Austrian Journal of Statistics
July 2023, Volume 52, 163–179.
http://www.ajs.or.at/
doi:10.17713/ajs.v52i4.1510

The Impact of COVID-19 on Relative Changes in Aggregated Mobility Using Mobile-phone Data

Georg Heiler
E-Commerce, TU Wien

Allan Hanbury
E-Commerce, TU Wien

Peter Filzmoser
CSTAT, TU Wien

Abstract

Mobile-phone data can be used to investigate the mobility of a big part of a population in a given period. Here we have analyzed this information for Austria in the first half year of the COVID-19 pandemic. Especially the period around the first lockdown was of interest, and our focus is on exploring possible differences between age groups and among females and males. The data is once treated from an absolute point of view, by analyzing the numbers as they are reported and from a relative point of view, with the help of compositional data analysis tools. Our goal is to compare analyses of the absolute values and of relative information, in order to reveal possible differences in the groups formed by age and gender. It turns out that both types of analyses provide different and partially complementary insights. This is also underlined when analyzing data from call durations, or subdata just for specific Austrian districts.

Keywords: compositional-data-analysis, mobility, pandemic, big-data, geospatial-data.

1. Introduction

The usage data of mobile-phones is used in a variety of different areas, such as during the COVID-19 pandemic (Pepe, Bajardi, Gauvin, Privitera, Lake, Cattuto, and Tizzoni 2020; Jia, Lu, Yuan, Xu, Christakis, Jia, and Nicholas 2020; Gao, Rao, Kang, Liang, Kruse, Doepfer, Sethi, Reyes, Patz, and Yandell 2020; Jeffrey, Walters, Ainslie, Eales, Ciavarella, Bhatia, Hayes, Baguelin, Boonyasiri, Brazeau *et al.* 2020; Yabe, Tsubouchi, Fujiwara, Wada, Sekimoto, and Ukkusuri 2020; Vollmer, Mishra, Juliette *et al.* 2020; Xu, Gutierrez, Mekaru, Sewalk, Goodwin, Loskill, Cohn, Hswen, Hill, Cobo, Zarebski, Li, Wu, Hulland, Morgan, Wang, O'Brien, Scarpino, Brownstein, Pybus, Pigott, and Kraemer 2020; Santamaria Serna Carlos, Sermi, Spyratos, Iacus, Annunziato, Tarchi, and Vespe 2020; Iacus Stefano, Serna, Sermi, Spyratos, Tarchi, and Vespe 2020a,b; Heuzroth 2020), customer segmentation (Aheleroff 2011), identification of personality traits and lifestyle (Chittaranjan, Jan, and Gatica-Perez 2011; Hillebrand, Khan, Peleja, and Oliver 2020), the analysis of large social networks (Aksu, Korpeoglu, and Ulusoy 2019; Al-Molhem, Rahal, and Dakkak 2019; Aledavood, Lehmann, and Saramäki 2018), hotspot detection (Nika, Ismail, Zhao, Gaito, Rossi, and Zheng 2016), prediction of movement (Dao, Le, and Yoon 2019), mode of transport identification (Zhao, Bucher, Martin, and Raubal 2020), credit scoring (Liu, Ma, Zhao, and Zou 2018), disaster

recovery (Andrade, Layedra, Vaca, and Cruz 2019; Marzuoli and Liu 2019), analysis of sleeping behavior of the population (Monsivais, Bhattacharya, Ghosh, Dunbar, and Kaski 2017), migration (Isaacman, Frias-Martinez, and Frias-Martinez 2018) and land usage classification (Shi, Lv, Seng, Xing, and Chen 2019; Lenormand, Picornell, Cantú-Ros, Louail, Herranz, Barthelemy, Frías-Martínez, Miguel, and Ramasco 2015).

The location of a mobile-phone is known for the Mobile Network Operator (MNO) of the Global System for Mobile Communication (GSM) network. We have partnered with an MNO in Austria to access such anonymized data. We have defined an aggregation method to understand the overall mobility of the whole population. Our data set, the aggregation, anonymization approach and the various phases of the lockdown in the first half year of the pandemic are outlined in detail in (Heiler, Reisch, Hurt, Forghani, Omani, Hanbury, and Karimipour 2020).

With the outbreak of COVID-19 and the subsequent lockdown in Austria, the mobility behavior of the population has changed significantly. This is reflected in the mobility data derived from mobile-phone information (Heiler *et al.* 2020). An appropriate measure needs to be established to measure mobility, which reflects the mobility. One possibility is the Radius of Gyration (ROG) (Gooch 2011). It is formally defined below, and refers to the time-weighted distance of the movement locations to the primary location. We compute it on a daily level. Its unit is meters and the values are strictly positive. In this work, we analyze the aggregated (median) ROG of the whole population of Austria for various groups as a time series. The groups are defined by gender- or age groups.

Traditionally, a comparison is made in terms of *absolute information*, i.e., the ROG time series values of the different groups are analyzed in their unit of meters. We have conducted such an analysis (Reisch, Heiler, Hurt, Klimek, Hanbury, and Thurner 2021) which focuses on gender differences.

An alternative is to compare *relative information*, for example the ROG of the males with respect to females, or in terms of the ratio males to females. This leads to a dimensionless time series, and to a different aspect of data analysis which emphasizes the differences between the individual groups. A joint increase or decrease in both groups may not lead to a big change of the ratio. On the other hand, the ratio will change if the values of one group increase, and at the same time they decrease in the other group, or vice versa. Here again, the relative change rather than the absolute change is important. For example, if the ROG changes from 1000m to 2000m in one group, and from 2000m to 1000m in the other group, the ratio would change from 1/2 to 2. The same change could be observed if the absolute values in both groups would be bigger by a factor of 10. Thus, absolute values are no longer relevant in this consideration, because a multiplication by any positive constant leads to the same ratio. This is still trivial in case of comparing two groups, but it is no longer straightforward when relative information of several groups, such as age classes, should be compared. *Compositional data analysis* is devoted to this problem of analyzing relative information (Aitchison 1986; Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015; Filzmoser, Hron, and Templ 2018). In fact, compositional data analysis is frequently used in geosciences, but also more and more in other fields such as biology (Espinoza, Shah, Singh, Nelson, and Dupont 2020), bioinformatics (Quinn, Erb, Richardson, and Crowley 2018), economics (Trinh, Morais, Thomas-Agnan, and Simioni 2019), marketing (Joueid and Coenders 2018), medicine (Dumuid, Pedišić, Palarea-Albaladejo, Martín-Fernández, Hron, and Olds 2020), in applications with spatially dependent data (Thomas-Agnan, Laurent, Ruiz-Gazen, Nguyen, Chakir, and Lungarska 2021), etc.

We analyze the movement data of the first half year of the COVID-19 pandemic. In this period, the lockdown starting on March 2020 had more substantial effects on mobility than subsequent lockdowns. Our goal is to compare analyses of the absolute values and of relative information, in order to reveal possible differences in the groups formed by age and gender. Potentially, groups at risk could be detected and special interventions placed to help these ensure their health and safety.

This work is structured as follows. In Section 2.1 we give a brief mathematical introduction to compositional data analysis. Section 2.2 provides more details about the mobile-phone data used and about the quantities derived. In addition to the mobility measured as the ROG, we will investigate the call duration per day, again aggregated by the median. Section 3 presents comparisons of the analysis based on absolute and on relative information, and the final Section 4 summarizes the findings.

2. Materials and methods

2.1. Compositional data analysis

From the point of view of compositional data analysis, a composition is defined as multivariate vector, consisting of strictly positive values, where the absolute numbers as such are not of interest, and only relative information is relevant for the analysis (Filzmoser *et al.* 2018). The median ROG values of different age categories for a particular day, and every age category can be considered as a composition. We use the notation $x_1, \ldots, x_D$ for the compositional parts of D categories, and the composition is written as the (column) vector $\mathbf{x} = (x_1, \ldots, x_D)'$. For every day in the data we will observe such a composition, which in fact leads to a multivariate compositional time series. The interest is in relative information in terms of the ratios, and thus all pairs x_j/x_k, for $j, k = 1, \ldots, D$, should be considered in the analysis. Obviously, the pairs for $j = k$ are not relevant, and pairs of the reverse ratio x_k/x_j do not contain potentially new information. This motivates to consider the logarithm of the ratios, $\ln(x_j/x_k)$, so-called log-ratios. The reverse ratios have a different sign, and thus do not need to be considered, and their variance is the same as for the original ratio. Moreover, log-ratios tend to be more symmetric than simple ratios without a logarithm (Pawlowsky-Glahn *et al.* 2015).

Still, the resulting $D(D-1)$ pairs $\ln(x_j/x_k)$, for $k > j$, can be represented by only $\leq D-1$ components (Filzmoser *et al.* 2018), and this motivates to aggregate this information. Consider an aggregation

$$y_1 = \frac{1}{D}\left(\ln\frac{x_1}{x_2} + \ldots + \ln\frac{x_1}{x_D}\right) = \ln\frac{x_1}{g(\mathbf{x})}\,, \tag{1}$$

where

$$g(\mathbf{x}) = \sqrt[D]{\prod_{j=1}^{D} x_j}$$

is the geometric mean of the composition $\mathbf{x}$. Then, y_1 represents all relative information about the part x_1 to the other parts in the composition in a form of an average of the log-ratios. This leads to the definition of so-called Centered Log Ratio (CLR) coefficients Aitchison (1986)

$$\mathbf{y} = (y_1, \ldots, y_D)' \quad \text{with} \quad y_j = \ln\frac{x_j}{g(\mathbf{x})}\,. \tag{2}$$

The vector $\mathbf{y}$ contains all relative information about $\mathbf{x}$. It consists of D components y_j which are associated with the relative information about the corresponding part x_j. However, it turns out that $y_1 + \ldots + y_D = 0$, and thus a representation of data in terms of CLR coefficients leads to singularity (Filzmoser *et al.* 2018). Although there are ways to circumvent this issue (Filzmoser *et al.* 2018), we will proceed with CLR coefficients for the following analysis for simplicity.

Consider now a multivariate compositional time series $\mathbf{x}_t = (x_{t1}, \ldots, x_{tD})'$, for the time points $t = 1, \ldots, T$, and the observations x_{tj} for each part $j \in \{1, \ldots, D\}$. The time series expressed in CLR coefficients is $\mathbf{y}_t = (y_{t1}, \ldots, y_{tD})'$, with $y_{tj} = \ln(x_{tj}/g(\mathbf{x}_t))$, with the geometric mean $g(\mathbf{x}_t) = (\prod_{j=1}^{D} x_{tj})^{1/D}$ per time point. Since this data representation only reflects relative information of the time series, an additional visualization of the absolute time series values can be interesting to get a more complete picture.

The CLR coefficients result in multivariate data that can be analyzed with the traditional multivariate statistical methods (Filzmoser *et al.* 2018). A prominent way to represent the information in a lower-dimensional space is to use Principal Component Analysis (PCA). Since PCA is sensitive to data outliers or inhomogeneous data, robust versions have been proposed, also in the compositional data analysis framework (Filzmoser *et al.* 2018). The resulting loadings and scores are commonly represented in a biplot to get an overview of the multivariate data (Aitchison and Greenacre 2002).

2.2. Mobile-phone data

In this work, we analyze two measures obtained from the mobile phone data, the call duration and the radius of gyration ROG. While the meaning of the former is straightforward, the latter needs to be defined.

Consider an individual $i := i(t)$ at a certain day $t \in \{1, \ldots, T\}$. For data privacy reasons, the individual's index will change every day. The data is made available to the researchers already anonymized with a daily changing key.

Furthermore, the current location of the individual's mobile phone is available at the time points $t_\tau = t + \tau_t$, for a number of time points T_t per day, where $\tau_t \in [0, 1)$. The corresponding x- and y-coordinates are denoted by $(\xi_{it_\tau}, \eta_{it_\tau})$. With this information, the stay duration l_{it_τ} for individual i at time point t_τ can be computed, which is used to calculate a weighted average $(\bar{\xi}_{it}, \bar{\eta}_{it}) = \left(\frac{\sum_{t_\tau} l_{it_\tau} \xi_{it_\tau}}{\sum_{t_\tau} l_{it_\tau}}, \frac{\sum_{t_\tau} l_{it_\tau} \eta_{it_\tau}}{\sum_{t_\tau} l_{it_\tau}} \right)$ for individual i for day t. These coordinates are in the middle of the area covered by all the locations which were visited during the day and are dominated by the two most prominently used (longest used) locations: home and work location.

Denote $d^2_{it_\tau} = (\bar{\xi}_{it} - \xi_{it_\tau})^2 + (\bar{\eta}_{it} - \eta_{it_\tau})^2$ as the squared Euclidean distance between the coordinates $(\bar{\xi}_{it}, \bar{\eta}_{it})$ and $(\xi_{it_\tau}, \eta_{it_\tau})$. This requires the coordinate system to be local to obtain valid, i.e. less distorted results. Otherwise, a Haversine[1] distance could be used instead in case of epsg:4326 WGS-84 projection of the coordinates.

The ROG for individual i and day t is then defined as

$$R_{it} = \sqrt{\frac{\sum_{t_\tau} d^2_{it_\tau}}{\sum_{t_\tau} l_{it_\tau}}}, \tag{3}$$

and it thus represents a distance to the center of all the places of stay during that day t weighted by the lengths of the stay duration at the different places.

Details are available and especially a description of how the large quantity of data was handled is available in (Heiler *et al.* 2020).

Using additional metadata, an individual i can be assigned to a gender group (female, male), to an age group (here we consider the age groups in 15 year intervals: 15-29, 30-44, 45-59, 60-74, and 75+), and to an Austrian district of the daily night location to derive the groups. Since the distributions are generally very right-skewed, we work with the median per group and day in the following and also ensure k-anonymity for each one.

The resulting time series can be directly investigated in terms of their absolute information, and they can be compared to an analysis based on relative information.

3. Results

3.1. Mobility measured by ROG

The results reported in this section refer to the median values of the ROG per group. To

[1] https://en.wikipedia.org/wiki/Haversine_formula, accessed: 2022-07-28

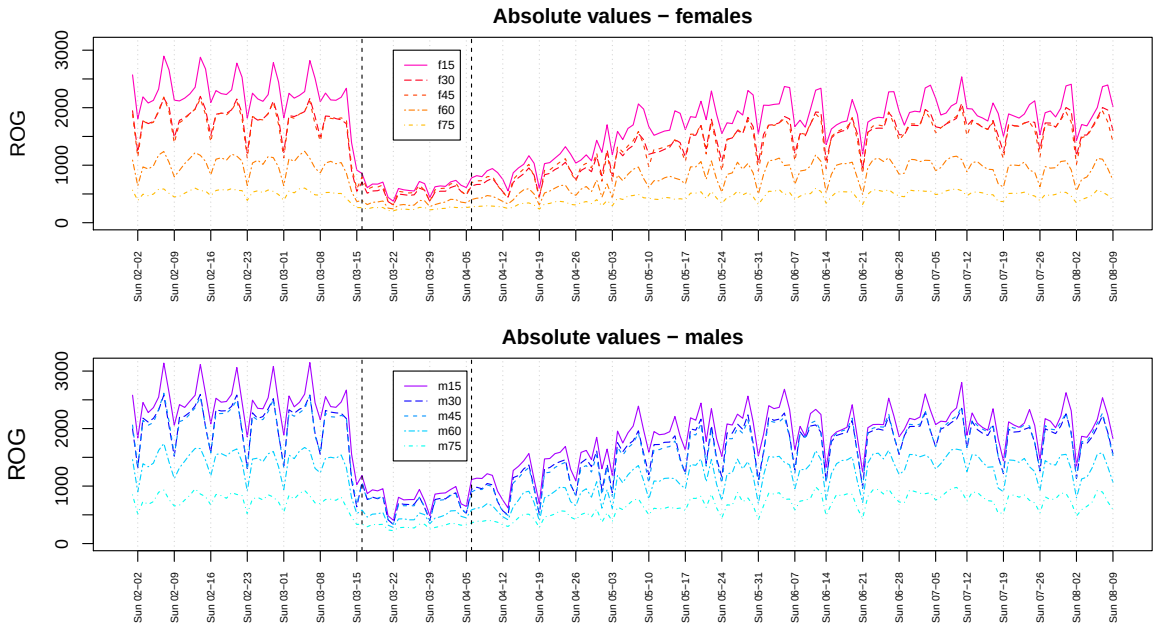

Figure 1: Median ROG values for different age groups over time for females (top) and males (bottom) in different age groups.

begin with, Figure 1 shows the absolute values for the females (top) and males (bottom) for different age groups. The legend indicates the considered age groups: 15 for age 15-29, 30 for age 30-44, 45 for age 45-59, 60 for age 60-74, and 75 for age elder than 75. For all of the following time series plots, the vertical dashed lines indicate the date March 16$^{\text{th}}$, 2020, when the restrictions came into action, and the date April 6$^{\text{th}}$, 2020, when they were relaxed. The data considered here are from the period February 1$^{\text{st}}$ until August 9$^{\text{th}}$, 2020. The plots clearly show the lockdown by an abrupt decay of the median ROG values in all age classes for both genders. After the lockdown, the order of the values remains the same, from the eldest group with the smallest values, and the youngest group with the highest values, but it is on a much smaller level. The level then increased more or less systematically until the middle of June. Afterwards, the level is not changing a lot, it is lower than at the beginning, and weekly patterns are visible. Note that these weekly time series patterns that are very regular at the beginning are getting somehow distorted, partially also due to holidays (April 13$^{\text{th}}$, May 1$^{\text{st}}$, May 21$^{\text{th}}$, June 1$^{\text{st}}$, June 11$^{\text{th}}$), and they never get back to this regularity.

In the subsequent analyses we consider the female age groups and the male age groups separately as two compositions. Note that it would also be possible to consider this data set as a multi-factorial composition, with age groups and gender as factors, and to analyze the complete composition jointly. Such an approach has been proposed in (Fačevicová, Filzmoser, and Hron 2022), and the mobility data set has been used as an illustration. Figure 2 focuses on the relative information contained in the median ROG values. The plots show the corresponding CLR coefficients for females (top) and males (bottom). While in Figure 1 we have essentially seen a decline of all values at the beginning of the lockdown phase, followed by an increase, we did not pay attention how differently the age groups declined and increased. This is the purpose of the relative view in Figure 2, where we mainly investigate the developments of the age groups to each other.

In both plots of Figure 2 we can see roughly the same pattern after the lockdown: the biggest relative changes are visible for the youngest and the oldest age group, but they go into different directions. While group 15 had the biggest decline, group 75+ increased the values relative to the other age groups. This seems to be counter-intuitive, but it can be explained by the fact that the geometric mean also went down significantly, and the ratio of the values of group 75+ to the geometric mean then even increased after the lockdown.

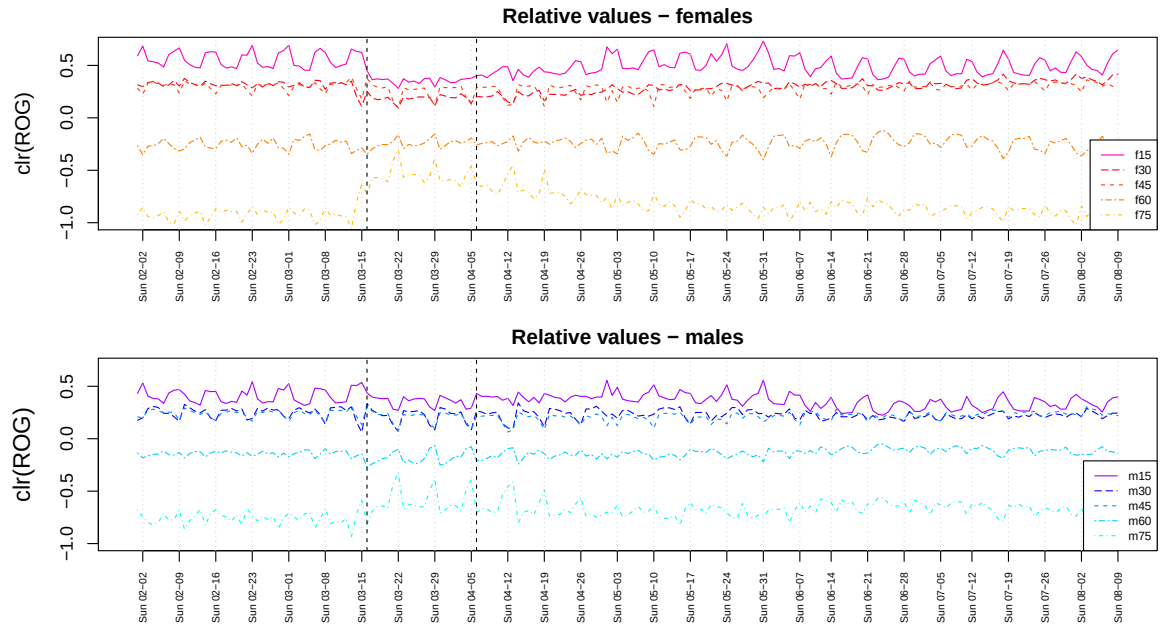

Figure 2: CLR coefficients of median ROG values for the female (top) and the male (bottom) composition. For the legend see Figure 1.

Another interesting phenomenon is that the groups 60 and 75+ show the biggest increase in mobility (in a relative sense) during the weekends in this lockdown period. Although on a different level, the values from July show a similar structure to those from February. Interestingly, the youngest age group 15 shows a somehow mirrored weekly pattern compared to the elder age groups. This is not visible when looking at the absolute values in Figure 1.

Relative information could also be understood in terms of data proportions. In particular, one could compute the proportion of a group on the total per time point, which in fact corresponds to normalizing the data per time point to a value of 1. Such a proportional presentation is shown in Figure 3 for the ROG values of the female age groups. Obviously, the information contained in this representation is different from CLR coefficients which focus on log-ratio information. One can hardly see any differences between the lockdown period and the remaining period, and thus this kind of "relative view" is not valuable for the analysis.

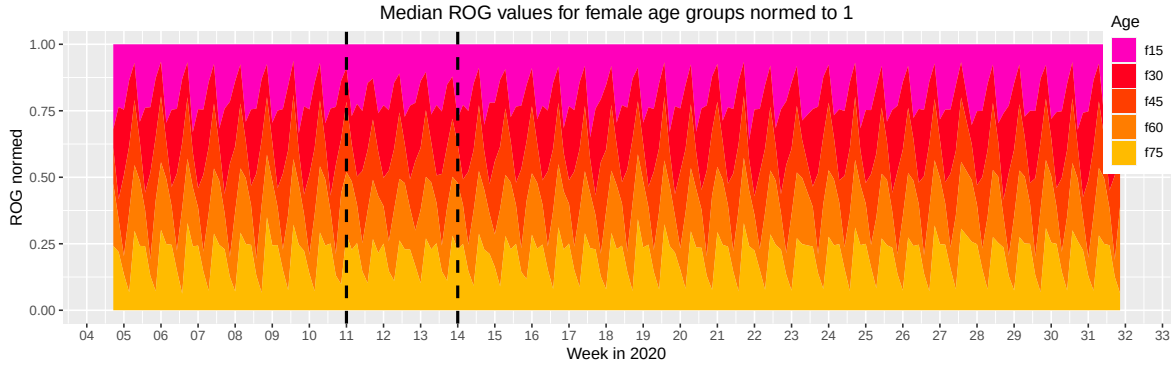

Figure 3: Proportional presentation of the median ROG values for the female age groups. For each time point, the data are normalized to a value of 1.

The median ROG values for the female and male age groups are analyzed in the following with PCA. Here, the method ROBPCA (Hubert, Rousseeuw, and Vanden Branden 2005) is taken, a robust version of PCA which downweights outlying observations. Figure 4 shows the biplot

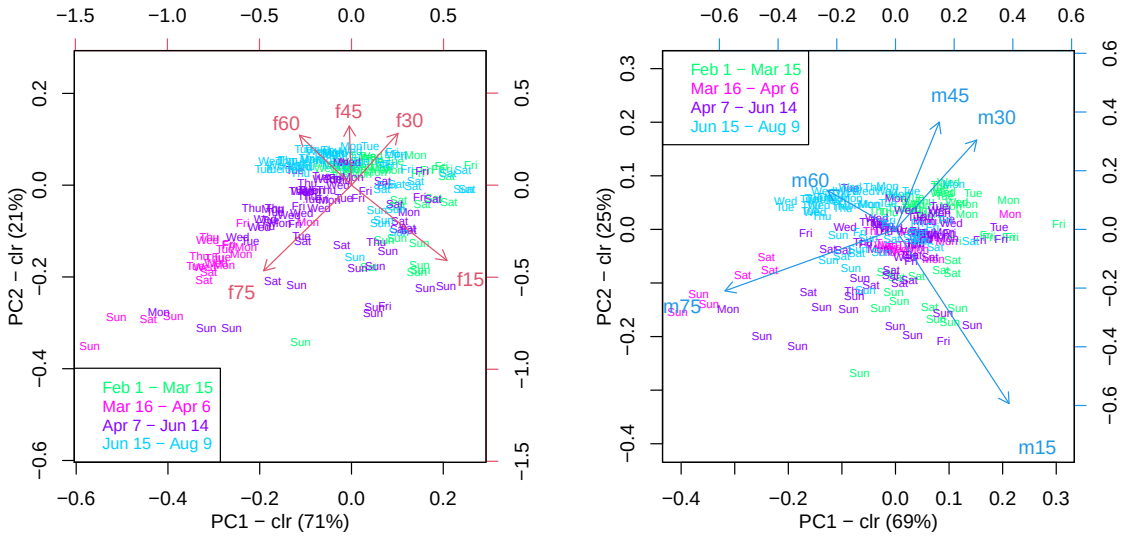

Figure 4: Biplots of the CLR coefficients of the median ROG values for females (left) and male (right) age groups. Green color for period before the lockdown, pink for lockdown period, purple after lockdown until mid of June, and light-blue after this period.

of the first two principal components (PCs) for the clr coefficients. This biplot presentation (as well as all subsequent biplots in this paper) is a so-called form biplot (Aitchison and Greenacre 2002), with favours the representation of the observations in the plot. The coloring is according to the time phases: green before the lockdown, pink during the lockdown period, purple after lockdown until mid of June, and light-blue after this period. The left biplot for the females identifies these four periods as clear clusters, while there is more overlap visible in the right biplot for the males. For the females, the direction of the first PC (71% explained variance) shows a transition of the relative ROG values from the young generation (f15, f30) before lockdown to the old (f75) one during lockdown, and then back to the center. Thus, younger and elderly females show a contrasting behavior in this time period, which was already observed in Figure 2 (top panel). The second PC (21% explained variance) shows also differences between the time periods, but it also reveals weekend effects. Especially on Sundays, the mobility for group f15 was bigger before and after lockdown, but it moved to group f75 during the lockdown phase.

The data structure in the biplot for the males (right plot) looks a bit different, but leads to similar conclusions. PC1 explains 69% and PC2 25% of the variance. Groups m75 and m15 have a similarly diverging behavior of Sunday mobility as observed for the females. The weekdays of the lockdown phase are in the center of the distribution, while for the females they were clearly moved towards group f75. On the other hand, the weekdays in the first time period (February 1st - March 15th) are better distinguishable from the weekdays of the last period (June 15 - August 9); a possible explanation is the fact that the working male population changed the mobility behavior more significantly than that of females due to home office.

A quite contrasting view is revealed in Figure 5, which shows the robust PCA results for the absolute values of ROG, for females (left) and males (right). In both analyses, PC1 explains 98% of the variability, and this direction essentially reflects the big change of the ROG over this period. Otherwise, there is not much information left in these analyses, reflecting the limited usefulness of absolute information if the task compares age groups.

3.2. Interaction measured by call duration

Figure 6 investigates the median call duration, reported in seconds, again for the two genders and the age groups. The upper plot shows the absolute values jointly for males and females.

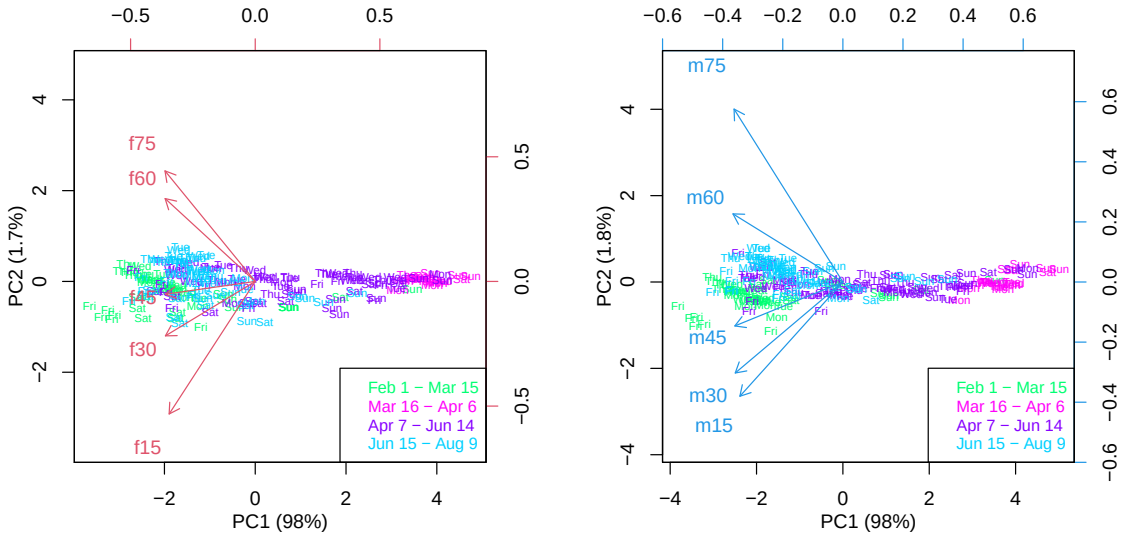

Figure 5: Biplots of the (absolute) median ROG values for females (left) and male (right) age groups. Green color for period before the lockdown, pink for lockdown period, purple after lockdown until mid of June, and light-blue after this period.

Here we observe the reverse ordering of the age groups compared to the plots for the ROG values: the lowest values are for the youngest group, and the biggest for the oldest group. The values of the females are systematically higher than those of the males. It is interesting to see that the call durations already started to increase one week before the lockdown. While the ROG time series had their peaks during the weekend, we have the opposite here. This pattern, however, seems to change after the lockdown for group f75 (uppermost line), and it went back to *normality* only later on.

The bottom plot of Figure 6 presents the CLR coefficients, which are separately calculated for females and males, but presented here jointly for easy comparison. Although the absolute values of the youngest age group also increased with the lockdown, the increase was smaller compared to the other groups, which is reflected by decreasing CLR coefficients. The pattern of f15 and m15 has also an interesting structure: Before the lockdown, the groups had quite different behavior within their gender-group, but during the lockdown phase they became quite similar. From June on, they show again a similar behavior as at the beginning. Another interesting phenomenon can be seen after the lockdown: the two oldest groups show a contrary behavior to the other groups during the weekends. Their decline in call duration during the weekends was much smaller than that of the other age groups.

Figure 7 presents biplots of a robust PCA for the CLR coefficients for the female (left) and male (right) age groups. The coloring is taken as in the previous biplots, green before lockdown, pink during, purple after lockdown, and light-blue from June 15$^{\text{th}}$ onwards. PC1 explains 72% of the variability for the females, 54% for males, and PC1 and PC2 together explain about 98% variance in both cases. The different groups which are visible in the biplots are essentially weekend-effects or affects due to the lockdown. These grouping effects are essentially caused by the youngest and oldest age groups. When comparing the first observed period with the last one, we can find clear differences in the corresponding PCA scores. These differences are essentially caused by the changing contrasting behavior between the youngest and elder groups; groups f75 and m75 (and also m30) do not seem to contribute to this difference. A possible explanation is the exploration of alternative communication methods, especially for the elder groups.

3.3. Interactions between source and destination

It can be recorded who is actively calling a person, and who is receiving a call. The former

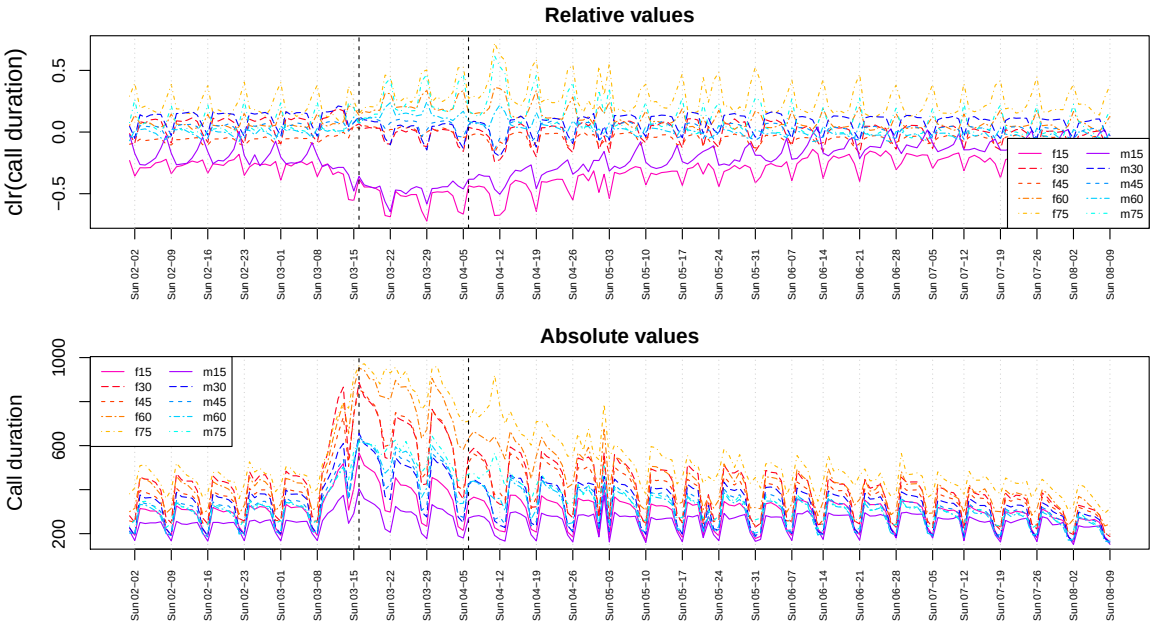

Figure 6: Median values of call durations per gender and age group over time (top), and CLR representations separately for female and male age groups (bottom).

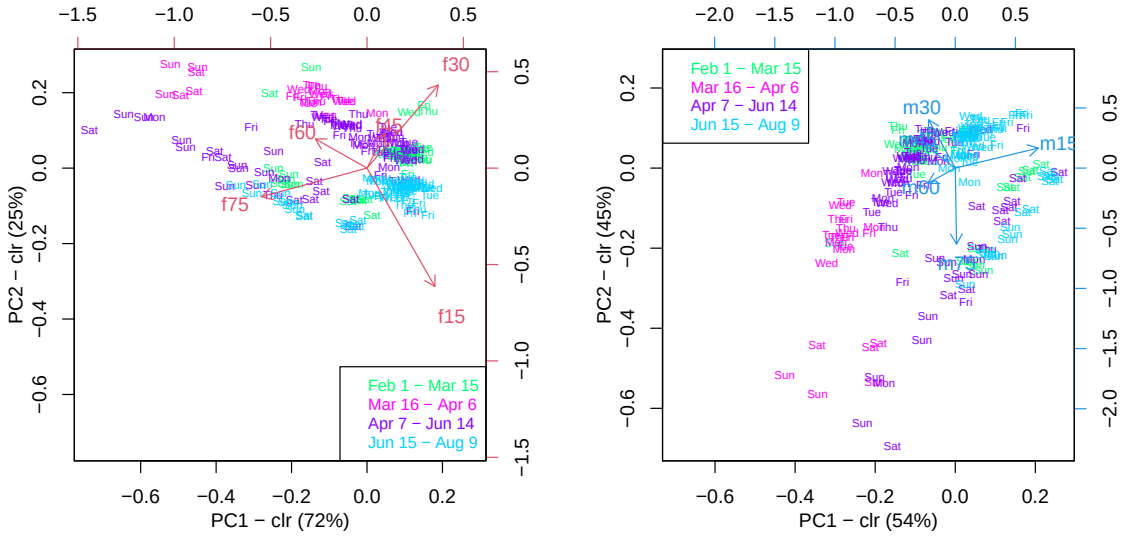

Figure 7: Biplots of the CLR coefficients of the median call duration values for females (left) and male (right) age groups. Green color for the period before the lockdown, pink for lockdown period, light-blue after this period.

person is called *source*, and the latter *destination*. Here we investigate the median ROG values for the different age groups of the females and males. However, the data set is more complex now, because a person from a specific age group can be source, while the destination can originate from a different age group. Moreover, both source and destination will have specific median ROG values.

Figure 8 illustrates these data for four specific cases: source f45 (f45_src) with destination f75 (f75_dst), and source f75 (f75_src) with destination f45 (f45_dst). In both cases, the median ROG values can be taken from the source group or from the destination group, see also figure legend. Throughout the whole period (here from February 1st - July 26th), the median ROG values from the source groups (solid lines) have slightly higher values than those of the destination groups (dashed lines) for the same age classes, which can be expected because people from the source groups might call from a place outside their usual environment. While the lines are on a similar level at the beginning and at the end of the considered period, the weekly periodicity changes, probably caused by the summer holidays.

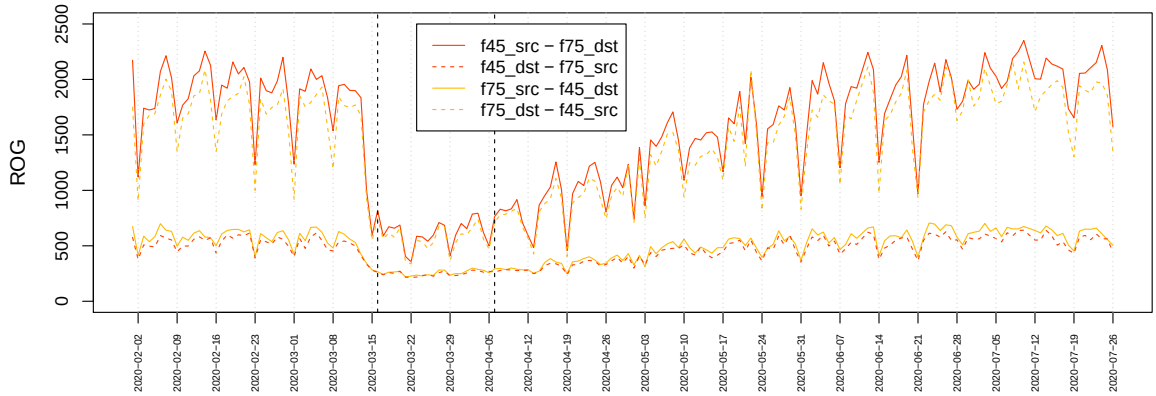

Figure 8: Median ROG values for the female age groups $f45$ and $f75$, depending whether they actively call (src) or they passively receive the call (dst). For example, line f75_src – f45_dst refers to the median ROG values for females in age group f75, actively calling females in age group f45.

In the following analyses we are interested in the similarity of the relative ROG values in terms of correlations, before lockdown (February 1st – March 15th) and after (March 16th – May 31th). In order to investigate relative information, the CLR coefficients are computed for a composition with all 25 age combinations of the source-destination groups and all 25 age combinations of the destination-source groups, separately for females and males. Figure 9 shows the resulting correlation matrix for the females as a heat map, left for time points before the lockdown, and right after lockdown. The row and column labels are referring to the group numbers. For example, src1-3 refers to the time series f15_src – f30_dst, or dst5-1 is the series f75_dst – f15_src. The heatmaps show that the correlation structure before and after lockdown has clearly changed. Afterwards, there are more blocks with higher (absolute) correlations, and thus more similarity or dissimilarity between certain age groups. In general, there is a more pronounced difference after lockdown in the mobility behavior between the younger and the elder age groups.

3.4. Incorporating spatial location

The mobile phone data also contain information about the location, in our case about the Austrian political district in which the phone has been used. The Austrian regions had different restrictions during the lockdown phase, and in particular people from all districts in Tirol had the most substantial movement restrictions. In this section we will provide some analysis examples based on this additional information, rather than going again into

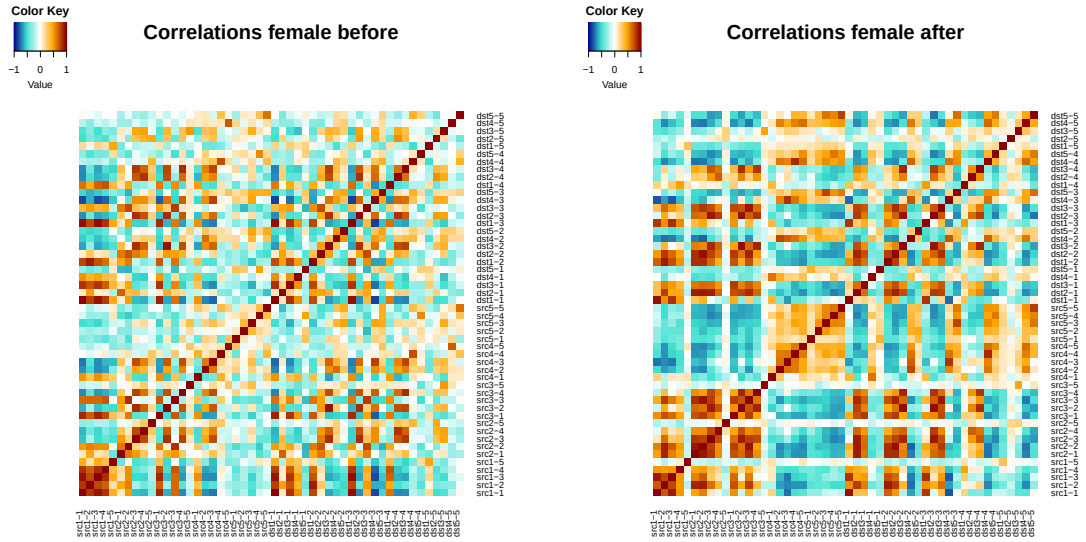

Figure 9: Correlations of the CLR coefficients for median ROG values for the female age groups (1 to 5, referring to 15 to 75+), when they are actively calling (src) or passively receiving a call (dst), recorded before March 16$^{\text{th}}$, 2020 (left), and afterwards (right).

detail with gender comparisons. Thus, in Figure 10 we compare the median ROG values for Kitzbühel, a district in Tirol, and Zell am See, which is also a rural district but located in Salzburg. The absolute values of the female age groups are shown in the upper plots, while the CLR coefficients are presented in the lower plots. Since the same scale is used along the vertical axes, one can clearly see the difference in mobility during the lockdown period in Kitzbühel and Zell am See, and this is also visible in the CLR coefficients. For Kitzbühel, there is much smaller variability of the values during lockdown, and also the relative differences between the age groups become much smaller. The change in the relative differences is not so pronounced for Zell am See. This means that also from a relative point of view, the data structure changes completely in Kitzbühel due to the restrictions.

Figure 11 focuses on the male age group m30, and compares the composition of all districts in Tirol with that of all districts in Salzburg. The dashed lines refer to the district capitals (Innsbruck and Salzburg, respectively). These districts behave differently compared to the other districts which are rural with many people commuting to their work place. The values of the districts in Tirol (except Innsbruck) get closer to each other after lockdown, and they start to diverge only in the middle of April. This may be explained by a similar mobility behavior of the m30 group within this period is probably caused by home-office or reduced working time. This seems different in districts of Salzburg, where the CLR coefficients show more variability after lockdown.

4. Discussion and conclusions

Since the pioneering work of Aitchison (1986) on compositional data analysis, this type of analysis has often been misunderstood as being only applicable to data for which the observations sum up to one – thus proportional data. However, the more recent literature made clear that the constant sum constraint is not at all important because (logarithms of) ratios between the variable values are the building blocks for this analysis, which are unchanged by rescaling an observation. Even if this aspect is acknowledged, there is often the question raised whether a data set is compositional or not. This means, should the data be analyzed traditionally, or should it be processed with the tools from compositional data analysis. The application in this paper has shown that both types of analyses are appropriate, but they provide answers to different questions, and accordingly the outcomes have a different meaning

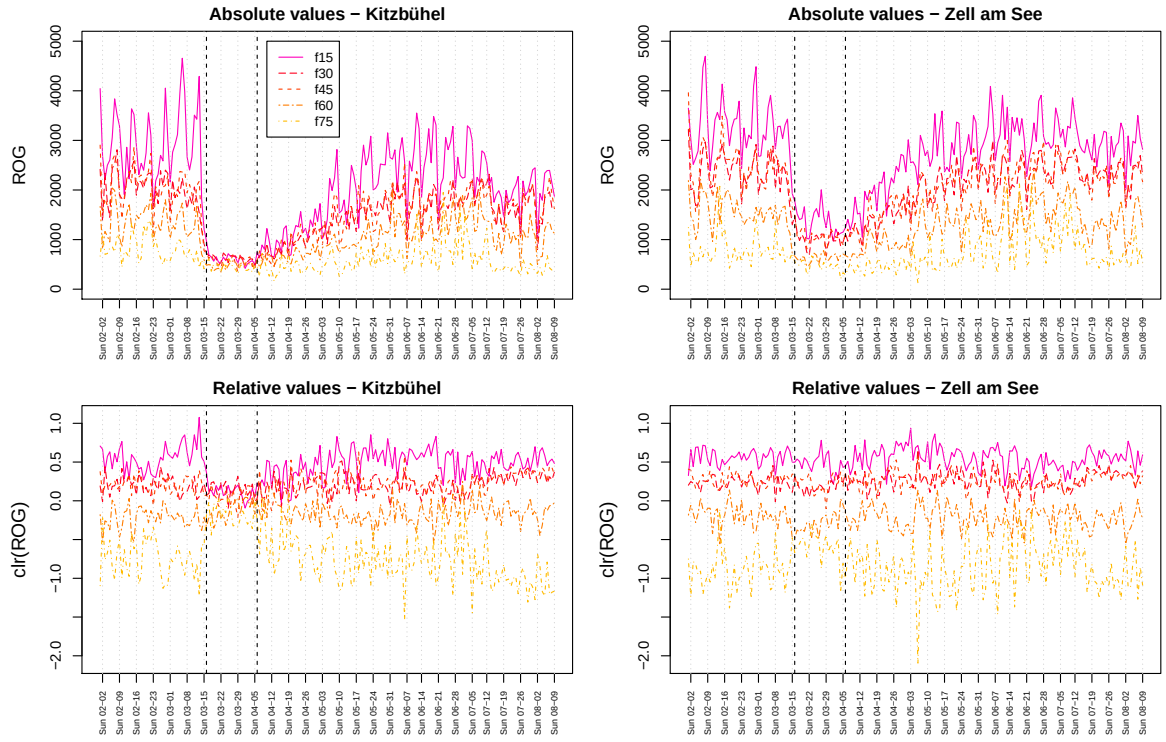

Figure 10: Median ROG values for Kitzbühel (left) and Zell am See (right) as absolute (top) and relative (bottom) information.

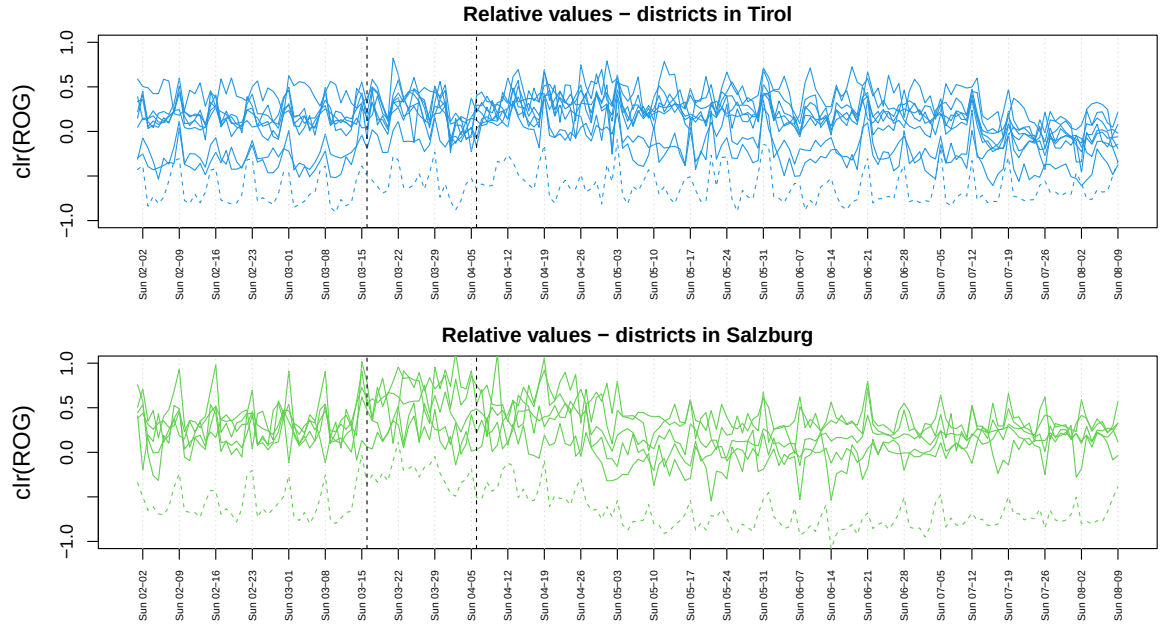

Figure 11: CLR coefficients of median ROG values for the male group m30 in all 9 districts of Tirol (top) and the 7 districts of Salzburg (bottom). The capitals are shown as dashed lines.

and interpretation.

Here the focus was on mobile-phone data, which allowed to analyze the mobility of people. Further, the duration of phone calls has been investigated, including the direction of the call. All this information has been recorded in the first half year of the COVID-19 pandemic, with a special focus on the first lockdown period. After more than two years of "experience" with this pandemic it is clear that this first lockdown had the most substantial effect on people's mobility. Here we investigated if the effect differs among gender and age groups and for different regions (political areas) in Austria. Specifically, we analyzed how these differences are expressed from the point of view of the absolute data, and also in terms of relative information by making use of the log-ratio analysis.

By analyzing relative changes using compositional data analysis methodologies, formerly hidden insights can be identified. We see that specific age groups of the population restrict mobility less than other population members. This is especially visible for elderly people who have been at high risk of infections, and the younger population especially during weekends. When analyzing the absolute values, this difference between the age groups is hardly visible. We just get the impression that overall, mobility strongly decreased for all age groups with the lockdown, and then slowly increased again. Also for call duration, the absolute numbers provide different insights than relative information, but both are interesting for the analyst.

In all our compositional analyses we have considered the age groups as the contributions to the composition, and the data have been analyzed separately for females and males. One could also consider both factors, age and gender, as splitting factors for the composition. However, in this case it is not at all obvious how the composition should be treated, since centered log-ratio coefficients, as an example, would have to be computed differently. An option would be using different coordinates, such as isometric log-ratio coordinates, where the choice of such coordinates would have to be done according to an appropriate interpretability (Pawlowsky-Glahn *et al.* 2015). Another attempt to link the information of both genders are multi-factorial compositional objects (Fačevicová *et al.* 2022), which are an extension of compositional tables, see Fačevicová, Hron, Todorov, and Templ (2018).

Acknowledgements

The WWTF funded this work under project COV COV20-035.

References

Aheleroff S (2011). "Customer Segmentation for a Mobile Telecommunications Company Based on Service Usage Behavior." *Proceedings - 3rd International Conference on Data Mining and Intelligent Information Technology Applications, ICMIA 2011*, (March), 308–313. ISSN 1314-4081. doi:10.5121/ijdkp.2015.5205.

Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London. (Reprinted in 2003 with additional material by The Blackburn Press). doi:10.2307/2982045.

Aitchison J, Greenacre M (2002). "Biplots for Compositional Data." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **51**(4), 375–392. doi:10.1111/1467-9876.00275.

Aksu H, Korpeoglu I, Ulusoy Ö (2019). "An Analysis of Social Networks Based on Tera-Scale Telecommunication Datasets." *IEEE Transactions on Emerging Topics in Computing*, **7**(2), 349–360. ISSN 21686750. doi:10.1109/TETC.2016.2627034.

Al-Molhem NR, Rahal Y, Dakkak M (2019). "Social Network Analysis in Telecom Data." *Journal of Big Data*, **6**(1). ISSN 21961115. doi:10.1186/s40537-019-0264-6.

Aledavood T, Lehmann S, Saramäki J (2018). "Social Network Differences of Chronotypes Identified from Mobile Phone Data." *EPJ Data Science*, **7**(1). ISSN 21931127. doi:10.1140/epjds/s13688-018-0174-4. 1709.06690.

Andrade X, Layedra F, Vaca C, Cruz E (2019). "RiSC: Quantifying Change after Natural Disasters to Estimate Infrastructure Damage with Mobile Phone Data." In *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pp. 3383–3391. ISBN 9781538650356. doi:10.1109/BigData.2018.8622374.

Chittaranjan G, Jan B, Gatica-Perez D (2011). "Who's Who with Big-five: Analyzing and Classifying Personality Traits with Smartphones." In *Proceedings - International Symposium on Wearable Computers, ISWC*, pp. 29–36. ISBN 9780769544380. ISSN 15504816. doi:10.1109/ISWC.2011.29.

Dao TN, Le D, Yoon S (2019). "Predicting Human Location Using Correlated Movements." *Electronics*, **8**(1), 54. ISSN 2079-9292. doi:10.3390/electronics8010054.

Dumuid D, Pedišić Ž, Palarea-Albaladejo J, Martín-Fernández JA, Hron K, Olds T (2020). "Compositional Data Analysis in Time-Use Epidemiology: What, Why, How." *International Journal of Environmental Research and Public Health*, **17**(7), 2220. doi:doi.org/10.3390/ijerph17072220.

Espinoza JL, Shah N, Singh S, Nelson KE, Dupont CL (2020). "Applications of Weighted Association Networks Applied to Compositional Data in Biology." *Environmental Microbiology*. doi:doi.org/10.1111/1462-2920.15091.

Fačevicová K, Filzmoser P, Hron K (2022). "Compositional Cubes: A New Concept for Multifactorial Compositions." *Statistical Papers*. doi:doi.org/10.1007/s00362-022-01350-8. To appear.

Fačevicová K, Hron K, Todorov V, Templ M (2018). "General Approach to Coordinate Representation of Compositional Tables." *Scandinavian Journal of Statistics*, **45**(4), 879–899. doi:doi.org/10.1111/sjos.12326.

Filzmoser P, Hron K, Templ M (2018). *Applied Compositional Data Analysis. With Worked Examples in R*. Springer Series in Statistics, Springer, Cham, Switzerland. doi:10.1007/978-3-319-96422-5.

Gao S, Rao J, Kang Y, Liang Y, Kruse J, Doepfer D, Sethi AK, Reyes JFM, Patz J, Yandell BS (2020). "Mobile Phone Location Data Reveal the Effect and Geographic Variation of Social Distancing on the Spread of the COVID-19 Epidemic." **586**. doi:doi.org/10.48550/arXiv.2004.11430. 2004.11430.

Gooch JW (2011). "Radius of Gyration." In *Encyclopedic Dictionary of Polymers*, pp. 607–607. Springer New York, New York, NY. doi:10.1007/978-1-4419-6247-8_9741.

Heiler G, Reisch T, Hurt J, Forghani M, Omani A, Hanbury A, Karimipour F (2020). "Country-wide Mobility Changes Observed Using Mobile Phone Data during COVID-19 Pandemic." In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 3123–3132. doi:10.1109/BigData50022.2020.9378374.

Heuzroth T (2020). "Corona-Pandemie: So hat Ischgl das Virus in die Welt getragen." https://www.welt.de/wirtschaft/article206879663/Corona-Pandemie-So-hat-Ischgl-das-Virus-in-die-Welt-getragen.html [Accessed: 2020-06-24].

Hillebrand M, Khan I, Peleja F, Oliver N (2020). "MobiSenseUs : Inferring Aggregate Objective and Subjective Well-being from Mobile Data." doi:10.3233/FAIA200297.

Hubert M, Rousseeuw PJ, Vanden Branden K (2005). "ROBPCA: A New Approach to Robust Principal Component Analysis." *Technometrics*, **47**, 64–79. doi:10.1198/004017004000000563.

Iacus Stefano, Serna CS, Sermi F, Spyratos S, Tarchi D, Vespe M (2020a). "How Human Mobility Explains the Initial Spread of COVID-19 (2)." doi:10.2760/61847.

Iacus Stefano, Serna CS, Sermi F, Spyratos S, Tarchi D, Vespe M (2020b). "Mapping Mobility Functional Areas (MFA) by using Mobile Positioning Data to Inform COVID-19 Policies (3)." doi:10.2760/076318.

Isaacman S, Frias-Martinez V, Frias-Martinez E (2018). "Modeling Human Migration Paterns during Drought Conditions in La Guajira, Colombia." In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2018*, 18. ISBN 9781450358163. doi:10.1145/3209811.3209861.

Jeffrey B, Walters CE, Ainslie KE, Eales O, Ciavarella C, Bhatia S, Hayes S, Baguelin M, Boonyasiri A, Brazeau NF, *et al.* (2020). "Anonymised and Aggregated Crowd Level Mobility Data from Mobile Phones Suggests that Initial Compliance with COVID-19 Social Distancing Interventions Was High and Geographically Consistent across the UK." *Wellcome Open Research*, **5**. doi:10.12688/wellcomeopenres.15997.1.

Jia JS, Lu X, Yuan Y, Xu G, Christakis A, Jia J, Nicholas A (2020). "Population Flow Drives Spatio-temporal Distribution of COVID-19 in China." *Nature*. doi:10.1038/s41586-020-2284-y.

Joueid A, Coenders G (2018). "Marketing Innovation and New Product Portfolios. A Compositional Approach." *Journal of Open Innovation: Technology, Market, and Complexity*, **4**(2), 19. doi:10.3390/joitmc4020019.

Lenormand M, Picornell M, Cantú-Ros OG, Louail T, Herranz R, Barthelemy M, Frías-Martínez E, Miguel MS, Ramasco JJ (2015). "Comparing and Modelling Land Use Organization in Cities." *Royal Society Open Science*, **2**(12). ISSN 20545703. doi:10.1098/rsos.150449. 1503.06152.

Liu H, Ma L, Zhao X, Zou J (2018). "An Effective Model between Mobile Phone Usage and P2P Default Behavior." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10861 LNCS, pp. 462–475. ISBN 9783319937007. ISSN 16113349. doi:10.1007/978-3-319-93701-4_36.

Marzuoli A, Liu F (2019). "A Data-driven Impact Evaluation of Hurricane Harvey from Mobile Phone Data." In *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pp. 3442–3451. ISBN 9781538650356. doi:10.1109/BigData.2018.8622641.

Monsivais D, Bhattacharya K, Ghosh A, Dunbar RI, Kaski K (2017). "Seasonal and Geographical Impact on Human Resting Periods." *Scientific Reports*, **7**(1). ISSN 20452322. doi:10.1038/s41598-017-11125-z. 1607.06341.

Nika A, Ismail A, Zhao BY, Gaito S, Rossi GP, Zheng H (2016). "Understanding and Predicting Data Hotspots in Cellular Networks." *Mobile Networks and Applications*, **21**(3), 402–413. ISSN 15728153. doi:10.1007/s11036-015-0648-6.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modeling and Analysis of Compositional Data*. Wiley, Chichester. doi:10.1002/9781119003144.

Pepe E, Bajardi P, Gauvin L, Privitera F, Lake B, Cattuto C, Tizzoni M (2020). "COVID-19 Outbreak Response: A First Assessment of Mobility Changes in Italy Following National Lockdown." *medRxiv*, p. 2020.03.22.20039933. doi:10.1101/2020.03.22.20039933.

Quinn TP, Erb I, Richardson MF, Crowley TM (2018). "Understanding Sequencing Data as Compositions: An Outlook and Review." *Bioinformatics*, **34**(16), 2870–2878. doi:10.1093/bioinformatics/bty175.

Reisch T, Heiler G, Hurt J, Klimek P, Hanbury A, Thurner S (2021). "Behavioral Gender Differences Are Reinforced during the COVID-19 Crisis." *Scientific Reports*, **11**(1), 1–12. ISSN 20452322. doi:10.1038/s41598-021-97394-1. 2010.10470.

Santamaria Serna Carlos, Sermi F, Spyratos S, Iacus S, Annunziato A, Tarchi D, Vespe M (2020). "Measuring the Impact of COVID-19 Confinement Measures on Human Mobility using Mobile Positioning Data (1)." doi:10.2760/913067.

Shi X, Lv F, Seng D, Xing B, Chen J (2019). "Exploring the Evolutionary Patterns of Urban Activity Areas Based on Origin-Destination Data." *IEEE Access*, **7**, 20416–20431. ISSN 21693536. doi:10.1109/ACCESS.2019.2897070.

Thomas-Agnan C, Laurent T, Ruiz-Gazen A, Nguyen THA, Chakir R, Lungarska A (2021). *Spatial Simultaneous Autoregressive Models for Compositional Data: Application to Land Use*, pp. 225–249. Springer International Publishing, Cham. doi:10.1007/978-3-030-71175-7_12.

Trinh HT, Morais J, Thomas-Agnan C, Simioni M (2019). "Relations between Socio-economic Factors and Nutritional Diet in Vietnam from 2004 to 2014: New Insights Using Compositional Data Analysis." *Statistical Methods in Medical Research*, **28**(8), 2305–2325. doi:10.1177/0962280218770223.

Vollmer M, Mishra S, Juliette H, *et al.* (2020). "Using Mobility to Estimate the Transmission Intensity of COVID-19 in Italy: A Subnational Analysis with Future Scenarios. Imperial College London. 2020." doi:10.1101/2020.05.05.20089359.

Xu B, Gutierrez B, Mekaru S, Sewalk K, Goodwin L, Loskill A, Cohn EL, Hswen Y, Hill SC, Cobo MM, Zarebski AE, Li S, Wu CH, Hulland E, Morgan JD, Wang L, O'Brien K, Scarpino SV, Brownstein JS, Pybus OG, Pigott DM, Kraemer MUG (2020). "Epidemiological Data from the COVID-19 Outbreak, Real-time Case Information." *Scientific Data*, **7**(1), 1–6. ISSN 20524463. doi:10.1038/s41597-020-0448-0.

Yabe T, Tsubouchi K, Fujiwara N, Wada T, Sekimoto Y, Ukkusuri SV (2020). "Non-Compulsory Measures Sufficiently Reduced Human Mobility in Japan during the COVID-19 Epidemic." pp. 1–9. doi:10.1038/s41598-020-75033-5.

Zhao P, Bucher D, Martin H, Raubal M (2020). "A Clustering-based Framework for Understanding Individuals' Travel Mode Choice Behavior." In *Lecture Notes in Geoinformation and Cartography*, pp. 77–94. ISBN 9783030147440. ISSN 18632351. doi:10.1007/978-3-030-14745-7_5.

Affiliation:

Peter Filzmoser
Computational Statistics
Institute of Statistics and Mathematical Methods in Economics
TU Wien
Wiedner Hauptstraße 8-10 1040 Vienna, Austria
E-mail: peter.filzmoser@tuwien.ac.at
URL: https://www.tuwien.at/mg/cstat

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 52 *Submitted:* 2022-03-23
July 2023 *Accepted:* 2022-08-12