on data of relationships and how to handle them – in general, in a fixed household and in an imputation process. Therefore, we combined algebraic, graph theoretical and statistical tools to construct a general framework.

Keywords: register-based census, family statistics, household relationship matrix.

1. Introduction

Household and family statistics as part of the population census have been generated in Austria since 1900. These statistics include information about types of households and families, as well as the specific status of a household and family member respectively (e.g. husband, wife, etc.).

However, the register-based census act of 16th March 2006 (cf. Registerzählungsgesetz 2006, §7 (1)-(4)) stipulated that the population census should be conducted completely by using administrative data sources. Such register-based statistics have a long tradition in the Nordic countries (see United Nations 2007) and hold several advantages in comparison to classical surveys. For example, such a procedure is very cost efficient and there is no respondent burden anymore. On the other hand such a kind of census is a challenge in statistical methodology, i.e. data editing, coding variables, matching and derivating attributes and finally estimating missing values or objects. For a detailed description of the Austrian Census see Lenk (2009). A general description of the register system and the methodological work can be found in Wallgren and Wallgren (2007).

Obviously, creating household and family statistics from administrative registers only, leads to several challenges. In a traditional census each member of a household has to fill out a questionnaire, which includes queries about the household status. Using that information it is - more or less - easy to deduce the type of household. For example if there is no relationship in the household, it is definitely a non-family household. On the other hand, in a register-based census the lack of relationships is not a sufficient criterion any more. Furthermore, a household

with at least three persons can become implausible because of incorrect relationships (e.g. two partnership relations in a three person household). In summary, the challenge in household and family statistics is to detect implausible households and to estimate missing relations.

The purpose of this paper is to describe a framework for family generation via relationships developed by Statistics Austria. As preparation, we explain in Section 2 some basic definitions on households and families. After that we describe the available data, in particular the data about relationships. In Section 3, we look at the household level and show how plausibility is checked. Additionally, there is a description of an imputation process for relations. In Section 4, we finish with a short discussion on the quality assessment for the considered statistics and some closing remarks.

2. Preparations

2.1. Family nucleus

Since we focus on a register-based census, a household is defined by the household-dwelling concept (see United Nations 2006), i.e. we consider all persons living in a housing unit to be members of the same household. We are interested in family statistics, so we limit our analyses to private households, whereas institutional households are not included in this analysis.

Child refers to a blood, step- or adopted son or daughter (regardless of age or marital status) who has usual residence in the household of at least one of the parents, and who has no partner or own child(ren) in the same household. Foster children are not included.

A *family nucleus* is defined as two or more persons who live in the same household and whose relationship is defined as either married or cohabiting partners, or as a registered same-sex couple, or as parent and child.

For these definitions and further information on households and families we refer to United Nations (2006).

2.2. Data sources

To derive a family nucleus, information on households, demography and relationships are needed.

Households in the Austrian census are generated by linking the Central Population Register (CPR) with the buildings and dwellings register (BDR). These registers contain the same addresses (numerical codes) for buildings, but not always the same information on door numbers and therefore dwellings. The BDR is highly reliable on building level. As far as dwellings are concerned, the linking of dwellings with people registered in the CPR is less successful due to some missing or wrong door numbers. In these remaining cases (about 1.9% of the Austrian population) additional sources are used to generate households, e.g. relationships.

The demographical information we need are sex, age and marital status. Further, a variable *age at registration* is needed, which can be derived from the date of registration in the CPR and the date of birth.

The basic data sources for relationships are:

- Central social security register (CSSR)
- Child allowance register (CAR)
- Tax register (TR)

The variable relationship occurs in more than one register (it is a so-called multiple attribute). In the CSSR, people who are co-insured through a family member's national health insurance are included. The kind of co-insurance implies the type of relationship. The CAR contains information about the parent-child relations for children up to 18, or if they are students, up to 27 years of age. Under certain conditions (e.g. if you get child allowance) you can request

tax allowance by the federal ministry of finance. Parts of these records can be used to derive relationships.

2.3. Relationships

Statistics Austria uses the following types of relationships.

Cou	couple relation (married or cohabiting partners or registered same-sex couple),
P-C	parent-child relation,
Sib	sibling relation (full-, half-, step- and adoptive-siblings),
Gp-Gc	grandparent-grandchild relation,
0	no relation.

A relation from a person p_1 to a person p_2 is denoted by $p_1 \to p_2$. The opposite relation is denoted by $p_2 \to p_1$. The opposite of the directed relation *P-C* resp. *Gp-Gc* is denoted by *C-P* resp. *Gc-Gp*. There is no extra notation for the opposite of the undirected relations *Cou*, *Sib* and 0. The set of relations and their opposite relations is denoted by $\mathcal{R}$.

Remark. In the data sources CSSR and CAR there is a further type of relationship. The foster parent-foster child relation. This information can be used to ensure that a person can not be the child of such a household member (since they are in a foster parent-foster child relation).

Obviously, a valid relation requires two different persons and between those the relation has to be well-defined, i.e. exactly one type of relationship can be valid. Hence, one has to define rules for plausibility for each type of relationship. To illustrate this process of data preparation, it is briefly described here. Depending on the type of relation it must satisfy certain requirements on sex, age and marital status, respectively. As example, Statistics Austria use the following (Table 1) to check a *Cou* relation $p_1 \to p_2$ between two persons p_1, p_2 with age a_1, a_2 ($a_1 \geq a_2$), sex s_1, s_2 and marital status m_1, m_2, respectively.

Table 1: Rules regarding demography.

sex	age	$a_1 - a_2$	marital status
$s_1 \neq s_2$	$a_2 \geq 16$	arbitrary	m_1, m_2 are opposite-sex partner
$s_1 = s_2$	$a_2 \geq 18$	arbitrary	m_1, m_2 are same-sex partner

If the relationship does not comply with those rules, it will be deleted. At present, if a relation in a source is not consistent with other sources, it will be deleted. It is also thinkable to develop certain rules to keep one of these relations.

The final step to prepare relationships is to derive new ones with the help of existing ones. Here it is crucial that the relationships are archived (Statistics Austria has been collecting data on relationships since 2006). Let p_1, p_2, p_3 be pairwise distinct persons and assume that there are relations $p_1 \to p_2, p_2 \to p_3$ and assume further that there exists no relation $p_1 \to p_3$. This relation $p_1 \to p_3$ is then derived by composing $p_1 \to p_2$ and $p_2 \to p_3$ (see Table 2).

Table 2: Rules to derive new relations.

$p_1 \to p_2$	$p_2 \to p_3$	$p_1 \to p_3$
C-P	*Cou*	*C-P*
C-P	*P-C*	*Sib*
P-C	*P-C*	*Gp-Gc*

This can be important if the relations $p_1 \to p_2, p_2 \to p_3$ do not exist any more in a household of the current population census because the involved person p_2 is absent, but p_1, p_3 are still present in the same household and there exists no direct relation $p_1 \to p_3$. This derivation goes beyond the census population level.

That way, Statistics Austria obtains over 9 million relations for the register-based census 2011.

3. Household level

3.1. Households and graphs

Let $n \in \mathbb{N}$. A *(abstract) household* $H = (P, R)$ is a non-empty finite set $P = \{p_1, p_2, \ldots, p_n\}$ of persons together with a set of relations $R = \{(p_i \to p_j) \in \mathcal{R} |\ 1 \leq i < j \leq n\}$. Hence, a household has at most $\binom{n}{2}$ relations unequal to 0.

To a household H we assign a simple graph G_H by taking P as the set of vertices and $\{r \in R |\ r \neq 0\}$ as the set of edges. A relation $p_i \to p_j$ induces a direction and a label (the type of relationship) to the corresponding edge. Since the labels *Cou* and *Sib* do not change if we reverse the direction, we skip their direction in G_H. Hence G_H is a simple, (partially) directed, labeled graph.

Definition. A household H is called *(weak-) connected* if G_H is connected. A household H is called *strong-connected* if the subgraph of G_H with the set of edges $\{r \in R |\ r \in \{Cou, P\text{-}C, C\text{-}P\}\}$ is connected. In particular, a one person household $H = (\{p_1\}, \{\})$ is strong-connected.

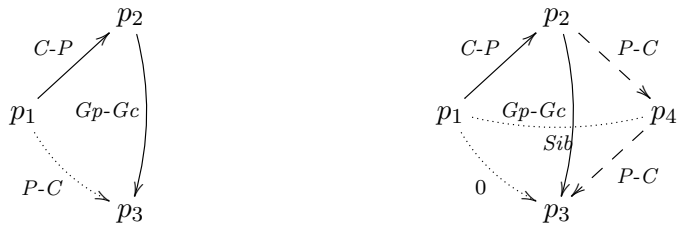
For the sake of completeness we define the *household relationship matrix* $A = (a_{ij})$ by

$$a_{ij} = \begin{cases} p_i \to p_j, & i < j, \\ 0, & i \geq j. \end{cases}$$

3.2. Algebraic structure of relations

We wish to define an operation $\circ$ on $\mathcal{R}$ by the natural way of composition. Unfortunately, composition is not a well-defined operation on $\mathcal{R}$, as the following example shows.

Example. Let p_1, p_2, p_3 be pairwise distinct persons in the same household and let $p_1 \to p_2 = C\text{-}P$, $p_2 \to p_3 = Gp\text{-}Gc$. Then there are two possibilities for the composition $C\text{-}P \circ Gp\text{-}Gc$. It seems natural that $p_1 \to p_3$ is a $P\text{-}C$ relation. Then p_1, p_2, p_3 are related in a direct-line as is shown in Figure 1a. But it is also possible that $p_1 \to p_3$ is a 0 relation. This makes sense if we assume that there is an unknown person p_4 who is in $P\text{-}C$ relation to p_3 and p_2 is in a $P\text{-}C$ relation to p_4. Then p_1 is the uncle/the aunt of p_3, i.e. a 0 relation (see Figure 1b).



(a) A direct-line household (b) A non-direct-line household

Figure 1: Some possible three-generations-households.

So in general, the operation $\circ$ is not well-defined which is caused by the fact that we involve relations between three generations.

However, there are at most two admissible values for a composition of two relationships. The first is always 0 and the second one depends on the composite relationships. Therefore, by defining the table of relationships operations (TRO) we label all of them by a. If there is one and only one admissible value, the composition will be unlabeled.

Table 3: Table of relationships operations (TRO).

$p_1 \to p_2$ \ $p_2 \to p_3$							
$\circ$	*Cou*	*P-C*	*C-P*	*Sib*	*Gp-Gc*	*Gc-Gp*	0
Cou	$\emptyset^c$	*P-C*	0	0	*Gp-Gc*	0	0
P-C	0	*Gp-Gc*	*Cou*^{b}	*P-C*	0	*C-P*^{a}	0
C-P	*C-P*	*Sib*	*Gc-Gp*	0	*P-C*^{a}	0	0
Sib	0	0	*C-P*	*Sib*	0	*Gc-Gp*	0
Gp-Gc	0	0	*P-C*^{a}	*Gp-Gc*	0	*Cou*^{a}	0
Gc-Gp	*Gc-Gp*	*C-P*^{a}	0	0	*Sib*^{a}	0	0
0	0	0	0	0	0	0	0

[a] There are two admissible values for the composition. The first is 0 and the second one is shown in the table.
[b] The persons p_1, p_3 have to fulfill conditions on sex, age and marital status, respectively like in Table 1.
[c] The symbol $\emptyset$ means that there is no admissible value.

The first column in TRO represents $p_1 \to p_2$ and the first row represents $p_2 \to p_3$. Additionally, $\emptyset$ means that there is no admissible value.

The *Cou* relation labeled by b in TRO requires special rules: We wish to calculate $p_1 \to p_3 = P\text{-}C \circ C\text{-}P$, which should be *Cou*. But in this case, relations alone are not able to guarantee the truth. More precisely, we have to check sex, age and marital status respectively like in Table 1. If these conditions are not fulfilled, then the whole household is called implausible (see Section 3.3).

The algebraic structure on $\mathcal{R}$ defined by composition in TRO is not associative as one can see by

$$(P\text{-}C \circ C\text{-}P) \circ Cou = Cou \circ Cou = \emptyset \neq Cou = P\text{-}C \circ C\text{-}P = P\text{-}C \circ (C\text{-}P \circ Cou).$$

3.3. Plausible households

Let $n > 2$, $H = (P, R)$ be a household and $r = p_1 \to p_2$, $s = p_2 \to p_3$, $t = p_1 \to p_3 \in R$, with pairwise distinct $p_1, p_2, p_3 \in P$. Assume that the following conditions on plausibility hold.

Plausibility conditions:

- $r \circ s \neq \emptyset$
- if $r \circ s \neq 0$, then $t \in \{0, r \circ s\}$
- if $r \circ s$ has label b, then Table 1 is satisfied

Then we can define a new operation $\diamond : R \times R \to R$ in the following way:

$$r \diamond s := \begin{cases} r \circ s, & \text{if } 0 \neq r \circ s \text{ has no label,} \\ t, & \text{if } r \circ s \text{ has label } a \text{ or if } r \circ s = 0, \\ r \circ s, & \text{if } 0 \neq r \circ s \text{ has label } b. \end{cases}$$

Now Statistics Austria uses the following **approach:**

1. Take pairwise distinct $p_1, p_2, p_3 \in P$. Check the plausibility conditions. If they are satisfied and $t \neq 0$, then do nothing. If they are satisfied and $t = 0$ and $r \diamond s \neq 0$, then replace t by $r \diamond s$. If they are not satisfied, stop and label H to be implausible. Do that for all pairwise distinct $p_1, p_2, p_3 \in P$.

2. Check whether a new relation $\neq 0$ in R has been derived by step 1. If no - stop. If yes - repeat step 1.

This approach overwrites the relations $0 \in R$ as long as new relations are derived and checks in addition if H is plausible.

Definition. A household H is called *plausible*, if $n \leq 2$ or H is not implausible by the approach. H is called *complete*, if no new relation $\neq 0$ can be derived.

Assume that H is plausible with size $|H| > 1$ and $p \in P$. The *household status of* p is ...

... *partner* if and only if there exists $q \in P$, such that $p \to q = Cou$.
... *child (not of lone parent)* if p is not a partner and there exists a partner $q \in P$, such that $q \to p = P\text{-}C$.
... *child (of lone parent)* if p is not a partner and there exists $q \in P$ who is not a partner, such that $q \to p = P\text{-}C$.
... *lone parent* if p is not a partner and there exists a child $q \in P$, such that $p \to q = P\text{-}C$.
... *not alone living* otherwise.

The following classification of private *household by type* is used in the Austrian Census 2011.

– Married couples without resident children
– Married couples with at least one resident child under 25
– Married couples, youngest resident son/daughter 25 or older
– Consensual union couples without resident children
– Consensual union couples with at least one resident child under 25
– Consensual union couples, youngest resident son/daughter 25 or older
– Lone father households with at least one resident child under 25
– Lone father households, youngest resident son/daughter 25 or older
– Lone mother households with at least one resident child under 25
– Lone mother households, youngest resident son/daughter 25 or older
– Two-or-more-family households
– One-person households
– Multi-person households (non-family)

Furthermore, the classification can be enlarged for one-family households, if one distinguishes such households with or without non-family members (see United Nations 2006).

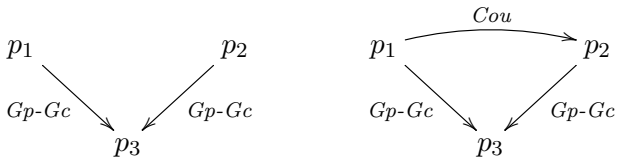
The household status implies the type of household and we get immediately the following sufficient criterion (this works well for the more detailed classification).

Criterion. *In a strong-connected, plausible household, the type of household is uniquely determined.*

Remarks. **1.** Note that the criterion above is not a necessary one; e.g. the type of $H = (\{p_1, p_2\}, \{r\}), r \neq 0$ is uniquely determined, but H is not strong-connected for $r \in \{Gp\text{-}Gc, Gc\text{-}Gp, Sib\}$. The condition – strong-connected – is strict for $n > 2$, as we can see in the following. Take $H_i = (\{p_1, p_2, p_3\}, \{p_1 \to p_2 = r_i, p_1 \to p_3 = Gp\text{-}Gc, p_2 \to p_3 = Gp\text{-}Gc\})$, $i = 1, 2$, $r_1 = 0$, $r_2 = Cou$. Then H_1, H_2 are connected, plausible households of different types. In this case it is possible to get H_2 (Figure 2b) from H_1 (Figure 2a) by estimating $\widehat{r_1} = Cou$, i.e. the type of household depends on the estimation process.

In practice, almost all connected households are strong-connected.

2. The type of household and the household status, respectively, are well-defined, which means that they are invariant under permutation of P, or equivalent, they just depend on the isomorphy class of G_H.



(a) Grandparents non-related (b) Grandparents related

Figure 2: Two connected, not strong-connected households.

If H is implausible, one has to redefine at least one $r \in R,\ r \neq 0$ to $r = 0$. It is not easy and sometimes impossible to determine which relations should be redefined to 0 (e.g. two partnership relations in a three person household). However, there are several ways to generate a plausible household from an implausible one. Some of these are:

- Redefine all relations in R to 0.
- If G_H is not connected, check (using the plausibility conditions) which connected component of G_H is implausible and redefine all relations in that component to 0.
- Check if the household becomes plausible by redefining just one certain relation and if there is no other relation with this property.

The occurrence of an implausible household is not very likely, i.e. in the Austrian register-based census 2011, only about 0.05% of all private households with three or more persons are implausible.

3.4. Estimation of Relationships

From now on we assume that H is a complete plausible, not strong-connected household. As we have seen, in such a household we are no longer able to guarantee the type of household. Hence, we have to impute relations in H such that the estimated household $\widehat{H}$ stays plausible. Our imputation method is a combination of a hot-deck technique based on demographic characteristics together with an ordering relation based on normalized frequencies and some static rules involving date of registration and external household relations. Since we are interested in strong-connection we just estimate relations of types $Cou, P\text{-}C$. Note that the imputation of $P\text{-}C$ types includes the estimation of $C\text{-}P$ types via permutation of the persons concerned. Before we handle the general case, let us consider the special case of a two persons household. An overview about the general data work flow of the Austrian census with special focus on the imputation process can be found in Kausl (2012).

Let $n = 2$, $H = (\{p_1, p_2\}, \{0\})$ and a_1, a_2 the age, $a_\Delta = a_1 - a_2$, s_1, s_2 the sex, m_1, m_2 the marital status of p_1, p_2. The variable *parents* indicates whether or not a person has at least a mother or a father in a separate household. For the variable *ages at registration* the date of the later registration of both persons was determined, then the minimum age of those persons at this date was computed. Relations are highly correlated with the *age-difference* a_Δ and with sex. To compute a probability distribution, depending on these variables and the relation type, we need a kind of non-relation to the relation type in question. Such non-relations imply the complementary probability.

***Cou*-distribution.** Let $s \in \{\text{male, female}\}$ and $a_\Delta \in \mathbb{Z}$ arbitrary but fixed. To compute the complementary probability we define non-*Cou* relations as follows: Take a three-person households H with persons $\{p_1, p_2, p_3\}$, such that $p_1 \to p_2 = Cou, p_2 \to p_3 = 0$ and the sub-household $\{p_2, p_3\}$ admits a *Cou* relation by demographical rules and $a_\Delta = a_2 - a_3, s_2 = s$. Since p_2 is already a partner, p_3 could not be a new one. Hence, $p_2 \to p_3$ forms a *non-Cou relation*. Restrict the relations of types $\{P\text{-}C, Gp\text{-}Gc, Sib\}$ in the stock of households to those which could be a *Cou* relation by demographical rules and which start with sex s and have age-

difference a_Δ. This set together with the non-*Cou* relations forms the set of complementary events. Comparing these events with the real *Cou* relations in the stock of households (who start with sex s and have age-difference a_Δ) leads to the probability distribution $d(Cou, s, a_\Delta)$ of *Cou*.

***P-C*-distribution.** Again, let $s \in \{\text{male, female}\}$ and $a_\Delta \in \mathbb{Z}$ arbitrary but fixed. Like the *Cou*-distribution, we compute the complementary probability by non-*P-C* relations defined as follows: Take a households H with persons $P = \{p_1, p_2, \dots\}$, such that $p_1 \to p_2 = 0$, the sub-household $\{p_1, p_2\}$ admits a *P-C* relation by demographical rules and $a_\Delta = a_1 - a_2, s_1 = s$. Further assume that there exists a person $q, q \notin P$ with sex s and relation $q \to p_2 = P\text{-}C$ (i.e. p_2 has at least a parent with sex s in a separate household). Then $p_1 \to p_2$ forms a *non-P-C relation*. Restrict the relations of types $\{Cou, Gp\text{-}Gc, Sib\}$ in the stock of households to those which could be a *P-C* relation by demographical rules and which start with sex s and have age-difference a_Δ. This set together with the non-*P-C* relations forms the set of complementary events. Comparing these events with the real *P-C* relations in the stock of households leads to the probability distribution $d(P\text{-}C, s, a_\Delta)$ of *P-C*.

Note that the result obtained that way heavily depends on the non-relations. Before one can use the distribution, one has to ensure that there are enough such non-relations. Perhaps one has to shrink the set of (complementary) events, e.g. restriction to households with $n \leq 3$.

Further rules. An imputed relation $p_1 \to p_2$ between two persons p_1, p_2 with sex s_1, s_2, respectively has to fulfill rules like those presented in Table 1 (perhaps some of them with enlarged restrictions). Further rules for an imputed relation (see Table 4) include the variables *parents* and *ages at registration*.

Table 4: Further rules for imputed relations.

$p_1 \to p_2$	parents	ages at registration
Cou	arbitrary	≥ 16 (opposite sex), ≥ 18 (same sex)
P-C	no parent with sex s_1 is known for p_2	arbitrary

Ordering relation. If a household allows to impute a relation for which more than one type ($\neq 0$) is possible or the household allows to impute two or more relations – which should be tried to be estimated first? The answer is crucial (in particular if $n > 2$), since an imputed relation affects the subsequent estimation procedure. Hence, we try to order the possible relations and types among themselves according to their probability, starting from the most probable. The easiest way to do that is to count the frequencies. Since we want to compare different types of relations and the number of *P-C* relations recorded by Statistics Austria is more then twice the number of *Cou* relations, we must normalize them. We take the whole stock of historicised relations as described in Section 2.3. Let $N(Cou, s, a)$ be the number of all relations $p_1 \to p_2 = Cou, s_1 = s, a = a_\Delta$ plus the number of all relations $p_1 \to p_2 = Cou, s_2 = s, a = -a_\Delta$ for arbitrary but fixed $s \in \{\text{male, female}\}, a \in \mathbb{Z}$. Further let $N(P\text{-}C, s_1, a_\Delta)$ be the number of all relations $p_1 \to p_2 = P\text{-}C$. Then the relative frequency distribution $f(r, s, a)$ of the relation $r \in \{Cou, P\text{-}C\}$ with sex $s_1 = s \in \{\text{male, female}\}$ and age-difference $a_\Delta = a$ is defined as

$$f(r, s, a) = \frac{N(r, s, a)}{\sum_a N(r, s, a)}. \tag{1}$$

Now assume that $p_1 \to p_2 = \widehat{r}$ can be either of type $r_1 \neq 0$ or $r_2 \neq 0$, then we define the preordering relation $r_1 \succeq r_2$ if and only if $f(r_1, s, a) \geq f(r_2, s, a)$.

a_Δ	rel. frequency in % *Cou*		*P-C*
17	0.31	$>$	0.30
18	0.25	$<$	0.46
19	0.20	$<$	0.78
20	0.16	$<$	1.27
$\vdots$		$\vdots$	

Example. Let $s_1 =$ male, $s_2 =$ female and $a_1 \geq 37,\ a_2 = 20$ and $m_1, m_2 =$ never married.

Hence, if $a_\Delta = 17$, we try to estimate a *Cou* relation first, whereas if $a_\Delta \geq 18$ we try to estimate a *P-C* relation first.

To estimate a relation $p_1 \to p_2 = \widehat{r}$ in a household $H = (\{p_1, p_2\}, \{0\})$ a uniformly distributed random variable x between 0 and 1 is produced and assigned to (p_1, p_2). If the type in question is r_1 and $x \leq d(r_1, s, a_\Delta)$ then we accept r_1. If not, we try whether we can estimate another type $r_2,\ r_2 \preceq r_1$ for the relation $\widehat{r}$.

Now we are able to enlarge this procedure to $n \geq 3$ by estimating relations stepwise.

Decomposition of H. Let $H = (P, R)$ be a fixed plausible household with $n > 3$ persons. Then H is a disjoint union of the strong-connected components $C_1, \ldots, C_m$ of H (i.e. the maximal strong-connected sub-households of H), $H = \bigsqcup_i C_i$. Let n_i be the number of persons in $C_i,\ i = 1, \ldots, m$. Obviously, there are $N = \sum_{1 \leq i < j \leq m} n_i n_j$ possibilities for choosing two persons $p_1, p_2 \in P$, who are not in the same component (by ignoring the order of the components). These possibilities define a set of pairs $\{(p_{i_1}, p_{i_2}) |\ i = 1 \ldots, N\}$. Each of these pairs (p_{i_1}, p_{i_2}) can be viewed as a two-person sub-household of H to which we can apply the procedure above (an example is shown in Figure 3). A further reduction of N is possible if we delete all pairs (p_{i_1}, p_{i_2}) which already define a relation of type *Gp-Gc*, *Gc-Gp* or *Sib*.

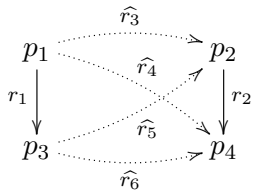


Figure 3: Possible relations in a certain household.

Each relation $\widehat{r_i}, i = 1, \ldots, N$ has at most two possible types of relation $\neq 0$. Let U be the (possibly empty) set of all combinations $(\widehat{r}, u)$, $\widehat{r}$ a possible relation, $u \neq 0$ a possible type of $\widehat{r}$. Let s_i, s_j be the sex of the first involved person of $\widehat{r_i}, \widehat{r_j}$ respectively and $a_{\Delta i}$, $a_{\Delta j}$ respectively, the age-difference of the corresponding pair. Then $(U, \succeq)$ is a total preordered set, defined by $(\widehat{r_i}, u) \succeq (\widehat{r_j}, u')$ if and only if $f(u, s_i, a_{\Delta i}) \geq f(u', s_j, a_{\Delta j})$.

Now we are ready to carry out the following procedure:

1. For a plausible household H compute the set U.
2. If $U = \{\}$ then stop, else take a (not necessary unique) maximal element $(\widehat{r}, u) \in U$ and generate a uniformly distributed random variable x between 0 and 1.
   (a) If $x \leq d(u, s, a_\Delta)$, assign the type u to $\widehat{r}$. This assignment generates a new household $\widehat{H}$.
      i. If $\widehat{H}$ is plausible, then replace H with $\widehat{H}$ and repeat step 1.
      ii. If $\widehat{H}$ is implausible, then remove $\widehat{r}$ with type u from $\widehat{H}$, delete $(\widehat{r}, u)$ from U and repeat step 2.

(b) If $x \not\leq d(u, s, a_\Delta)$, then delete $(\widehat{r}, u)$ from U and repeat step 2.

The procedure terminates if $U = \{\}$, i.e. if no new relation can be imputed. The maximal number of imputed relations is the number m of strong-connected components of H.

Remark. It is possible that in huge households (e.g. $n > 10$) this procedure generates many households of the type *Two-or-more-family household*. To prevent this event, one can introduce further restrictions (e.g. limit the number of families in H by certain rules).

4. Quality assessment

It is of general interest to asses the quality of statistics based on administrative sources. This is a broad field which reaches from the quality of the data sources to the outcome statistics (see Berka, Humer, Lenk, Moser, Rechta, and Schwerer 2010, for a structural quality framework on this topic). Here we wish to measure the quality of the produced statistics for families by the following measure.

Quality measure. Let $H = (P, R)$ be a plausible household. The imputation process described in Section 3.4 finally leads to $H \rightsquigarrow \widehat{H} = (P, \widehat{R})$.

We define $q : \widehat{R} \to Q$ by

$$q(r) = \begin{cases} 1, & r \text{ is not imputed,} \\ 0, & r \text{ is imputed.} \end{cases}$$

Here $Q = (\{0, 1\}, \cdot)$ is a Boolean algebra. The set $\widehat{R}$ generates by $\diamond$ a complete household $\widetilde{H} = (P, \widetilde{R})$ on which we enlarge q by

$$q(r \diamond s) := \begin{cases} q(r) \cdot q(s), & \text{if } 0 \neq r \circ s \text{ has no label or label } b, \\ q(t), & \text{if } r \circ s \text{ has label } a \text{ or if } r \circ s = 0, \end{cases}$$

where $r = p_1 \to p_2,\ s = p_2 \to p_3,\ t = p_1 \to p_3 \in \widehat{R}$, with pairwise distinct $p_1, p_2, p_3 \in P$.

Assume that $F = (P_F, R_F)$ is a family of $\widetilde{H}$ (i.e. a sub-household of $\widetilde{H}$ which forms a family). Then we can compute a quality measure $\mu(F) \in [0, 1]$ of F as

$$\mu(F) := \frac{1}{|R_F|} \left(\sum_{r \in R_F} q(r) \right).$$

So μ measures how reliable a family actually is. Counting all families F by $\mu(F)$, assesses the quality of the family statistics.

In the Austrian register-based census 2011 there are 2,306,650 families. About 80% of them have $\mu(F) = 1$, about 6% have $\mu(F) \in (0, 1)$ and the remaining 14% have $\mu(F) = 0$.

Classification rate. A correct classification rate shows further details of the accuracy of an imputed model. Usually the input data is split into a training and testing sample. Then the testing sample can be used to compute the classification rate. Unfortunately, this approach is very difficult in the family topic. Since we have no observed 0 relations in certain cases (e.g. s_1 = male, s_2 = female, p_1, p_2 adult), we cannot involve them in the training sample as well as in a testing sample.

However, instead of this approach we compute several correct classification rates by using the results of the last traditional population census 2001. There are ensured 0 relations. We take the demographical attributes, use the developed rules and compare the outcome with the result of the census 2001. For example, the type of household coincides in about 92% of all private households. If we restrict the private households to those with two or more persons

Table 5: Classification rates by household types.

Type of household[1]	correct classification rates
Married couples without resident children	96%
Married couples with at least one resident child under 25	92%
Married couples, youngest resident son/daughter 25 or older	87%
Consensual union couples without resident children	85%
Consensual union couples with at least one resident child under 25	96%
Consensual union couples, youngest resident son/daughter 25 or older	78%
Lone father households with at least one resident child under 25	70%
Lone father households, youngest resident son/daughter 25 or older	54%
Lone mother households with at least one resident child under 25	92%
Lone mother households, youngest resident son/daughter 25 or older	90%
Multi-person households (Non-family)	57%

[1] The table does not include the type *one-person households* since in such households there is no estimated relation, but all of them must be correct classified. Further the type *two-or-more-family households* is omitted for the following reason: to prevent an overestimation of such households, we limited the number of families generated by estimated relations by one. So a classification rate cannot be computed for this household type.

(since a one-person household obviously must be classified correctly), about 89% still match. The classification rates for household types are listed in Table 5.

5. Closing remarks

As we have seen in Section 4, most of the families are verified by administrative data sources. Of course, it is unrealistic to have the measure μ equal 1 for all families, but one can try to improve the present quota. Obviously, this will happen in the future since the relations are archived. One can speed up this event by involving other data sources - for example, relations derived from a new central civil status register, which will be established at the ministry of interior in the year 2014. Other issues arise in some optional topics, such as foster children, stepchildren, reconstituted families etc. However, in a register-based census these topics are not easy to handle.

References

Berka C, Humer S, Lenk M, Moser M, Rechta H, Schwerer E (2010). "A Quality Framework for Statistics Based on Administrative Data Sources Using the Example of the Austrian Census 2011." *Austrian Journal of Statistics*, **39 (4)**, 299–308.

Kausl A (2012). "The Data Imputation Process of the Austria Register-Based Census." UN-ECE, Conference of European Statisticans, Work Session on Statistical Data Editing.

Lenk M (2009). "Methods of the Register-based Census in Austria." Seminar on Innovations in Official Statistics United Nations, New York.

Registerzählungsgesetz (2006). "Durchführung von Volks-, Arbeitsstätten-, Gebäude- und Wohnungszählungen und Änderung des Bundesgesetzes, mit dem das Postgesetz 1997, das Meldegesetz 1991 und das Bildungsdokumentationsgesetz geändert werden." BGBl. I, Nr. 33/2006. Vienna.

United Nations (2006). *Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing.* United Nations Publication, New York and Geneva.

United Nations (2007). *Register-Based Statistics in the Nordic Countries: Review of Best Practices with Focus on Population and Social Statistics*. United Nations Publication, New York and Geneva.

Wallgren A, Wallgren B (2007). *Register-Based Statistics: Administrative Data for Statistical Purposes*. John Wiley & Sons, Ltd.

Affiliation:

Henrik Rechta, Eliane Schwerer, Christoph Waldner
Registers, Classification and Methods Division
Statistics Austria
Guglgasse 13
A-1110 Vienna, Austria
E-mail: Henrik.Rechta@statistik.gv.at
E-mail: Eliane.Schwerer@statistik.gv.at
E-mail: Christoph.Waldner@statistik.gv.at
URL: http://www.statistik.at