Alma Mater Studiorum · University of Bologna

School of Science
Department of Physics and Astronomy "Augusto Righi"
Master Degree in Physics

Characterisation of the sewage microbiome in four European cities throughout the COVID-19 pandemic by means of ecological and network modelling

Supervisor:
Prof. Claudia Sala

Co-supervisor:
Prof. Daniel Remondini

Submitted by:
Ettore Rocchi

Academic Year 2021/2022

Abstract

This dissertation analyses time series data derived from sewage metagenomic samples, available within the VEO (Versatile Emerging infectious disease Observatory) project, a European initiative intended to increase and improve the generation and distribution of biological data with the scope of studying anti-microbial resistance (AMR) and monitoring emerging infectious diseases (EIDs).

The main purpose of this work was to investigate possible spatial and temporal patterns occurring in the sewage bacterial content of four cities (Bologna, Budapest, Rome, and Rotterdam) over time (from March 2020 to November 2021), also considering the possible effects of the lockdown periods due to the COVID-19 pandemic.

First of all, we started the analyses by evaluating the between samples diversity, looking for similarities (or dissimilarities) among the four cities, as well as among different time periods (seasonality). To this aim, we computed both similarity networks and Principal Coordinate Analysis plots based on the Bray-Curtis metric, a measurement of dissimilarity commonly used in ecology to compare samples based on their taxonomic composition.

Ecological techniques were also taken into account to estimate the α-biodiversity of the samples. This was achieved by means of different diversity indices (e.g. Shannon, Pielou, Chao, etc.), which take into account different ecological features: species richness, evenness, and taxonomic distance. By looking at the temporal behaviour of the biodiversity in the four cities, we noticed an abrupt decrease in both Rome and Budapest in the Summer of 2020. This collapse of biodiversity was further investigated.

The first interesting result is that the Rotterdam samples seem to be very different with respect to those from the other cities, in terms of both variability and stationarity. In particular, we observed a peculiar low variability in the Rotterdam samples, which seems to be related to the species of *Pseudomonas* genus. Such species are in fact highly variable and plentiful in the other cities, but are not among the most abundant in Rotterdam.

Secondly, we observed that no seasonality effect emerged from the time series of the four cities.

These results are confirmed by the Bray-Curtis-based Principal Coordinate Analysis, where three clusters (corresponding to the samples of Budapest, Rome, and, especially, Rotterdam) can be recognised, while no evident separation among samples collected during different seasons is observed.

In terms of temporal behaviour of the within-sample α-biodiversity, the most important observed feature is the occurrence of a minimum of biodiversity in the Summer of 2020 in Rome and Budapest, which is related to two different aspects: the prevalence of some species when the minimum occurred, namely the *Pseudomonas* spp., and the change in correlations among species, which is enriched in the period of minimum biodiversity. Moreover, the Rotterdam time series is proved to be stable and stationary also in terms of $\alpha-$biodiversity.

The last consideration concerns the impact of the periods of lockdown imposed by the COVID-19 pandemic: unfortunately data on several of these periods are not available. From the limited data available, no effect of the lockdown on the time series considered emerges.

Within the VEO project, sewage samples have been collected also for the following periods (and are still being collected); however they have not been sequenced yet. Once the samples will be sequenced, the analyses proposed in this dissertation will be performed also on those data, so that the possible effects of lockdowns may be studied.

Contents

Chapter 1

Introduction

This dissertation analyses time series data derived from sewage metagenomic samples, available within the VEO project.

In this introduction, a brief overview of what metagenomics is, as well as, a presentation of the VEO project are given. Lastly, the main idea behind the work rationale is reported.

1.1 From genomics to metagenomics

The word *genome* was created by the German botanist Hans Winkler in his book *Dissemination and Cause of Parthenogenesis in the Plant and Animal Kingdom* [1], but its etymology is not certain (it could be the blend of the words gene and chromosome or could derive from the Greek verb γίγνομαι, which means "to become"). The *genome* is the nucleotide sequence of DNA of an organism, and thus contains the genetic information of an organism.

Genomics, which can be considered a branch of molecular biology, deals with the study of the entire genome of living organisms. It can be said that genomics was born in 1980, when the entire genome of a virus, the Φ-X174 phage, was sequenced. The sequencing of the first complete genome of a bacterium (precisely *Haemophilus influenzae*) dates back to 1995.

Since that time, the sequencing of the whole genome of living organisms has progressed on increasingly complex organisms, both thanks to the use of increasingly advanced sequencing techniques, and thanks to bioinformatics tools for the management of large amounts of data.

In 1986 the Human Genome Project was launched for the sequencing of the entire genome of humans. This project led to a first publication in 2000 (which however concerned about 90% of the entire genome with a not always satisfactory accuracy), to be substantially completed in 2003 with the publication of 99% of the genome with an accuracy of 99.9%.

Parallel to the development of genomics, metagenomics was born between the

end of the 70s and the beginning of the 80s (thanks to the pioneering work of Woese [2], to which Pace and colleagues gave concrete application [3]). Metagenomics is also defined as community or environmental genomics, and is based on the direct extraction, amplification, and sequencing of the microbial DNA derived from a biological sample. The initial consideration was based on the evidence that most microorganisms were not (and are not) culturable. Pace and colleagues, hence, suggested to study the microbial content of a biological sample exploiting the sequencing of highly conserved genes containing hyper-variable regions. Specifically, the authors suggested the use of the 16S rRNA gene as phylogenetic marker[4]: comparing the measured 16S rRNA sequences with previously annotated databases allows to reveal the microbial content of a sample.

At first, metagenomics was based on two classical approaches: the first one, called function-based screening, was based on the heterologous expression of the gene sequences obtained from the environmental sample; the second one, called sequence-based screening, acted selecting the clones on the basis of the presence of a specific sequence of interest (which must be already known, at least partially).

Later, these approaches were replaced by an approach called Whole Metagenome Shotgun Sequencing, which do not focus on the single gene or genome, but examines the entire biodiversity of the microbial community. Among other things, this approach allows to evaluate both the microbial content (i.e. which species are present and with which relative abundance), and the genetic content (in terms of microbial genes) of a sample. Another great advantage of this approach is its ability to identify homology-free genes with already known sequences, which leads to enormous potential in terms of biodiscovery.

Shotgun sequencing is based on random splitting of DNA into numerous small segments; these segments are sequenced using the chain termination procedure, which consists in the random incorporation of dideoxynucleotides through the action of DNA polymerase during the in vitro DNA replication. The method then involves electrophoresis of the DNA fragments. However, nowadays, the most used method is the one called Next Generation Sequencing. This method is developed in the following phases: first of all, the DNA to be sequenced is extracted, which is then broken into fragments usually between 350 and 500 bp. The -3' ends are adenylated so that they can bind to a thymine placed at the end of an adapter. At this point the DNA amplification takes place with the Polymerase Chain Reaction (PCR) technique. The Illumina method, which is one of the most widely used, involves the use of a slide to which specific DNA fragments are adhered which allow the anchoring of the sections of DNA to be amplified (and subsequently to be sequenced). When the DNA copies have been amplified by PCR, resulting in a so-called DNA cluster, sequencing begins. In the presence of DNA polymerase and labeled nucleotides, each time a nucleotide is added, it emits a specific fluorescence

which is detected and recorded.

In this way, in shotgun sequencing the small DNA fragments are sequenced, obtaining the so-called reads. By repeating these operations of random fragmentation and sequencing of the fragments obtained, the superposition of the reads allows the reconstruction of the whole starting DNA.

By exploiting these procedures (which are gradually being refined) the Whole Metagenome Shotgun Sequencing is achieved; it allows to identify, using marker genes present in a differentiated way in all organisms, the entire structure of the community. It was thus possible to identify, for example, the community structure of the human gut microbiome. Whole Metagenome Shotgun Sequencing is computationally very complex; moreover, it requires reads of adequate length, which can be combined either with a de novo assembly or mapping on a reference database.

The last step is therefore to identify what has been sequenced. However, this task is not trivial, since most of the sequences will not find a 100% correspondence with sequences already known and stored in dedicated databases.

It is then possible to use marker genes (single genes or gene families) to evaluate, with phylogenetic analyses, which are the most represented taxa. For bacteria, the marker of choice is the already cited 16S rRNA gene, that is a sequence universally shared by all prokaryotes and presents extremely conserved regions, interspersed with highly variable regions (numbered from V1 to V9, of different length, with different variability and therefore very useful for phylogenetic reconstructions). These variable regions can be amplified and sequenced thanks to the use of primers.

Sometimes it is necessary to assemble the reads de novo to generate so-called super-contigs (i.e. large pieces of genome) or even complete genomes to be characterised.

The most common strategy is the one called binning, or the assignment of redundant sequences to the same OTU (Operative Taxonomic Unit, a term based on the arbitrary definition of a taxonomic unit exclusively based on the criterion of sequence divergence).

The term binning means the grouping of taxa (species, genera and families) on the basis of shared characteristics obtainable from the analysis of some properties of their variable sequences: frequency of observation of di-, tri- or tetra-nucleotides, content in GC (Guanine-Cytosine), etc.

In general, binning concerns the grouping of sequences deriving from the sequencing of phylogenetic markers on the basis of similarity criteria. The comparison takes place between the reconstructed sequences and a reference database. A minimum similarity threshold is set above which it is possible to proceed with the assignment to a certain taxonomic group (for example $> 97\%$ for the species, $> 90\%$ for the genus, $> 80\%$ for the family). It is easy to understand that these thresholds are chosen arbitrarily, as a "rule of thumb", derived from empirical and common sense considerations: in fact, higher thresholds could lead to an overestimation of

biodiversity, while the use of lower thresholds could lead to incorrect classifications.

All sequences included in an OTU must have a level of mutual similarity $> 97\%$. Furthermore, among all the sequences a reference one must be identified, which should hopefully be the most abundant. It should also be noted that among the OTUs a continuous spectrum of similarity between sequences can sometimes be generated and this can determine the possible classification of some sequences in more than one OTU.

Once the OTUs with their respective frequencies have been defined (essentially by exploiting the count of reads that fall into each OTU and relating it to the total number of informative reads) the reference sequences of each OTU must be compared with a reference database (for example SILVA or NCBI). If the similarity between the reference sequence of the OTU and that of the database exceeds a certain threshold it can be reached at the species level, otherwise the classification stops at higher taxonomic levels.

1.2 The VEO project

The Versatile Emerging infectious disease Observatory (VEO) is a European initiative intended to increase and improve the generation and distribution of biological data within the scope of studying anti-microbial resistance (AMR) and monitoring emerging infectious diseases (EIDs).

The employed strategy of VEO in studying and analysing these subjects is the innovative One Health perspective.

One Health is an integrated, unifying approach to health studies, based on the fact that human health is closely linked to the healthiness of food, animals and the environment. Thus, it "aims to sustainably balance and optimise the health of people, animals and ecosystems" [5].

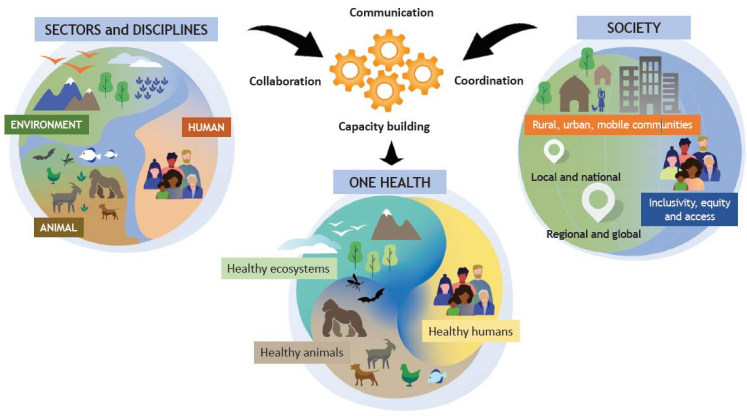


Figure 1.1: Schematic description of One Health approach

While the traditional approach on AMR and EIDs studies is human-centred

and usually based on ecological models subsequent to spillover occurrence (from animals to humans), the VEO approach works on an extended and longsighted domain (as can be seen in Figure 1.2), taking advantage of collaborations among multiple scientific branches.

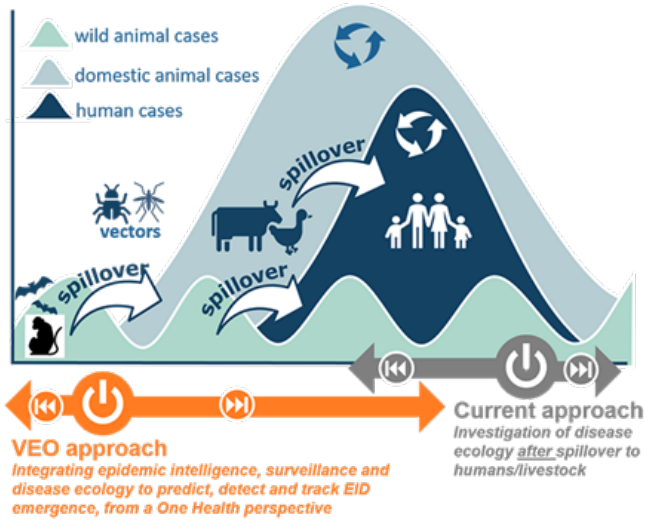


Figure 1.2: Schematic description of VEO approach

The VEO project provides diverse data types which may be analysed from many different perspectives; this underlying interdisciplinary may contribute to predict, detect, and track many global health threats.

The work presented here is based on data available within the VEO project, that is the metagenomic compositions (in terms of abundances) of samples taken from sewage of different cities, and in different times. The analyses of this dissertation and their purposes will be briefly presented in the next section, but deeply described later on, in chapter 2.

1.3 Work rationale

The main purpose of this work was to investigate possible spatial and temporal patterns occurring in the sewage bacterial content of four European cities, starting from the metagenomic data available within the VEO project and taking into account the possible effects of the lockdown periods due to the COVID-19 pandemic.

In order to achieve this goal, several time series techniques were used [6], and an ecological study on the biodiversity was performed, both within and between samples.

First of all, the samples were investigated in search of particular periodicities, for instance related to seasonality. In order to do so, the Principal Coordinate Analysis, starting from the Bray-Curtis distance, was performed considering separately

the samples collected in each city. Moreover, a measure of similarity among all the samples was computed (and visualised through PCoA), so that to investigate possible spatial patterns.

Other ecological analyses have been performed: the α-diversity of each sample was computed by means of biodiversity indices. In this way, the temporal behaviour of biodiversity could be explored, taking into account both species richness and evenness, but also the taxonomic distances inside the bacterial content of sewage. The indices considered to evaluate the within sample biodiversity were the well known Shannon, Pielou, Gini-Simpson and Chao indices, the Hill numbers, but also the two taxonomic distinctness indices proposed by Clarke and Warwick.

Then, an in-depth investigation of sub-periods of the time series was proposed in order to analyse the drops of biodiversity occurring in the Rome and Budapest time series. This analysis was based on a correlation network technique, used to highlight changes in the relationships among species during the sub-periods, as well as on a statistical inspection of the temporal trend of the most abundant species, in search of significant changes in terms of abundance that could explain the drop in biodiversity.

Finally, some brief considerations about the effects of the lockdowns due to the COVID-19 pandemic were reported.

Chapter 2

Material and Methods

In this chapter, a sketch of the data used in this dissertation is presented, along with the preprocessing steps required to work properly with them.

Then, an in-depth theoretical description of the different performed analyses is provided.

2.1 The data

Among the diverse data collected within the VEO project, this dissertation focused on the results of metagenomic data derived from the sewage of four European cities: Bologna, Budapest, Rome, Rotterdam. The sampling was repeated for a period of at least 40 weeks, with a frequency of one sample every two weeks, on average.

The extracted metagenome was then sequenced and aligned to databases in order to identify the taxa composition of each sample. The considered taxonomic levels are seven: species, genus, family, order, class, phylum and superkingdom. The data are then organised in abundance tables, one for each taxonomic level; a column represents a certain taxon, while a row represents a certain sample (characterised by city and time of isolation). The file is thus filled with the number of fragments of DNA counted in each sample, for each taxon.

In this way, seven *tsv* file were built, one for each taxonomic levels. In this dissertation, the analyses will be carried out starting from those data. Table 2.1 summarise the number of samples and the time period covered by each time series, that is by each city's dataset.

More details on the metagenomic samples are collected in the appendix A.

As an example, Figure 2.1 shows a piece of one *tsv* file (the phylum one). It is worth noticing that the first column contains the sample ID, which in turn may be used to get useful information interpolating it in the metadata *tsv* file. This file collects sample IDs and city and date of isolation.

The power of these data resides not only in the fact that they are longitudinal

City	Samples	Starting time	Final time
Bologna	28	2020-03-12	2021-04-27
Budapest	26	2020-05-18	2021-05-17
Rome	20	2020-03-17	2020-12-09
Rotterdam	39	2020-04-08	2021-11-03

Table 2.1: Time series information

(time series may be seen as one-dimensional longitudinal data) but also in their variety in terms of place of origin at European scale. Hence, both spatial and temporal considerations may be taken into account.

```
sample  Proteobacteria  Bacteroidetes   Cercozoa        Actinobacteria  unknown Streptophyta    Nematoda        Firmicutes      Fusobacteria    Discosea        Ciliophora
Platyhelminthes Tenericutes     Chordata        Chlorophyta     Apicomplexa     Euglenozoa      Euryarchaeota   Ascomycota      Cyanobacteria   Chlorobi        Acidobacteria   Synergistetes
Basidiomycota   Thermotogae     Zoopagomycota   Candidatus Cloacimonetes        Spirochaetes    Annelida        Arthropoda      Chloroflexi     Elusimicrobia   Planctomycetes
Bacillariophyta Verrucomicrobia Rotifera        Cremarchaeota   Rhodophyta      Mucoromycota    Onychophora     Candidatus Gracilibacteria      Lentisphaerae   Candidatus Campbellbacteria
Armatimonadetes Chlamydiae      Candidatus Saccharibacteria     Candidatus Sumerlaeota  Kiritimatiellaeota      Candidatus Peregrinibacteria    Candidatus Parcubacteria        Deinococcus-
Thermus Cryptomycota    Candidatus Hydrogenedentes      Candidatus Kapabacteria Candidatus Aenigmarchaeota      Thaumarchaeota  Candidatus Shapirobacteria      Caldiserica     Candidatus
Moranbacteria   Candidatus Woesebacteria        Candidatus Nomurabacteria       Nitrospinae     Candidatus Yonathbacteria       Gemmatimonadetes        Fibrobacteres   Rhodothermaeota
Candidatus Dadabacteria Nanoarchaeota   Candidatus Curtissbacteria      Abditibacteriota        Candidatus Beckwithbacteria     Candidatus Collierbacteria      Perkinsozoa     Candidatus
Falkowbacteria  Mollusca        Candidatus Kaiserbacteria       Cryptophyta     Tubulinea       Foraminifera    Candidatus Liptonbacteria       Candidatus Jacksonbacteria
Ignavibacteriae Candidatus Hydrothermae Cnidaria        Candidatus Omnitrophica Candidatus Melainabacteria      Hemichordata    Candidatus Pacebacteria Evosea  Candidatus Magasanikbacteria
Candidatus Colwellbacteria      Candidatus Azambacteria Candidatus Berkelbacteria       Candidatus Adlerbacteria        Candidatus Wolfebacteria        Candidatus Levybacteria Candidatus
Roizmanbacteria Candidatus Kerfeldbacteria      Porifera        Nitrospirae     candidate division KD3-62       Candidatus Fermentibacteria     Balneolaeota    candidate division
Zixibacteria    Candidatus Micrarchaeota        Candidatus Gottesmanbacteria    Candidatus Vogelbacteria        Candidatus Atribacteria Candidatus Doudnabacteria       Candidatus
Latescibacteria Candidatus Margulisbacteria     Candidatus Yanofskybacteria     Chytridiomycota candidate division WWE3 Candidatus Spechtbacteria       Tardigrada      Candidatus
Bipolaricaulota Candidatus Poribacteria Blastocladiomycota      Candidatus Uhrbacteria  Bryozoa Candidatus Staskawiczbacteria   Xenacoelomorpha Candidatus Daviesbacteria       candidate
division CPR3   Aquificae        Candidatus Kuenenbacteria       Gastrotricha    Candidatus Giovannonibacteria   Candidatus Kryptonia    Calditrichaeota Candidatus Microgenomates
Echinodermata   Deferribacteres Candidatus Tectomicrobia        candidate division NC10 Candidatus Raymondbacteria      Candidatus Nealsonbacteria      Microsporidia   Heterolobosea
Candidatus Amesbacteria Thermodesulfobacteria   Candidatus Zambryskibacteria    Candidatus Chisholmbacteria     Candidatus Abawacabacteria      Entoprocta      Candidatus
Hydrothermarchaeota     Candidatus Marinimicrobia       Coprothermobacterota    Candidatus Lloydbacteria        Candidatus Altiarchaeota        Candidatus Korarchaeota Hemimastigophora
candidate division CPR1 Candidatus Wildermuthbacteria   Candidatus Komeilibacteria      Candidatus Buchananbacteria     Candidatus Calescamantes        Candidatus Ryanbacteria Candidatus
Schekmanbacteria        Olpidiomycota   Candidatus Firestonebacteria    Candidatus Portnoybacteria      Candidatus Jorgensenbacteria    Nematoda        Rhombozoa       Candidatus
Terrybacteria   Dictyoglomi     Ctenophora      Candidatus Diapherotrites       Brachiopoda     Picozoa Candidatus Aminicenantes        Candidatus Edwardsbacteria      Chrysiogenetes
Candidatus Harrisonbacteria     Candidatus Verstraetearchaeota  Chaetognatha    Candidatus Odinarchaeota        Candidatus Woykebacteria        Candidatus Veblenbacteria       Candidatus
Geothermarchaeota       Candidatus Blackburnbacteria    Acanthocephala  Candidatus Aerophobetes
DTU_2022_1015109_1_MG_IT_RO_201209_GSI_373      93206   32914   174     658     1158    683     388     5202    180     3       47      37      519     181     12      2       15      176
57      171     8       44      70      22      5       2       6       17      634     57      71      500     43      6       115     9       2       5       4       1       11      3
15      5       13      16      3       31      1       2       2       9       14      3       1       1       4       4       3       1       2       3       1       5       1       5
2       1       1       1       1       1       1       2       2       8       2       2       4       1       4       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
DTU_2022_1015078_1_MG_IT_RO_200421_GSI_B4_S28_L001_R1_001       87086   12887   76      2270    1105    637     1874    12466   1254    0       69      12      217     241     34      2
6       198     64      285     3       111     472     40      15      1       8       63      174     201     115     541     71      2       483     104     3       15      7       0
11      10      1       31      12      33      13      199     2       3       6       6       23      4       0       11      4       3       9       2       3       20      0       11
24      2       4       1       1       0       0       2       0       1       29      4       4       4       5       0       2       1       0       2       0       15      0       5
1       7       0       0       0       0       0       10      4       0       0       12      0       0       2       1       0       0       0       0       0       0       0       0
4       0       0       4       2       1       1       1       1       1       0       0       0       0       0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
0
```

Figure 2.1: Example of abundance table in *.tsv* format

Some preprocessing operations are required on these data, in order to appropriately perform analyses on them. Next section will examine the main issues occurring when dealing with this data type.

2.1.1 Data preprocessing steps

First, it is worth pointing out that each sewage sample is composed of not only bacteria, but also eukaryotes, archaea ad viruses. Nevertheless, only the bacteria population is examined in the performed analyses, since it is of much more interest when the aim is to detect and monitor potential health threats or to study antimicrobial resistance[1].

[1]Brief clarification: also viruses may be interesting to be analysed, of course. However, metagenomic samples can only take into account DNA viruses, because of the inherently weaker structure of RNA with respect to DNA. Hence, this results in an incomplete and not-large-enough portrayal of virus communities in the samples.

Hence, the first preprocessing step is to filter the samples, removing all those taxa which does not belong to bacteria (at each taxonomic level, except for superkingdom). To achieve this goal, the NCBI (National Center for Biotechnology Information) database [7] and the GitHub repository "*taxonomy_ranks*" [8] were used. In fact, a taxonomic reconstruction of every taxa is needed in order to establish its membership to the Bacteria superkingdom.

Unfortunately, during this step, some taxa were lost, i.e. the tool was not able to find them in the NCBI database; the list of missing taxa is reported in appendix A.

Also, it is worth mentioning that, at each taxonomic level, some DNA fragments were not identified at all, during the alignment procedure. Although those fragments give no information about the sample composition, their counts are collected in each file into the "unknown" column. These counts are omitted in the analyses, but further information may be found in appendix A.

All those steps, as well as, all the analyses that will be described later on, were performed via python.

Working with compositional data

Abundances of taxa in a sample are example of compositional data, i.e. of data representing "proportions of some whole", as stated by the Scottish statistician Aitchison (1982) [9]. As described in that article, this type of data requires to be handled carefully, since they live in a particular mathematical space called simplex.

A $(k-1)$-dimensional Aitchison simplex[2] is a a space described by:

$$\mathcal{S} = \left\{ \mathbf{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k \middle| \, x_i > 0, \; i = 1, \ldots, k; \; \sum_{i=1}^{k} x_i = \alpha \right\} \tag{2.1}$$

This space is equipped with an operation called closure:

$$\mathscr{C}(x_1, \ldots, x_k) = \left(\frac{x_1}{\sum_{i=1}^{k} x_i}, \; \ldots \; , \frac{x_k}{\sum_{i=1}^{k} x_i} \right) \quad . \tag{2.2}$$

which allows to normalise the data on the simplex 2.1 leading to the so-called probability simplex with the constraint $\sum_{i=1}^{k} x_i = 1$.

Another useful operation is the centred log-ratio(**clr**), which map the Aitchison simplex into a subset of the real space:

$$\begin{aligned} \mathbf{clr} : \; & S \longrightarrow U, \;\; U \subset \mathbb{R}^k \\ \mathbf{clr}(\mathbf{x}) = & \left(\log \frac{x_1}{g(\mathbf{x})}, \; \ldots \; , \log \frac{x_k}{g(\mathbf{x})} \right) \end{aligned} \tag{2.3}$$

[2]Here, the name *Aitchison simplex* is used to describe the sample space of compositional data and thus to distinguish it from the more general definition of simplex. Further details about simplices and *Aitchison simplices* may be found in appendix B.

where k and $g(\mathbf{x})$ are the dimension and the geometric mean of the vector $\mathbf{x}$, respectively.

Further details on Aitchison simplex may be found in appendix B.

The main obstacle which occurs when dealing with compositional data living in this space is the constraint on the $\mathbf{x}$'s components. In fact, x_i (for each $i = 1, \ldots, k$) has to be strictly positive and this means that:

- the Aitchison simplex does not include the boundary, thus it can be defined as an *interior* simplex;
- the Aitchison simplex may be thought as the set of probability distributions on a k-dimensional dataset where zero probabilities are forbidden; notice that, as a consequence, also probabilities equal to 1 are left out.

As a consequence, each sample containing zero DNA fragments of certain taxa is not suitable for being studied through Aitchison simplices, since the "probability" (that is the relative abundances, in this context) of those taxa will be null. Moreover, forcing those data to lie in that simplex can not work, because many operations may not be mathematically performed (for instance, the centred log-ratio will raise an error when zeros occur).

This problem may be solved through the multiplicative replacement strategy[3] [10]. Let us $\mathbf{x} = (x_1, \ldots, x_k) \in \mathcal{S}$ be a composition, i.e. a vector in the Aitchison simplex in 2.1, with Z zero components; the multiplicative replacement maps $\mathbf{x}$ into a new composition $\mathbf{r} \in \mathcal{S}$ without zeros:

$$r_i = \begin{cases} \delta & \text{if } x_i = 0 \\ \left(1 - \dfrac{Z\delta}{\alpha}\right) x_i & \text{if } x_i > 0 \end{cases} \qquad \forall i = 1, \ldots, k \tag{2.4}$$

where δ is a small positive correction parameter and α is the sum constraint in 2.1.

The parameter δ may be a constant set a-priori or may be a function of the number of $\mathbf{x}$ components. One usual choice is to set δ equal to $1/k^2$; this choice helps to avoid negative components in the new composition $\mathbf{r}$, since they have no meaning.

The multiplicative replacement strategy has some interesting properties:

1. the parameter δ does not depend on the number of zeros in the composition, Z (and, if it has been fixed to a constant value, neither depends on he dimension k);

[3]There are other ways to tackle the problem of zeros abundances, but, here, the multiplicative replacement strategy is chosen among them because it guarantees coherence with the Aitchison simplex's structure described above.

2. all the basic operations on the simplex (such as **clr** and closure) are preserved and coherent in the sub-composition of non-zero components;

3. ratios are preserved, that is:

$$\frac{r_i}{r_j} = \frac{x_i}{x_j} \tag{2.5}$$

for all non-zero values x_i and x_j; this implies that the covariance structure of non-zero components is preserved.

All the above discussion about Aitchison simplex has the purpose of highlight the importance of using the appropriate operations when dealing with compositional data, especially when correlation-based analyses are performed: this is a consequence of equation 2.5.

Coming back on the sewage case study introduced in this dissertation, the centred log-ratio transform was performed, as well as the multiplicative replacement strategy, in order to avoid spurious correlations among samples and to (approximately) maintain the properties and structure of the original data.

Working with time series

The time series structure of the data analysed in this dissertation allows several exploratory analyses regarding the bacterial community in sewage. In fact, different methods may be used to study significant features of the community, such as its stability over time or its response to external effects and perturbations.

Unfortunately, among the variety of possible techniques, some of them were not suitable for the provided dataset because of its properties and structure.

Of course, one of the main issues was the compositional data structure and the presence of zero components in the samples; this part was tackled by adopting the Aitchison simplex formalism and using the functions defined in this space, as already described in the previous section.

However, this was not the only problem; a non-exhaustive list of the questions concerning the data and their features is shown below:

1. The time series are composed of a small number of samples and the sampling frequency is not sufficiently large; this means that every analysis which requires a certain time series length are not suitable for the data (examples are the Hurst exponent and the Lyapunov exponent);

2. The sampling rate is not uniform, i.e. the samples belonging to the time series of a city were not extracted with a constant frequency (see 2.2 for further details);

3. Each composition, i.e. each one of the 113 samples, exhibits many zero components (zero counts associated to certain taxa, no matter the considered taxonomic level). Of course, this is an intrinsic feature of these type of data; nevertheless it may cause spurious and biased results in correlation analyses;

4. The time series cover a short period of time (at most 1 year and 7 month, in Rotterdam); this allows only minimal (and often non - significant) periodicity analyses.

The non-uniform sampling rate mentioned in point 2 may result in the unfeasibility of many correlation or periodicity - based techniques. Some of these types of analysis, namely the Fourier spectral analysis and the cross (and auto) - correlation, have been performed despite the (non-satisfied) hypothesis of equidistant time points. In fact, in many cases (and for specific techniques) the sampling rate may be approximated to uniform because of its narrow variability; the average time gap between two consecutive samples, as well as the standard deviation, are reported in Table 2.2, expressed in days.

City	**Average time gap [days]**	**Standard deviation [days]**
Bologna	15.2	2.9
Budapest	14.6	6.8
Rome	14.1	3.9
Rotterdam	15.1	4.5

Table 2.2: Descriptive statistics of time gaps

Obviously, this approximation may lead to (hopefully minor) distortions.

About point 3, instead, it is worth pointing out that the zero components is a tricky problem when dealing with Pearson's correlation. The Pearson's correlation coefficient among two vectors (Let us say X_1 and X_2) is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}} \tag{2.6}$$

where $\bar{x}_j$ is the mean of the n-dimensional $X_j = (x_{j1}, \ldots, x_{jn})$ vector's components.

In the context of taxa abundances in a sample, the vectors X_j (that is the sample composition) may have many zero components, which, in turn, will be transformed

via centred log-ratio (after the multiplicative replacement strategy). These "transformed zero components" may lead to biased correlations, since, as a consequence of equation 2.6, vectors with many identical components result in higher Pearson's correlation coefficient.

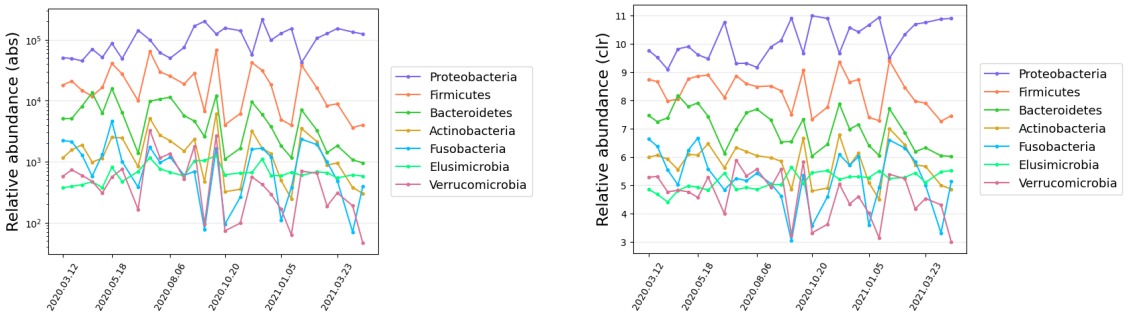
For the same reasons, the same problem arises when correlations analyses are performed not between sample compositions, but between time series of taxa abundances, particularly in the presence of rare OTUs.

Finally, a more precise description of the consequences of point 1 will be presented in the conclusive chapter 4.

From data to analyses

The above-described preprocessing steps as well as all the issues presented in the previous sections are intended to bring to light the intrinsic complexity of this type of data, especially when dealing with some techniques.

Before going on with a detailed description of the methods used to analyse the sewage compositions, a data example is depicted in Figures 2.2 a-b where the time-distribution of the seven most representative phyla of the Bologna's samples is shown.



(a) Absolute abundances (b) Relative abundances (centred log-ratio)

Figure 2.2: Example of time series (Bologna, at phylum level)

Few considerations about the time-distribution of taxa will be made in the next chapter (3), especially comparing the cities' composition and time series.

Instead, in the next sections, a theoretical overview of the techniques carried out on the data will be presented.

2.2 Auto-correlation and cross-correlation

Auto-correlation and cross-correlation are two techniques widely used in the analysis of time series. They allows to analyse the correlations within the same time series

(auto-correlation) or within different time series (cross-correlation) as a function of the separating interval between two observations (delay time or time lag).

Both techniques can be thought of as an evolution of the Pearson coefficient, which measures the correlation, i.e. the degree of linear link between two variables X_1 and X_2:

$$r = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}} \tag{2.7}$$

where $\bar{x}_j$ is the mean of the n-dimensional $X_j = (x_{j1}, \ldots, x_{jn})$ vector's components.

Let us introduce more details about both these techniques.

2.2.1 Auto-correlation

As already said, auto-correlation is a technique that exploits the Pearson correlation coefficient by applying it to the same time series at various shift levels, i.e. at various time lags τ. Thus, the classic Pearson correlation coefficient becomes

$$\gamma_\tau = \frac{\sum_{t=1}^{n-\tau}(x_t - \bar{x}_0)(x_{t+\tau} - \bar{x}_\tau)}{\sqrt{\sum_{t=1}^{n-\tau}(x_t - \bar{x}_0)^2 \sum_{t=1}^{n-\tau}(x_{t+\tau} - \bar{x}_\tau)^2}} \tag{2.8}$$

where $\bar{x}_0 = \frac{1}{n-\tau}\sum_{i=1}^{n-\tau} x_i$, while $\bar{x}_\tau = \frac{1}{n-\tau}\sum_{i=\tau+1}^{n-} x_i$.

This index is defined as the auto-correlation index at τ interval. Obviously, when the lag time is set to 0, this index measures the correlation of a series with itself and thus returns the value 1.

A plot that places the time lag values τ on the x-axis and the values of γ_τ on the y-axis is called auto-correlogram; it can be useful to suggest possible models to be applied to the time series under examination.

Of course, the auto-correlation function is symmetric around $\tau = 0$.

It is also possible to estimate the standard error for the auto-correlation coefficient γ_τ [11]:

$$SE(\gamma_\tau) = \sqrt{\frac{1}{m_0}\left(1 + \sum_{t=1}^{\tau-1}\gamma_t^2\right)} \tag{2.9}$$

where m_0 is the number of non-missing values in x.

2.2.2 Cross-correlation

Cross-correlation is the correlation between two different time series, whose component at time t are denoted by x_t, y_t. In this case, the observations of one series are

correlated with the observations of the other one at different time lags (sometimes called advances).

Hence, the cross-correlation coefficient is a function of time lag τ and it is defined as:

$$r_{xy|\tau} = \frac{\sum_{t=1}^{n-\tau}(x_t - \bar{x}_0)(y_{t+\tau} - \bar{y}_\tau)}{\sqrt{\sum_{t=1}^{n-\tau}(x_t - \bar{x}_0)^2 \sum_{t=1}^{n-\tau}(y_{t+\tau} - \bar{y}_t)^2}}, \qquad \text{for } \tau = 0, 1, 2, \ldots \tag{2.10}$$

$$r_{xy|\tau} = \frac{\sum_{t=1}^{n+\tau}(y_t - \bar{y}_0)(x_{t-\tau} - \bar{x}_\tau)}{\sqrt{\sum_{t=1}^{n+\tau}(x_{t-\tau} - \bar{x}_\tau)^2 \sum_{t=1}^{n+\tau}(y_t - \bar{y}_\tau)^2}}, \qquad \text{for } \tau = -1, -2, \ldots \tag{2.11}$$

where the same notation of the auto-correlation is used.

It is worth mentioning that two definitions of cross-correlations are given because one time series may be shifted towards or backwards with respect to the other one. In fact, in contrast with what happens in the auto-correlation case, the cross-correlation function is not symmetric about $\tau = 0$.

When the lag time is set to 0, the cross-correlation index coincides with Pearson's correlation coefficient r.

The values of $r_{xy|\tau}$ plotted as a function of the (discrete) time lag values τ is called cross-correlogram.

It is possible to estimate the standard error of the cross-correlation coefficient $r_{xy|\tau}$ [12]:

$$SE(r_{xy|\tau}) \cong \sqrt{\frac{1}{n - |\tau|}}. \tag{2.12}$$

2.3 Correlation Networks

A useful tool to study and visualise the relationships among variables is the so-called correlation network.

Let us consider a network $G(N, E)$, having N nodes representing general variables $X_1, \ldots, X_N$; it is a correlation network if an edge between two variables X_i and X_j (with $i \neq j$) occurs when the correlation between X_i and X_j is greater than a certain threshold, properly chosen depending on the considered case study.

This type of network depends not only on the choice of the threshold, but also on the choice of the measure of correlation used (Pearson's r, Spearman's ρ, Kendall's τ, ...).

In this dissertation, correlation networks have been built using, as variables $X_1, \ldots, X_N$, both the overall samples' composition and the taxa belonging to a

fixed sample. The former case results in a depiction of the relationships among the samples of the considered city; the latter is instead a sketch of the connections and dependencies among taxa inside one given sample. Also, the Pearson's correlation coefficient is used, and the edges has been equipped with a weight, corresponding to the given correlation of the variables (thus higher weights correspond to higher degrees of correlation).

Some of the network analyses used in this work were based on the concept of centrality measures which, according to the chosen definition, characterise in some way the nodes.

In this dissertation, two centrality measures have been used:

- Degree Centrality: it measures the centrality of a node v, by counting the number of links attached to it:

  $$C_D(v) = \deg(v)$$

  where $\deg(v)$ represents the number of links having the node v as an end; when the network is weighted, the degree of a node becomes:

  $$\deg(v) = \sum_{j=1}^{E_v} w_j$$

  where w_j is the weight of the j-th link, while E_v is the number of links having v as an end;

- Betweenness Centrality: it measures the centrality of a node v, starting from the concept of shortest path. The betweenness centrality is defined as follows:

  $$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

  where $\sigma_{st}(v)$ is the number of shortest path connecting nodes s and t and passing through v, while σ_{st} is the total number of shortest path between s and t; thus, each term of the sum is ≤ 1. For a weighted network the shortest path between two nodes s and t is defined as the path, going from s to t, that minimises the sum of the links' weights.

It is worth pointing out that a filtering procedure is often required before the construction of the correlation networks, when working on abundances of taxa. In fact, some taxa are barely present in most of the samples and this results in a lot of zero components. As already said in the previous sections, this may cause problems when dealing with correlations, since zeros correlates between them and the results are biased.

Sometimes, it is useful to visualise the degree distribution of the nodes of a network. In order to do this, there are at least two ways: by means of the degree

histogram or by means of the degree rank plot. The former is the histogram having the possible degrees in the x-axis and the number of nodes in the y-axis. The latter, instead, has the rank on the x-axis and the degree on the $y - axis$, and is the distribution of the degree by rank, in decreasing order of degree. This means that each node is ordered based on its degree and one point is drawn in the plot for each node at the corresponding pair of rank and degree.

Lastly, it is important to remember that, in this work, correlations are computed between variables transformed via centred log-ratio.

2.4 Fourier Spectral Analysis

The discrete Fourier spectral analysis (also called harmonic analysis) is a time series technique which allows to study the periodicities of a series.

The analysis is based on the Fourier transform, a mathematical operation which maps a function, Let us say $h(t)$, in a complex function $H(\omega)$:

$$H(\omega) = \int_{-\infty}^{+\infty} h(t)\, e^{i\omega t}\, dt \tag{2.13}$$

Equation 2.13 should be adapted to the case of a function for which N sampled values are given[4], with a certain sampling interval Δ:

$$h_j = h(t_j) = h(j\Delta) \quad j = 0, 1, \ldots, N-1$$

The purpose of the Fourier analysis is to estimate $H(\omega)$; of course, since N values of h are given, the procedure can only result in the estimation of N values of H, associated to the following ω values:

$$\omega_k = \frac{2\pi}{N\Delta} k \quad k = -\frac{N}{2}, \ldots, \frac{N}{2} \tag{2.14}$$

where k "discretises" the ω_k values (k may only assume values like $N/2, N/3$, $N/4$, ..., 0 and the negative ones).

Notice that the ω_k values lie in the range $[-\omega_c, \omega_c]$, where:

$$\omega_c = \frac{\pi}{\Delta} \tag{2.15}$$

is a critical value called Nyquist frequency[5].

One may notice that the sequence of ω values in 2.14 is composed of $N+1$ values, and not of N values. This is not in contrast with what was previously stated; in

[4]For the sake of simplicity, Let us assume that N is even; an odd value of N requires some minor adjustment in the consecutive steps.

[5]To be more precise, the Nyquist frequency is $\nu_c = \frac{1}{2\Delta}$, i.e. one-half of the sampling rate $\frac{1}{\Delta}$

fact, the extreme values of k are not independent because of a phenomenon related to the Nyquist frequency and called aliasing[6].

Now, equation 2.13 is approximated to its discrete version:

$$H_k = H(\omega_k) = \sum_{j=0}^{N-1} h_j \, e^{\frac{2\pi i k j}{N}} \tag{2.16}$$

As can be seen through equation 2.16, the discrete Fourier transform is a mapping of N numbers (the input h_j) to N numbers (the output H_k).

For what has been written so far, k can go from $-N/2$ to $N/2$; however, equation 2.16 is periodic in k with period N, i.e. $H_{-k} = H_{N-k}$. Thus k may be considered as an index ranging from 0 to $N-1$, just like the j index does. Using this notation:

$$\begin{aligned} 0 < k < N/2 \quad &\longmapsto \quad 0 < \omega < \omega_c \\ N/2 < k < N \quad &\longmapsto \quad -\omega_c < \omega < 0 \end{aligned}$$

while $k = N/2$ corresponds to both $\omega = \omega_c$ and $\omega = -\omega_c$, which in turn results in coincident components as a consequence of the discretisation.

Once the Fourier transform is performed, the so-called Fourier spectrum may be plotted by putting the (harmonic) k values on the x-axis (or its relative frequency ω_k or ν_k) and the amplitude of the k-th harmonic, i.e. $A_k = \sqrt{\mathfrak{Re}(H_k)^2 + \mathfrak{Im}(H_k)^2}$ (since H_k generally is a complex number), on the y-axis.

When the spectrum presents a peak, the corresponding harmonic indicates the existence of a possible periodicity in the time series. If confirmed, by means of a specific statistical test (which will be described below), the period will be:

$$T = \frac{N}{k} \, . \tag{2.17}$$

Fourier spectral analysis presents a relevant limit in the search for periodicities of a time series. In fact, the analysis of the spectrum does not take into consideration all the possible periodicities, but only those corresponding to the harmonics k.

Lastly, it is worth mentioning that, for the proper application of this technique, the following two conditions must be verified:

- the time series must be completely filled, i.e. it must not have missing data;
- the time series must be scanned at regular intervals.

In the work presented in this dissertation, the second constraint is not perfectly guaranteed, especially in some of the four cities considered. In other words, the time interval between two consecutive samples was not always exactly 14 days (see table 2.2 in section 2.1.1). However, the Fourier analysis was performed approximating each time gap to a value of 14 days.

[6]The aliasing is basically a translation of all the spectral amplitudes $H(\omega)$ of frequencies ω, such that $|\omega| > \omega_c$, from outside to inside the band $[-\omega_c, \omega_c]$.

2.4.1 Fast Fourier Transform

Before going on with the description of the Fisher test, used to check the significance of a periodic component of the Fourier spectrum, the Fast Fourier Transform (FFT) is presented.

One problem of the Fourier analysis presented in the previous section is the computational time required to compute the transform 2.16.

Let us consider a time series (h) composed of N values, so that:

$$H_k = \sum_{j=0}^{N-1} {W_N}^{kj} h_j \tag{2.18}$$

where $W_N = e^{\frac{2\pi i}{N}}$. This operation may be seen as an application of a matrix to a vector and thus requires N^2 operations to be performed (plus the operations required for the evaluation of the powers of W); thus the discrete Fourier transform is $O(N^2)$ computational expensive.

However, in 1965, Cooley e Tukey [13] proposed an algorithm to speed the computation up, based on the recursive "decomposition" of the discrete Fourier transform. This algorithm allows to reduce the computation time to $O(N\log_2 N)$.

Many other algorithms have been proposed but each reduces the computation time of the same amount of the Cooley-Tukey algorithm.

The Cooley-Tukey FFT algorithm is the one used in this work to perform the Fourier analysis.

2.4.2 Fisher test

Before describing the Fisher test, some useful formulae regarding the Fourier analysis are presented, following the same notation used in the previous sections:

$$A_k = \sqrt{\mathfrak{Re}(H_k)^2 + \mathfrak{Im}(H_k)^2} \tag{2.19}$$

$$\phi_k = \arctan \frac{\mathfrak{Re}(H_k)}{\mathfrak{Im}(H_k)} \tag{2.20}$$

$$s_k^2 = \frac{\mathfrak{Re}(H_k)^2 + \mathfrak{Im}(H_k)^2}{2} \tag{2.21}$$

where A_k, ϕ_k, s_k^2 are the amplitude, the phase and the variance of the k-th harmonic, respectively, and $\mathfrak{Re}(\cdot)$ and $\mathfrak{Im}(\cdot)$ are the real and imaginary part operators.

In particular, the definition of the variance of a harmonic equation 2.21 is important for the Fisher test.

In order to test the dominant periodic component of a Fisher spectrum, Fisher [14] developed a test: its purpose is precisely to evaluate whether the peak of a

spectrum represents a significant periodic component, or if it can be considered a random fluctuation.

The test statistic is:

$$\hat{g} = \frac{s^2_{max}}{s^2} \tag{2.22}$$

where s^2_{max} is the maximum variance, i.e. the maximum value of equation 2.21 (which, in turn, corresponds to the same k giving the maximum value of A_k), and s^2 is the variance of the entire time series. Since s^2 is given by the sum of the variances corresponding to all the harmonics of the spectrum, the test statistic may be re-written as:

$$\hat{g} = \frac{s^2_{max}}{\sum_{j=0}^{N-1} s^2_k} \ . \tag{2.23}$$

The critical value of the Fisher's test statistic $\hat{g}$ is given by

$$g_c \simeq 1 - \left(\frac{\alpha}{N}\right)^{\frac{1}{N-1}} \tag{2.24}$$

where α is the significance level (usually 0.01 or 0.05).

If the evaluated statistic $\hat{g}$ exceeds the critical value g, it is possible to conclude that the peak represents a significant periodic component; on the other hand, a statistic $\hat{g}$ less than the critical value g means that the observed peak is due to random fluctuations.

2.5 Biodiversity

Biodiversity is a term coined in 1988 by Edward O. Wilson [15] to indicate the variety of life forms present in a given community. Several authors have proposed indices to measure this variety; they are based on different aspects: the species richness (i.e., the number of species belonging to the considered community), the evenness (i.e., the homogeneity of abundance of the different species) and the taxonomic distinctness among the living forms present in the considered environment (that is the biological distance between the different species). From what has been said, it clearly emerges that biodiversity increases as each of these factors increases.

The various biodiversity indices proposed in the literature can take into account all these aspects or just some of them.

Before going on with the definitions of some of the most known indices, it is useful to distinguish two different frameworks when dealing with biodiversity:

- The first one is the so-called α-biodiversity, which is a measure of diversity applied separately to each sample, i.e. it is a within sample (or intra-sample)

biodiversity measure. However, it allows the comparison of the values of the considered index obtained from samples belonging to environments with similar features;

- The second one is the so-called β-biodiversity, which measures the difference in biodiversity between different samples, i.e. it is a inter-samples biodiversity measure.

In the next sections, these two types of biodiversity indices are presented.

2.5.1 α - biodiversity

As already said, α-biodiversity measures the diversity within a certain sample. In this dissertation, various indices referring to α-biodiversity were used. They are all presented below, along with their definitions and their main characteristics.

For the sake of clarity, the notation used in this section (unless otherwise indicated), is listed:

S is the number of observed species;

N is the number of observed individuals;

n_i is the number of observed individuals of the i-th species;

p_i is the relative observed abundance of the i-th species, i.e. $p_i = n_i/N$.

The Shannon index

The Shannon-Wiener index (or simply Shannon index), named after the two scholars in the information field who came to describe (separately) this index [16], assumes that individuals are randomly sampled from an "indefinitely large" population, and that all the species in the community have an equal probability of being represented in the sample.

This index is usually indicated with H and is defined a follows:

$$H = -\sum_{i=1}^{S} p_i \log p_i \tag{2.25}$$

where the notations used are those already described.

As suggested by Shannon, there are no particular constraints for the choice of the base of the logarithm but, obviously, a comparison between two values of the indices is only possible if the same base has been used. In the present work the base e, thus the natural logarithm, has been used.

The index takes into account both the species richness and relative abundances (and therefore the evenness), summarising the information in a single diversity value. H reaches its minimum value (i.e. 0) when all individuals belong to a single species.

On the other hand, its maximum is reached when the individuals are uniformly distributed among all the species present in the sample. For the minimum H is:

$$H_{min} = -\sum_{i=1}^{S} p_i \log p_i = -1 \log 1 = 0.$$

while for the maximum:

$$n_1 = n_2 = \cdots = n_S = n$$
$$p_i = \frac{n_i}{N} = \frac{n}{S \cdot n} = \frac{1}{S}$$
$$H_{max} = -\sum_{i=1}^{S} p_i \log p_i = -\sum_{i=1}^{S} \frac{1}{S} log \frac{1}{S} = -S\frac{1}{S}(-\log S) = \log S.$$

Finally, it is worth pointing out that H index is strictly related to the weighted geometric mean[7] of the relative abundances p_i, performed using the same relative abundances p_i as weights; in fact,

$$H = -\sum_{i=1}^{S} p_i \log p_i = -\sum_{i=1}^{S} \log {p_i}^{p_i} = -\log \prod_{i=1}^{S} {p_i}^{p_i} \tag{2.26}$$

Now, remembering that $\sum_{i=1}^{S} pi = 1$ by definition, and that p_i are used as weights, the index H can be thought of as the opposite of the logarithm of the weighted geometric mean of the relative abundances p_i.

Lastly, a list of useful properties of the Shannon-Wiener index are presented:

1. it is continuous w.r.t to p_i;
2. if all $p_i = 1/S \ \forall i$, then H is monotonically increasing w.r.t. S;
3. it does not change if the p_i values are re-ordered.

Point 3 has an important consequence: if the values of the Shannon index of two samples are equal, then the only reasonable conclusion that one may draw is that they have the same biodiversity. If the two samples are also composed of the same S species, then the information encoded by the index is that the values of relative abundance of the S species are the same in both the samples, but this does not necessarily imply that same values of relative abundance correspond to the same species in the two different samples.

[7]The weighted geometric mean (g_w) of an S-dimensional vector $\mathbf{x}$ is $g_w(x) = \left(\prod_{i=1}^{S} {x_i}^{w_i}\right)^{\frac{1}{\sum_{j=1}^{S} w_i}}$ where w_i are the weights.

The Pielou index

Pielou [17] proposed a normalisation of the Shannon-Wiener index so to make it more comparable even when it comes from different contexts. By exploiting the property of the Shannon-Wiener index for which its theoretical maximum equals the logarithm of the number of species, the Pielou index is defined as the ratio between the Shannon-Wiener index and the logarithm of S:

$$J = \frac{H}{H_{max}} = \frac{-\sum_{i=1}^{S} p_i \log p_i}{\log S}. \tag{2.27}$$

Hence, the Pielou index lives in the range between 0 and 1. Being a derivation of the Shannon-Wiener index, this index also depends on both species richness and evenness.

The Gini-Simpson index

In 1949 Simpson [18] developed an index capable of measuring biodiversity by calculating the probability that two individuals randomly taken from a sample belong to the same species. His index is therefore the following:

$$\lambda = \sum_{i=1}^{S} p_i^2. \tag{2.28}$$

However, this index is impractical as it is inversely proportional to the observed biodiversity. In fact, it reaches the its maximum, equal to 1, when all the individuals belong to the same species. The index does not have a minimum but it has a lower limit of zero. Because of its nature, it is known as (Simpson) dominance index (rather than diversity index).

In his paper, Simpson himself proposed to consider as direct biodiversity index the reciprocal of the dominance index. By doing so, he has built an index which increases as biodiversity increases, but which has the disadvantage of have a minimum value of 1 and, especially, the absence of an upper limit:

$$\frac{1}{\lambda} = \frac{1}{\sum_{i=1}^{S} p_i^2}. \tag{2.29}$$

This index is known as the Simpson inverse dominance index.

Some time later, a little-known paper by Gini (1912) [19] was rediscovered: in this paper he simply proposed as an index of diversity, albeit in a different and more general context, the following:

$$GS = 1 - \sum_{i=1}^{S} p_i^2 \tag{2.30}$$

which in fact can be thought as the complementary (to 1) of the Simpson dominance index:

$$GS = 1 - \lambda. \tag{2.31}$$

It has a minimum, equal to 0, corresponding to the minimum possible biodiversity (i.e. in the case of only one species belonging to the sample) and it tends to 1 when biodiversity is maximum. This version of the index, known as the Gini-Simpson index (GS), was used in this dissertation.

All the presented versions of the Simpson index take into account both species richness and evenness.

The Hill numbers

Another type of indices has been proposed considering the reciprocal of the weighted generalised mean of the relative abundances of the species with exponent $q-1$ (indicated as $1/M_{q-1}$), where the weights are the relative abundances p_i; these indices are usually indicated with ${}^qD$, they depend on the choice of q and are called Hill numbers of order q [20]:

$${}^qD = \frac{1}{M_{q-1}} = \frac{1}{\sqrt[q-1]{\sum_{i=1}^{S} p_i p_i^{q-1}}} = \left(\sum_{i=1}^{S} p_i^q\right)^{\frac{1}{1-q}}. \tag{2.32}$$

It is immediate to prove that the Hill number of order 2 (${}^2D$) coincides with the Simpson inverse dominance index 2.29:

$${}^2D = \left(\sum_{i=1}^{S} p_i^2\right)^{\frac{1}{1-2}} = \frac{1}{\sum_{i=1}^{S} p_i^2} = \frac{1}{\lambda}. \tag{2.33}$$

It is much less immediate to verify that the Hill number of order 1 (${}^1D$) tends (for q approaching 1) to the exponential of the Shannon-Wiener index computed with the natural logarithm:

$${}^1D = \lim_{q\to 1}\left(\sum_{i=1}^{S} p_i^q\right)^{\frac{1}{1-q}} = \frac{1}{\prod_{i=1}^{S} p_i{}^{p_i}} = \exp\left(-\sum_{i=1}^{S} p_i \ln p_i\right). \tag{2.34}$$

The mathematical proof is given in appendix C.

The Chao index

Starting from the consideration that estimating the number of species present in a community is a difficult task due to the fact that less abundant species have a low probability of being part of the sample, Chao [21] proposed a non-parametric method to estimate the number of species actually present in a given community,

starting from the number of species observed. Her approach starts from the abundances of the rarest species, precisely from the number of singletons and doubletons, defined as the number of species with absolute abundance equal to 1 and equal to 2, respectively.

Denoting with F_1 the number of singletons, with F_2 the number of doubletons and with S_{obs} the number of species observed in the sample, the unbiased version of the Chao index[8] results:

$$S_{Chao} = S_{obs} + \frac{F_1(F_1 - 1)}{2(F_2 + 1)}. \tag{2.35}$$

where the second term of the right hand side estimates the number of "unsampled" species, based on the number of low abundance species.

As can be deduced the Chao index is just a measure of the species richness.

The Taxonomic Distinctness indices

Many biodiversity indices (including all those described so far) are based on species richness and evenness, completely neglecting taxonomic diversity. To capture this aspect of biodiversity as well, Clarke and Warwick [22] proposed two new indices that included, in addition to species richness and relative abundances, also information on taxonomic relationships between species.

The first index, called Taxonomic distinctness index and denoted with Δ^* can be thought as a variant of the Gini-Simpson index, to which taxonomic distances between species have been added. It is the following:

$$\Delta^* = \frac{\sum\sum_{i<j} \omega_{i,j} n_i, n_j}{\sum\sum_{i<j} n_i, n_j} \tag{2.36}$$

where $\omega_{i,j}$ is defined as a distinctness weight, measuring the length of the path linking each pair of species i, j along the taxonomic classification tree (considered up to the common ancestor, see Figure 2.3 for a visual explanation), and n_i was already defined at the beginning of this section.

This index can be interpreted as the expected distance of the path along the taxonomic tree between any two individuals (belonging to different species) randomly chosen from the sample. It is also interesting to note that Clarke and Warwick have shown that this index is invariant with respect to changes in scale, so that instead of absolute abundances it is possible to consider other variables of ecological interest, such as biomass.

[8]The very first definition given by Chao was $S_{Chao} = S_{obs} + \frac{{F_1}^2}{2 \cdot F_2}$ but it is not defined for $F_2 = 0$ and it was proved to be biased.

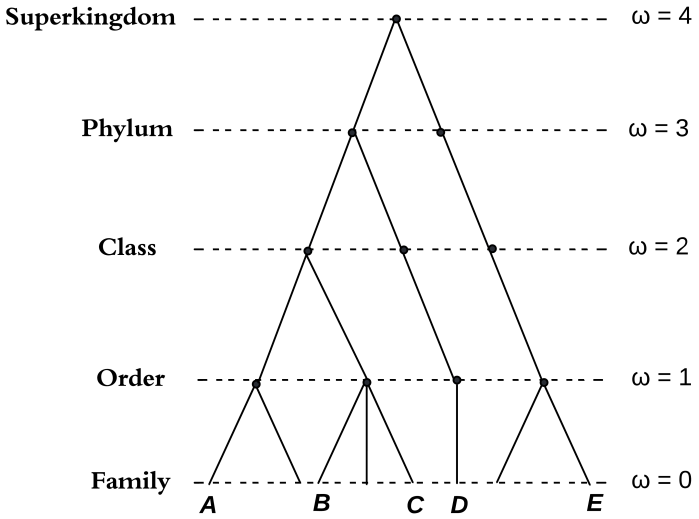


Figure 2.3: Schematic example of how the weights for the Taxonomic Distinctness indices are computed. For each pair of families, the weight is fixed on the basis of the common ancestor (e.g. between A and B $\omega = 2$, between B and C $\omega = 1$, between A and D $\omega = 3$, and between A and E $\omega = 4$)

The second index, called Binary taxonomic distinctness index and denoted by Δ^+) can be considered as a special case of the previous one, which is obtained when, for each species, only the binary information of presence/absence is available:

$$\Delta^+ = \frac{\sum \sum_{i<j} \omega_{i,j}}{S(S-1)/2}. \tag{2.37}$$

Therefore, when the data are expressed in terms of presence/absence, the Δ^* index becomes Δ^+, which represents the average distance between any two species within the taxonomic tree.

It is worth mentioning that Δ^+ is the average value of the taxonomic weights.

Since the evaluation of the indices is computationally expensive and time consuming, it has been performed starting from the family level instead of the species level. This choice does not change the original idea behind these indices of taking into account the taxonomic hierarchical structure.

2.5.2 β - biodiversity

β-diversity takes into account the differences between samples (two or more), both temporal (same spatial unit, different times) and spatial (same time, different spatial units). In this context, various indices have been proposed that make it possible to quantify, in different situations, the similarity or dissimilarity of two samples.

In this dissertation, the Bray-Curtis dissimilarity index was chosen, and its results were, in turn, used in a so-called ordination technique, the Principal Coordinate

Analysis (PCoA).

2.5.3 The Bray-Curtis dissimilarity index

Let us consider two samples denoted by k and l; the following notations will be used:

S_k and S_l are the number of species (with abundances $\neq 0$) observed in the k and l site, respectively;

N_k and N_l are the number of individuals observed in the k and l sample, respectively;

$n_{i,k}$ and $n_{i,l}$ are the number of individuals observed for the i-th species in the k and l sample, respectively, so that the same index i indicates the same species in the two samples; thus $i = 1, 2, \ldots, S_{tot}$ (where S_{tot} is the number of species present in at least one sample).

The Bray-Curtis dissimilarity index (denoted by BC) [23] may be written as follows:

$$BC = \frac{\sum_{i=1}^{S_{tot}} |n_{i,k} - n_{i,l}|}{\sum_{i=1}^{S_{tot}} (n_{i,k} + n_{i,l})} \ . \tag{2.38}$$

The BC index ranges in the interval $[0, 1]$; the value 0 (minimum dissimilarity) occurs when the abundances in the two samples are identical for each species:

$$n_{i,k} = n_{i,l} \qquad \forall i = 1, 2, \ldots, S_{tot}$$

and the value 1 (maximum dissimilarity) occurs when all the species represented in one sample are absent in the other one and vice versa:

$$n_{i,k} \neq 0 \iff n_{i,l} = 0 \quad \text{and} \quad n_{i,k} = 0 \iff n_{i,l} \neq 0 \qquad \forall i = 1, 2, \ldots, S_{tot} \ .$$

Another way to represent the Bray-Curtis index is the following:

$$BC = 1 - \frac{2 \sum_{i=1}^{S_{tot}} \min\{n_{i,k}, n_{i,l}\}}{N_k + N_l}. \tag{2.39}$$

In other words, in this formulation the term $\sum_{i=1}^{S_{tot}} \min\{n_{i,k}, n_{i,l}\}$ is the sum of the minimum abundance (of each species) between the two samples.

Definitions 2.38 and 2.39 are equivalent.

2.5.4 Ordination techniques

The term ordination refers to a set of multivariate techniques used for dimensionality reduction on a dataset, in order to make it viewable in a two- or three-dimensional space. This procedure allows to highlight and find patterns or clusters that are not easily identifiable directly from the data or from the statistics that can be derived from them.

In this work, in particular, a technique known as Principal Coordinate Analysis (PCoA) was applied. Before proceeding with the discussion about PCoA, an overview of the Principal Component Analysis (PCA) is proposed, since PCoA is an "adjustment" of PCA under certain conditions.

Principal Component Analysis

The Principal Component Analysis is an ordination technique which operates exclusively a rigid rotation of the axes of the multidimensional space of the data, in order to orient them in such a way as to maximise the dispersion of the data. This allows a set of data to be represented more effectively even in a reduced number of dimensions, i.e. in a system of orthogonal axes (called Principal Components) defined as linear combinations of the original variables.

Let us consider the data matrix $\mathbf{X} \in \mathbf{M}_{n\times p}$, where n is the number of the samples and p is the number of the (original) components (i.e. variables). The elements $x_{i,j}$ of the matrix $\mathbf{X}$ are then transformed in their difference from the arithmetic mean of the respective column. Thus, the resulting matrix, $\mathbf{Y} \in \mathbf{M}_{n\times p}$, is defined by:

$$y_{i,j} = x_{i,j} - \frac{\sum_{i=1}^{n} x_{i,j}}{n}$$

where, of course, $y_{i,j}$ is the element of the i-th row and j-th column of $\mathbf{Y}$. The next step is to obtain the covariance matrix as follows:

$$\mathbf{\Sigma} = \frac{1}{n}\mathbf{Y}^T\mathbf{Y} \ .$$

and from it, its p eigenvalues λ_k and eigenvectors u_k $(k = 1, 2, \ldots, p)$. The k eigenvectors may be arranged as columns of a matrix, denoted by $\mathbf{U} \in \mathbf{M}_{p\times p}$.

These eigenvectors represents the new directions of the new system of axes.

However, in order to reduce the dimensionality, it is necessary to take only m eigenvectors, with $m < p$; in particular, the ones associated to the greatest m eigenvalues. The choice of m depends on different factors; usually, m is chosen to be equal to 2 or 3 so that the results of the PCA may plotted and visualised in search of clusters.

The matrix of eigenvector modified by taking only m of them is denoted by $\mathbf{U}' \in \mathbf{M}_{p\times m}$ and thus, its element $u_{j,k}$ represents the j-th component of the k-th eigenvector (with $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, m$).

Now, Let us define $\mathbf{F} \in \mathbf{M}_{n\times m}$ as the matrix which gives the coordinates of the original data points in the new system of axes (i.e. to the Principal Components); its elements $f_{i,j}$ are calculated by multiplying the matrix $\mathbf{Y}$ by the matrix of the taken eigenvectors $\mathbf{U}'$, so that:

$$\mathbf{F} = \mathbf{Y} \times \mathbf{U}' \ .$$

Thus, $\mathbf{F}$ works as a map from the input data to the low-dimensional space defined by the m eigenvectors.

The quality of the representation obtained from the PCA may be evaluated on the basis of the eigenvalues corresponding to the considered eigenvectors. In fact, the percentage of variance explained by the a certain principal component is equal to the ratio between the the associated eigenvalue and the trace of the matrix $\mathbf{\Sigma}$, which, in turn, is equal to the some of all the eigenvalues[9].

For a proper application of the Principal Component Analysis, it is required to deal with quantitative variables, whose distribution is normal and that the data matrix does not contain an excessive number of zeros.

Principal Coordinate Analysis

Datasets do not always possess the properties necessary for a proper application of Principal Component Analysis. An example is given by the case considered in this dissertation, that is a list of species observed in a certain number of samples: the counts (abundances) are not necessarily distributed according to a normal distribution. Moreover, the number of zeros in the data matrix, which correspond to the absence of species from a sample, is very often even higher than the number of non-null values.

Thus, the idea is to consider another ordination technique which allows to deal with this type of data; an example is given by the Principal Coordinate Analysis (PCoA) [24].

The Principal Coordinate Analysis is based on a metric, which allows to evaluate the distances among the data. In the working case study, this metric is usually called measure of similarity (or dissimilarity), because it computes the distances between two samples by evaluating a sort of degree of similarity (or dissimilarity). In the literature there are numerous measures of similarity and of dissimilarities, but in this dissertation, the Bray - Curtis dissimilarity matrix is the chosen metric for PCoA.

PCoA starts from the matrix $\mathbf{D} \in \mathbf{M}_{n\times n}$ of the dissimilarities between the n samples, which is transformed into the matrix $\mathbf{\Delta} \in \mathbf{M}_{n\times n}$:

$$\mathbf{\Delta} = -\frac{1}{2}\mathbf{D} \ .$$

The matrix $\mathbf{\Delta}$ is, in turn, transformed into the matrix $\mathbf{C} \in \mathbf{M}_{n\times n}$ by centring $\mathbf{\Delta}$ in such a way that the origin of the new axes is located in the centroid of the samples:

$$c_{i,j} = \delta_{i,j} - \frac{1}{n}\sum_{h=1}^{n}\delta_{i,h} - \frac{1}{n}\sum_{k=1}^{n}\delta_{k,j} - \frac{1}{n^2}\sum_{h=1}^{n}\sum_{k=1}^{n}\delta_{h,k}, \tag{2.40}$$

[9]The eigenvalues of the $p \times p$ matrix Σ are p, of course. In this context, this means that all those eigenvalues with multiplicity κ is considered as κ eigenvalues.

where $c_{i,j}, \delta_{i,j}$ are the elements of the matrices $\mathbf{C}, \mathbf{\Delta}$, respectively.

It is worth pointing out that the second and third terms of the expression of $c_{i,j}$ represent the row and column means of the $\mathbf{\Delta}$ matrix, while the last term represents the total mean of the same matrix.

Then, the eigenvalues of $\mathbf{C}$ are computed, and the greatest m are taken (usually $m = 2$ or $m = 3$, which correspond to the dimensions to which the original dataset is intended to be reduced). The m corresponding eigenvectors of the matrix C are then arranged in the same way as in the PCA, obtaining the matrix $\mathbf{U'} \in \mathbf{M}_{n \times m}$ whose elements are $u_{i,j}$ with $i = 1, \ldots, n, \quad j = 1, \ldots, m$.

The Principal Coordinate $f_{i,j}$ of the samples are obtained by multiplying the eigenvectors by the square root of the corresponding eigenvalue:

$$f_{i,j} = u_{i,j} \sqrt{\lambda_j}.$$

As for the PCA, also in this case the quality of the ordination obtained for each Principal Coordinate can be evaluated on the basis of the ratio between the corresponding eigenvalue and the sum of all the eigenvalues.

2.6 Tests for comparing time series

In some cases, comparisons among some features of different time series, or even among subsets of a given time series, were necessary.

Depending on the feature to be compared, different tests statistics have been used in this dissertation, among which there are the Augmented Dickey-Fuller test (for evaluating the stationarity of a time series), the Brown-Forsythe test (for comparing the variances of time series), and the one-way ANOVA approach, followed by the Tukey's Honestly Significant Difference (HSD) test[10] (for comparing the mean of different time series).

All of these tests are described in the following sections.

2.6.1 The Augmented Dickey-Fuller test for time series analysis

The Augmented Dickey-Fuller (ADF) test is a widely used test to verify the stationarity of a time series. It represents an extension of the original Dickey-Fuller (DF) test [25], which had similar purposes. Therefore, before describing the ADF test, its original version is presented.

[10] To be more precise, the one-way ANOVA approach is followed by the Tukey's HSD test only if the former gives significant results.

The original Dickey-Fuller test

The model starts from a simple autoregressive model (AR(1) model) which can be written in the following form:

$$y_t = \rho y_{t-1} + e_t \tag{2.41}$$

where y_t is the variable studied in the time series, t is the time instance, e_t is the error term, which is supposed to have the characteristics of a white noise, with mean 0 and standard deviation σ.

The ρ parameter is such that if $|\rho| = 1$ then the time-series is non-stationary. This is referred to as a unit root situation.

If $|\rho| < 1$ the time series converges (for $t \to \infty$) to stationarity. On the other hand, if $|\rho| > 1$, the series diverges as t increases.

Subtracting y_{t-1} from both sides of the equation 2.41, one gets:

$$y_t - y_{t-1} = \rho y_{t-1} - y_{t-1} + e_t \tag{2.42}$$

which can be re-written as:

$$\Delta y_t = (\rho - 1)\, y_{t-1} + e_t \ . \tag{2.43}$$

Now, by setting $\rho - 1 = \delta$ the model may be written in the following form:

$$\Delta y_t = \delta y_{t-1} + e_t. \tag{2.44}$$

Hence, testing the null hypothesis of unit root is equivalent to setting the null hypothesis as:

$$H_0 : \ \delta = 0 \quad \Longrightarrow \quad H_0 : \ \rho = 1 \ .$$

Thus, the null hypothesis is the unit root of 2.41, while the alternative hypothesis is:

$$H_1 : \ \delta \neq 0 \quad \Longrightarrow \quad H_1 : \ \rho \neq 1 \ .$$

To be more precise, there are two other versions of the model to be tested via the Dickey-Fuller test:

1. Test for unit root with constant:

$$\Delta y_t = b_0 + \delta y_{t-1} + e_t$$

2. Test for unit root with constant and deterministic trend:

$$\Delta y_t = b_0 + b_1 t + \delta y_{t-1} + e_t$$

The test statistic is the same for all models and is given by:

$$DF = \frac{\hat{\delta}}{SE(\hat{\delta})} \ . \tag{2.45}$$

However, the critical values for these tests are different, and depend not only on the model but also on the sample size.

It is worth pointing out that because of the definition of the test statistic, the hypothesis of stationarity is basically equivalent to the alternative hypothesis when DF is negative.

The theoretical distribution of DF under the null hypothesis has not yet been identified, but Dickey and Fuller provided the tables of the critical values, through the application of Monte Carlo methods.

The Augmented Dickey-Fuller test

Dickey and Fuller also proposed an extension of the test (called Augmented Dickey-Fuller test, ADF [26]) that evaluates the same null hypothesis of unit root, but excluding the structural effects related to auto-correlations with lag greater than 1.

The model becomes the following:

$$\Delta y_t = b_0 + b_1 t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + e_t, \tag{2.46}$$

where b_0 is a constant, b_1 is the coefficient of a time trend, and p is the order of the autoregressive process. As for the (simple) Dickey-Fuller test, there are three version of the test depending on the presence or absence of the b_0 and b_1 terms.

In this dissertation, the model without the trend component is used:

$$\Delta y_t = b_0 + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + e_t. \tag{2.47}$$

The parameter to be tested is just γ, and that is true for all the different versions; also, the null hypothesis is always the same:

$$H_0 : \ \gamma = 0$$

which, again, corresponds to a unit root process.

Also in the ADF case, as in the DF, the test statistic is the same for all the models and is given by:

$$DF' = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \ . \tag{2.48}$$

Also in this case, the theoretical distribution of DF'under the null hypothesis has not yet been identified, but Dickey and Fuller provided the tables of the critical values, through the application of Monte Carlo methods, for all the versions, and for various sample sizes.

2.6.2 Brown-Forsythe test

In addition to the evaluation of the stationarity of the time series, sometimes the aim is to compare the variability of a time series with respect to the variability of the others. In order to do so, the Levene's test can be used. In this dissertation, the Brown-Forsythe's version of this test is used.

Levene's test, in its original version, can be described as follows: Let us consider k groups on which a variable X is measured, so that

$$x_{i,j} = \mu_j + \epsilon_{i,j}$$

are the i-th observation ($i = 1, 2, \ldots, n_j$) of the j-th group ($j = 1, 2, \ldots, p$), and where μ_j is the unknown mean of the j-group, while $\epsilon_{i,j}$ is an error term, assumed with 0 mean and possibly different variances. Let us also define $n = \sum_{j=1}^{p} n_j$ the total number of data.

The purpose of Levene's test [27] is to verify that the variances of the p groups are not significantly different (in other words, Levene's test wants to verify the homoskedasticity of the distributions of the populations from which the groups come). In order to do this, Levene started from the average absolute deviation:

$$\bar{z}_j = \frac{\sum_{i=1}^{n_j} z_{i,j}}{n_j} = \frac{\sum_{i=1}^{n_j} |x_{i,j} - \bar{x}_j|}{n_j}$$

and from:

$$\bar{\bar{z}} = \frac{\sum_{j=1}^{p} \sum_{i=1}^{n_j} z_{i,j}}{n},$$

which is the general mean of all the $z_{i,j}$, i.e. considering all data of all groups.

The test statistics is:

$$W_0 = \frac{\dfrac{\sum_{j=1}^{p} n_j (\bar{z}_j - \bar{\bar{z}})^2}{p-1}}{\dfrac{\sum_{j=1}^{p} \sum_{i=1}^{n_j} (z_{i,j} - \bar{z}_j)^2}{n-p}} ,$$

and has many similarities with the one-way ANOVA (which will be described in the following section). This test statistics follows a Fisher-Snedecor F-distribution with $(p-1,\, n-p)$ degrees of freedom, under the null hypothesis, which is:

$$H_0 : \ \sigma_1^2 = \sigma_2^2 = \ \ldots \ = \sigma_p^2 \ .$$

Brown and Forsythe [28] showed that if the absolute deviations $z_{i,j}$ are evaluated with respect to the median $\tilde{x}_j$, instead of the mean $\bar{x}_j$, i.e.

$$z_{i,j} = |x_{i,j} - \tilde{x}_j| \ ,$$

the homoskedasticity test was more robust than the original Levene's test.

2.6.3 One-way ANOVA

The purpose of the Analysis of Variance (ANOVA) is to allow a simultaneous comparison between the means of more than 2 samples. In fact, comparing p groups by means of t-tests, each at a significance level α, results in a probability equal to $(1-\alpha)^c$ of not committing any type I error (i.e. false positive error), where c is the number of tests. Therefore the risk of committing at least one type I error (which is called family-wise error rate, FWER) will be:

$$\text{FWER} = 1 - (1-\alpha)^c \ .$$

One way to overcome this problem is to use the Bonferroni correction, that is based on adjusting the significance level α dividing it by the number of comparisons to perform, c, so that each of the c t-tests will be performed at a significance level of α/c. This results in a risk of committing at least one type I error equal to:

$$\text{FWER} = 1 - \left(1 - \frac{\alpha}{c}\right)^c \leq \alpha \ .$$

The Bonferroni correction is very conservative and thus increases the probability to find false negatives, i.e. it reduces the statistical power.

Another way to counteract the problem of multiple testing's significance level is the Fisher's ANOVA approach.

Let us consider p groups, each one providing an average, denoted as $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$. The null hypothesis is that there are no differences between the averages of the populations from which the individual samples are extracted, that is:

$$H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_p \ ,$$

while the alternative hypothesis will be that there is at least one mean significantly different from the others, i.e. that not all means coincide:

$$H_1 : \ \exists\, \mu_l, \mu_m : \ \mu_l \neq \mu_m \, , \quad l, m = 1, 2, \ldots, p \ .$$

In the one-way ANOVA design, a set of n data-points is divided into p groups; in the considered case there are n time measurements divided into three periods ($p = 3$).

The following notation will be used:

$x_{i,j}$ are the single observations; so $x_{i,j}$ is the value of the variable X, detected in the i-th data-point of the j-th group;

i is the position indicator within the group, $i = 1, \ldots, n_j$;

j is the group indicator $j = 1, \ldots, p$;

n_j is the size of the j-th group;

$n = \sum_{j=1}^{p} n_j$ is the total number of data in the experiment;

$\bar{x}_j$ is the mean of the j-th group;
$\bar{\bar{x}}$ is the total mean of all data.

One of the simplest Analysis of Variance model is the one with only one "experimental factor", Let us call it α_j (this way the test is called one-way ANOVA):

$$x_{i,j} = \mu + \alpha_j + \epsilon_{i,j} \ . \tag{2.49}$$

In other words, this model states that the value $x_{i,j}$ of the i-th data-point of the j-th group depends on:

- μ: an average effect common to all groups;
- α_j: a characteristic effect of the j-th group;
- $\epsilon_{i,j}$: a deviation due to random fluctuations; this difference is called residual and it is equivalent to the difference between the observed value and the expected value based on the model. It could indicate the effect of unknown factors or anyway not kept under control in the experiment.

Moving from the theoretical model to the experimental data, Equation 2.49 becomes:

$$x_{i,j} = \bar{\bar{x}} + (\bar{x}_j - \bar{\bar{x}}) + \epsilon_{i,j} \ . \tag{2.50}$$

The Analysis of Variance is based on the decomposition of Sum of Squares. Three different source of variations, expressed as Sum of Squares, need to be considered:

1. a total Sum of Squares, i.e. the Sum of Squares calculated from the totality of the data with respect to the total mean:

$$SS_{tot} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{i,j} - \bar{\bar{x}})^2 \ ;$$

   this Sum of Squares has $n - 1$ degrees of freedom; in fact, n is the total number of data-points and 1 is the parameter (the total mean) estimated from the data;

2. a Sum of Squares between groups, i.e. the Sum of Squares of the group means from the total mean, weighted by the group size:

$$SS_{between} = \sum_{j=1}^{p} n_j (\bar{x}_j - \bar{\bar{x}})^2 \ ;$$

   this Sum of Squares has $p-1$ degrees of freedom; in fact, p is the total number of considered data (the p means of the groups), while 1 is the parameter (the total mean) estimated from the data;

3. a Sum of Squares within groups, i.e. the Sum of Squares of each data, with respect to the mean of the group to which it belongs:

$$SS_{within} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j)^2 \ ;$$

this Sum of Squares has $n-p$ degrees of freedom; in fact, n is the total number of the data considered and p are the parameters (the means of each treatment group) estimated from the data.

The following decomposition of the Sum of Squares holds:

$$SS_{tot} = SS_{between} + SS_{within} \ ,$$

and the same decomposition may be applied to the corresponding degrees of freedom:

$$(n-1) = (p-1) + (n-p) \ .$$

Starting from the different Sum of Squares, the corresponding variances can be obtained by dividing them by the associated degrees of freedom, and in particular:

- a variance between groups, also called explained variance (since it represents the portion of variability that is precisely "explained" by the belonging to a group):

$$Var_{between} = \frac{SS_{between}}{p-1};$$

- a variance within groups, also called unexplained variance or residual variance (since it represents the portion of variability due to random fluctuations):

$$Var_{within} = \frac{SS_{within}}{n-p}.$$

It is now possible to proceed with the statistical test, called Fisher's F test. The basic idea of the test is the following: if the null hypothesis is true (i.e. if all the means of the groups are equal) then the variance between groups and the variance within groups are estimates of the same true variance, and should therefore assume the same value (so their ratio should be close to 1).

On the contrary, if the alternative hypothesis is true, then the variance between the groups should be greater than that within the groups (and therefore the ratio between the two variances should be greater than 1). In other words, the explained variance should be higher than (and not comparable to) the residual variance, that is, the unexplained one.

The test-statistic is therefore given by the ratio between these two variances:

$$F = \frac{Var_{between}}{Var_{within}} \ .$$

This test statistic is distributed as a Fisher-Snedecor distribution with $(p-1;\ n-p)$ degrees of freedom, under the null hypothesis.

2.6.4 Tukey's HSD (post-hoc) test

Once the Analysis of Variance has given a significant result, and therefore it has been established that not all the true means of the groups are equal, it is legitimate to wonder between which pairs of groups there is a significant difference.

The test used in this dissertation for evaluate the significance of this difference was the HSD (Honestly Significant Difference) developed by Tukey.

This test applies to all pairs of groups; in other words, each mean of a group is compared with the means of all the other groups.

The Tukey test is basically a correction of the Student's t-test, aimed at ensuring the maintenance of the significance level (usually $\alpha = 0.05$) established for the entire family of tests.

Tukey's test is structured as follows: two means $(\bar{x}_l, \bar{x}_m)$ are significantly different at an α family-wise significance level if their difference in absolute value equals or exceeds the so-called minimum critical difference (MSD), i.e. if:

$$|\bar{x}_l - \bar{x}_m| \ \geq \ MSD_{l,m} \ .$$

The $MSD_{l,m}$ depends on groups l and m and is defined as follows:

$$MSD_{l,m} = Q_{\alpha,[p,n-p]} \cdot SE_{(\bar{x}_l - \bar{x}_m)}$$

where $Q_{\alpha,[p,n-p]}$ is the critical value of the Studentised range distribution at an α significance level with p and $n-p$ degrees of freedom, while the Standard Error $SE_{(\bar{x}_l - \bar{x}_m)}$ is defined as

$$SE_{(\bar{x}_l - \bar{x}_m)} = \sqrt{\frac{Var_{within}(\frac{1}{n_l} + \frac{1}{n_m})}{2}}.$$

Chapter 3

Results and Discussion

In this chapter, the results deriving from the different analyses proposed in this dissertation are presented by means of different plots and tables.

First of all, we started the analyses by evaluating the between samples diversity, looking for (dis)similarities among the four cities. Also, the effects of seasonality is inspected, by means of Principal Coordinate Analysis. Then, a temporal description of the time series was given, by analysing the periodicities and studying the correlations among taxa.

Ecological techniques were then taken into account to describe the (α) biodiversity of the samples, by means of nine indices, based on different features: species richness, evenness, and taxonomic distance. By looking at the temporal behaviour of the biodiversity in the four cities, we noticed an abrupt decrease in both Rome and Budapest in the Summer of 2020. This collapse of biodiversity was further investigated.

Finally, a correlation network analysis was performed, so that a depiction of the relationships among species was provided.

All non-reported outcomes and images emerged from this work and cited in the text are collected in the appendix D (or in the GitHub repository [29].

3.1 Between samples diversity

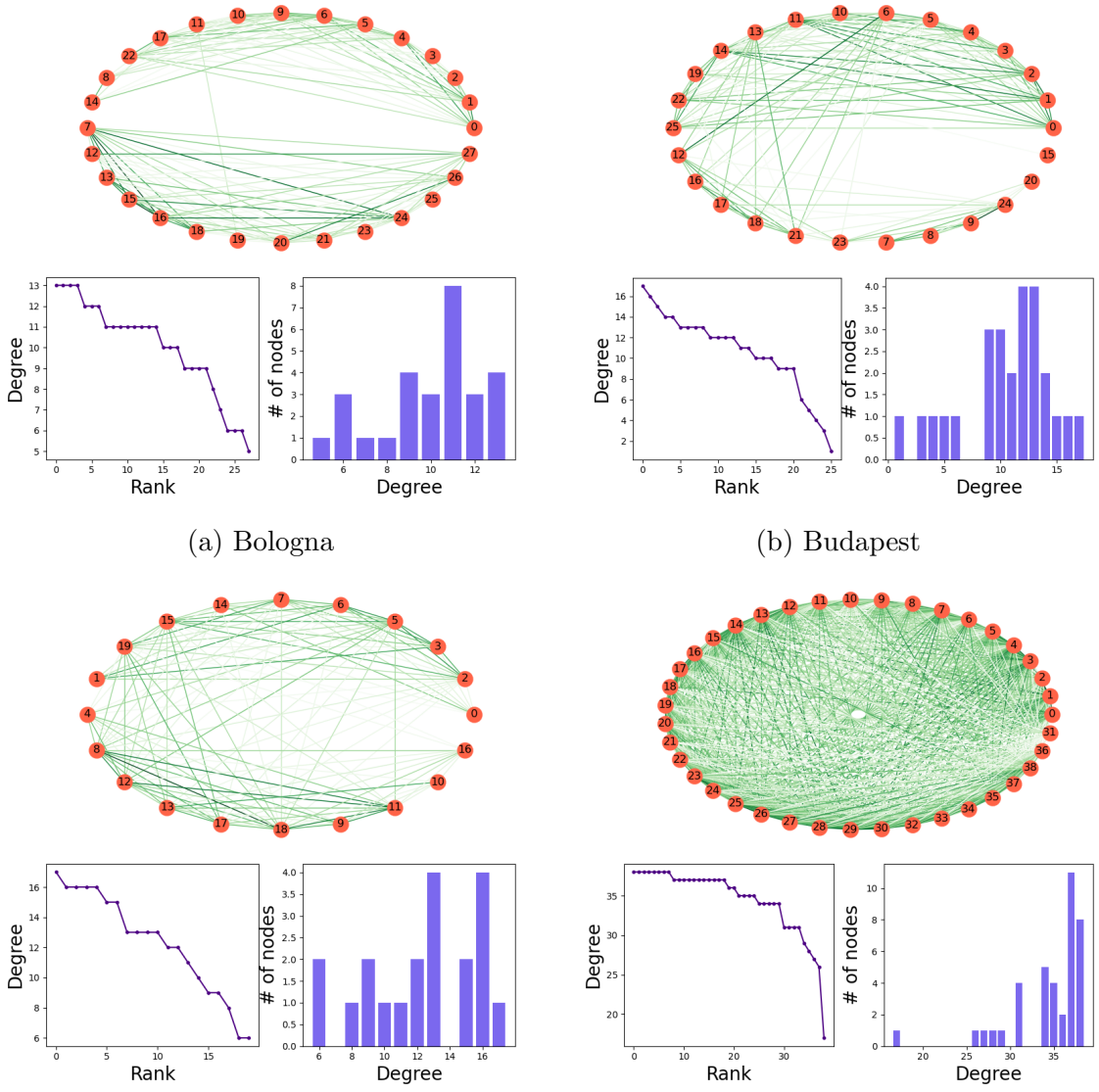
First of all, the samples (dis)similarities was quantified, in terms of bacterial composition (β-biodiversity), by computing the Bray-Curtis distance between each pair of samples based on the absolute species abundances. The computation was performed first considering all the samples, and then each city's samples separately. In both cases, the β-biodiversity matrix was exploited to explore the samples (dis)similarities using two different techniques:

1. building similarity networks (using the complement to 1 of the Bray-Curtis index);

2. performing the Principal Coordinate Analysis (PCoA), in order to plot the samples in a space with reduced dimensionality.

Similarity networks

Let us first consider the similarity networks, which are shown, at the species level and for the four cities, in Figures 3.1 a-d.



(a) Bologna

(b) Budapest

(c) Rome

(d) Rotterdam

Figure 3.1: Similarity networks based on Bray-Curtis dissimilarity metric computed at species level, for the four cities. The degree rank plot and the degree histogram, giving information on the degree distribution of nodes in the network, are also shown.

The networks were built drawing a link between two samples only if their distance (according to the Bray-Curtis metric) was smaller than a threshold value. Given

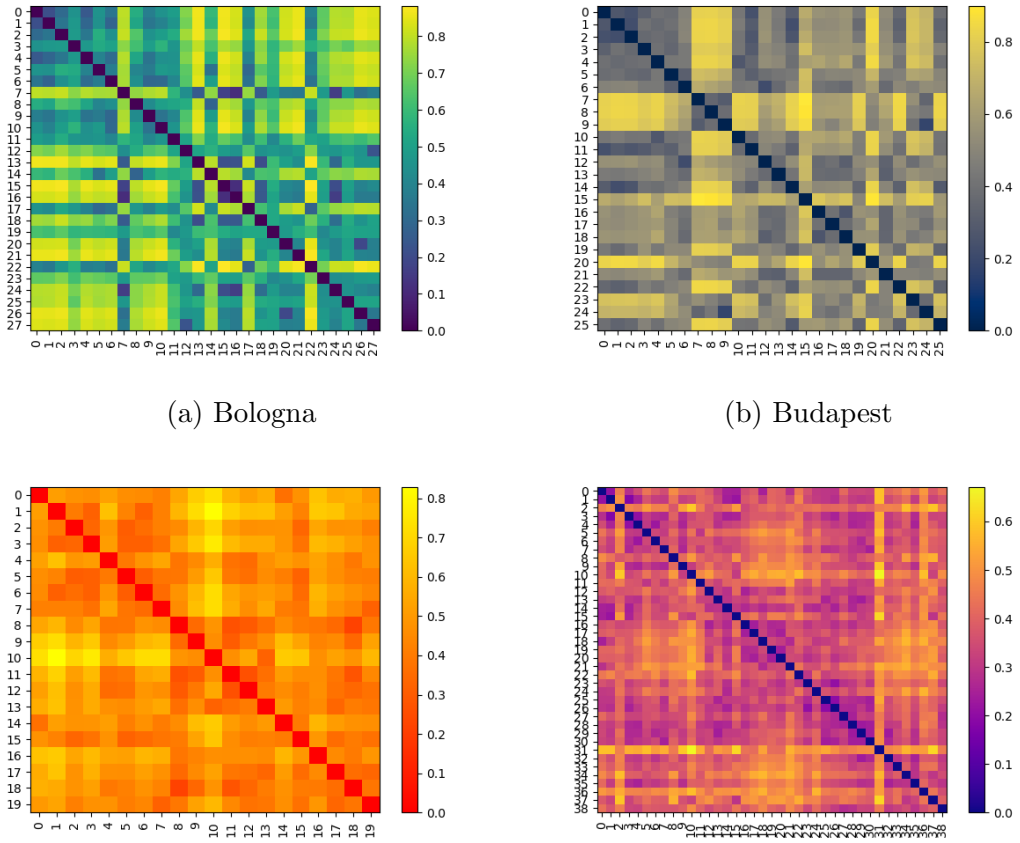
that the Bray-Curtis distance ranges from 0 to 1, we chose to fix the threshold to 0.5.

At least two important features may be noticed looking at those networks:

1. Rotterdam is much richer in terms of links than the other cities, meaning that many of the samples collected in Rotterdam have a similar bacterial composition; we will show that this peculiarity of Rotterdam is confirmed by the time series analyses.

2. In the other cities, particularly in Bologna and Budapest, the networks highlight two main components, which include samples collected during different time periods.

To better visualise the results, the heat map of the Bray-Curtis dissimilarities between samples, at species level, of the time series was generated, and are shown in Figures 3.2 a-d.



(a) Bologna

(b) Budapest

(c) Rome

(d) Rotterdam

Figure 3.2: Heat maps of the Bray-Curtis dissimilarities between samples of the four time series at species level

Notice that samples 7, 8, and 9 in Budapest, which were collected during the Summer of 2020 are very different from the other samples. The same behaviour may be noticed for the samples 9 and 10 in Rome, which, again, are the samples collected in the same Summer. Also this point will be confirmed by the following time series analyses.

Principal Coordinate Analysis

As already said, the dissimilarity matrices were also used to describe the samples in a reduced dimensional space, performing the PCoA.

The PCoA was used to analyse the samples diversities from two different perspectives: on one hand, one of the aim was to find potential patterns and clusters due to the variability of the samples' composition over time, thus inspecting possible effects due to seasonality; on the other hand, also spatial effects have been examined by comparing the cities in a unique PCoA.

Let us first consider the Bray-Curtis-based PCoA plots obtained within each city. The four PCoA plots are shown in Figures 3.3 a-d.

No seasonality effect seem to be detected from this analysis. Actually, only the Rome samples exhibit a few recognisable clusters, specifically those related to the Autumn and Summer seasons. It is however worth noticing that the time series of Rome is the shortest one, with only 20 samples and covering only 10 months; thus, as the samples do not even cover a whole year, it is difficult to explain and justify the seasonality clusters appearing in the plots.

The second purpose of the PCoA, instead, was to verify whether samples from the same cities collected at different time points cluster together.

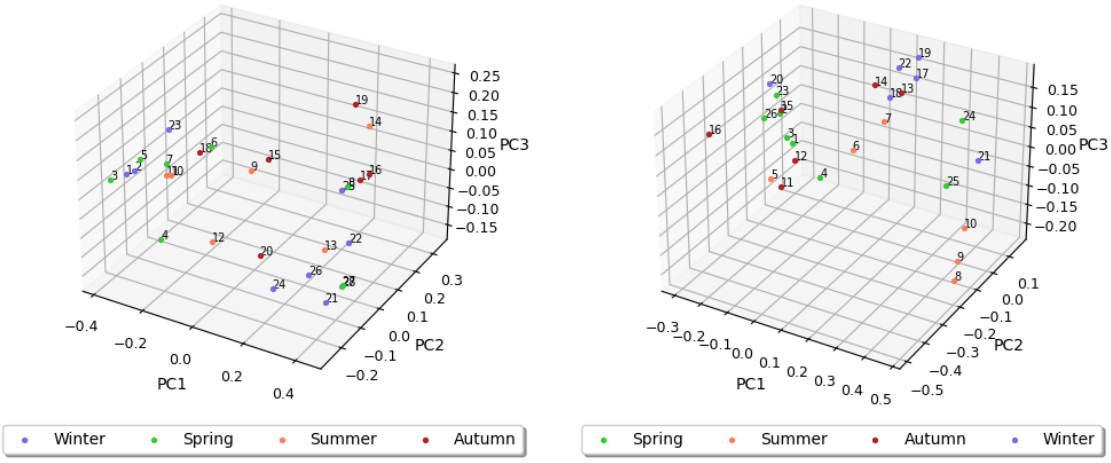
The 3-dimensional PCoA plot is shown in Figure 3.4.

It is worth noticing that samples collected from the same city tend indeed to cluster together, with the exception of samples from Bologna. In particular, it should be noticed that the time series of Rotterdam clusters on the positive values of the first principal coordinate.

3.2 Temporal characterisation of the sewage microbiome in the four cities

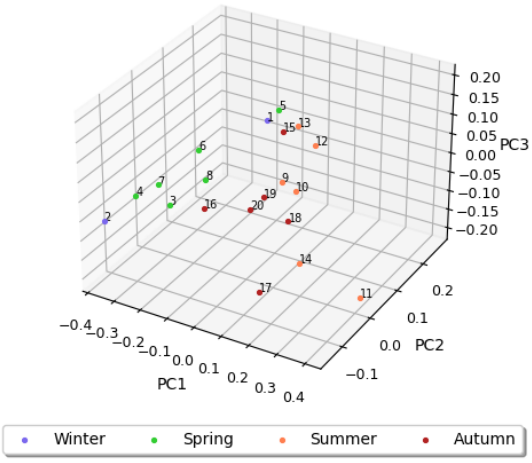
Time series of most abundant taxa

In order to explore the temporal trends of the sewage microbiome, we first plotted the time series of the most representative taxa in each of the four cities (Bologna, Budapest, Rome, Rotterdam).

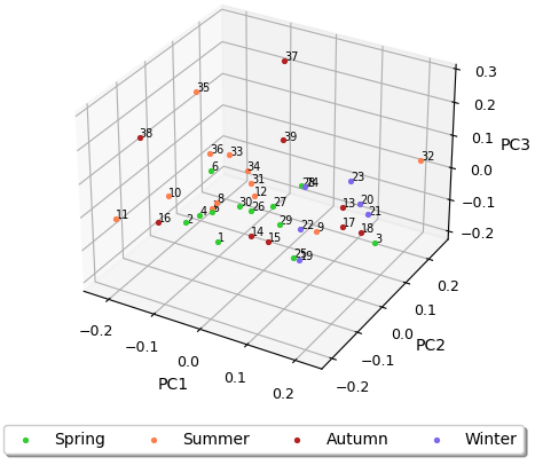


(a) Bologna

(b) Budapest



(c) Rome



(d) Rotterdam

Figure 3.3: PCoA 3-dimensional plot based on Bray-Curtis dissimilarity between samples in the four cities. Samples are coloured according to the season in which they were collected.

To this aim, we considered the bacterial relative frequencies obtained transforming the original absolute abundances with the centred log-ratio method, due to the compositional nature of the data.

Figures 3.5 a-d depict the time series of the seven most abundant species in each city. Similar plots for higher taxonomic levels (phylum, class, order, family, and genus), and for the four cities, are shown in Figures D.1, D.2, D.3, D.4, and D.5, in appendix D.

Overall, the time-series plots show that *Klebsiella pneumoniae* is the most abundant species in all cities and that its relative abundance is generally stable over time. On the other hand, other species appear among the seven most abundant ones only for specific cities (e.g. *Neisseria zalophi* is one of the top 7 species only in Budapest,

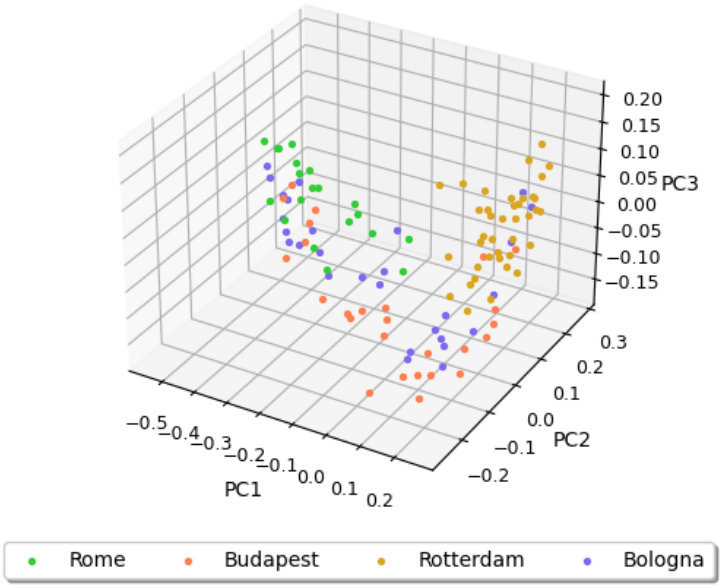
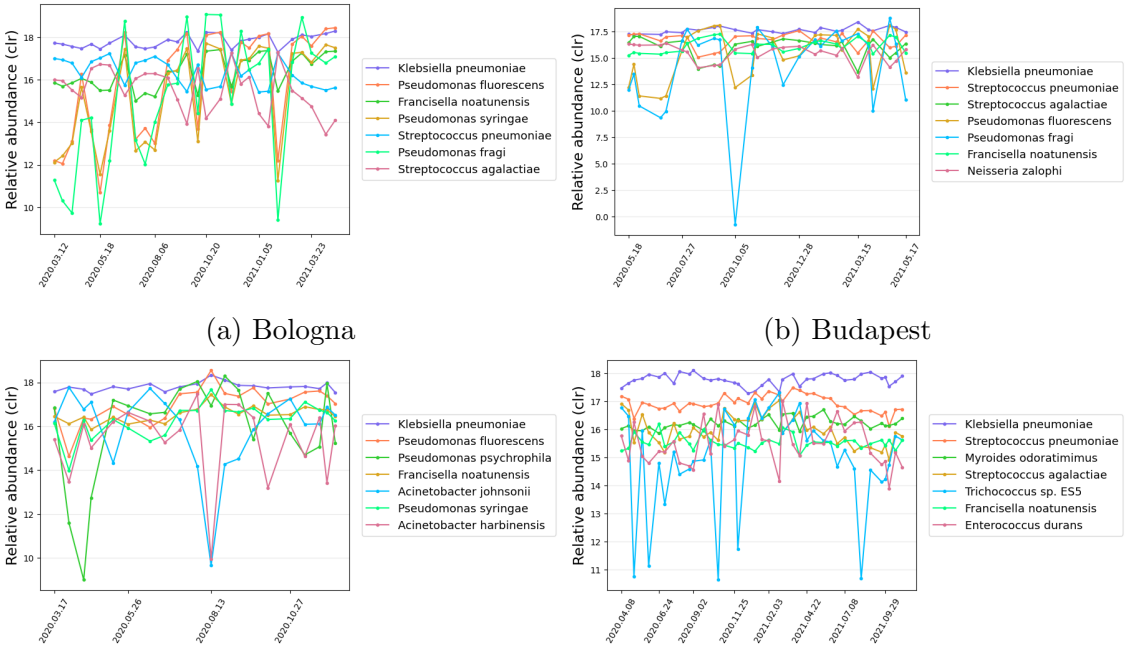


Figure 3.4: PCoA 3-dimensional plot based on Bray-Curtis dissimilarity between samples collected from all four cities at all time points. Samples are coloured according to their city of origin.



(a) Bologna (b) Budapest

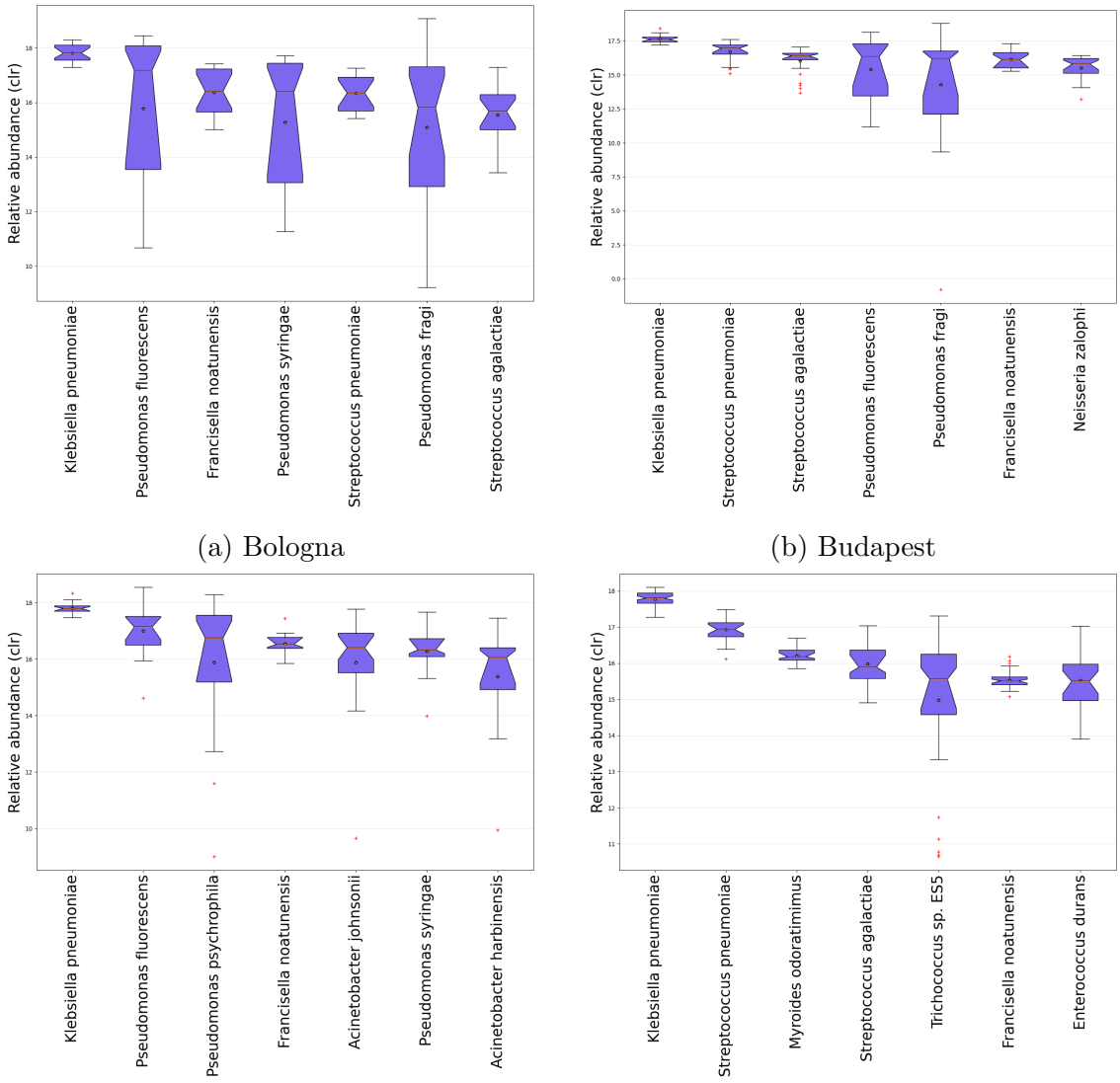
(c) Rome (d) Rotterdam

Figure 3.5: Time series of the 7 most representative species in the four cities (abundances expressed in centred log-ratio)

while *Pseudomonas fluorescens* is in the top 7 species in all cities but Rotterdam), and show a much higher variability over time.

In order to better highlight the variability of the relative abundances within each series, the box-plots[1] of the 7 most abundant species in each city are reported in Figures 3.6 a-d.

The box-plots show that among the most representative species, those with the highest variability belong to the *Pseudomonas* genus, which is not in the top 7 species in Rotterdam.



(a) Bologna

(b) Budapest

(c) Rome

(d) Rotterdam

Figure 3.6: Box-plots summarising the distributions within the time series of the 7 most representative species (expressed in centred log-ratio) in the four cities.

[1]In the box-plots: the box represents the interquartile (IQR) range; the upper (lower) whisker represents the largest (smallest) value no further than $1.5 \cdot IQR$ from the 3-rd (1-st) quartile; the horizontal line is the median value; the red crosses indicate the outliers; the star represents the mean value; the notches represent the 95% confidence interval of the median obtained by "bootstrapping" (1000 re-sampling).

Periodicity of the time series

The fluctuations of the time series observed in Figure 3.5, suggested the presence of some periodicities. In order to verify the presence of periodicities, and in particular of seasonality, an harmonic analysis of the time series through the Fourier Spectral Analysis was performed.

An approximation was required for this analysis, due to the fact that the sampling dates were not uniformly separated; however, this approximation appears reasonable, taking into account the fact that sampling was scheduled every two weeks and that the difference between the predicted time gaps and those actually observed was quite small (see table 2.2 in section 2.1.1).

The peak of each time series was analysed using the Fisher's test described in section 2.4.2 of chapter 2, and the results showed some significant periodicities ($p < 0.05$) and some highly significant peaks ($p < 0.01$) (see Table 3.1).

For higher taxonomic ranks, the results of the Fisher's test for periodicities are reported in appendix D.

Species	**City**	**Period [weeks]**	**Significance**
Pseudomonas fluorescens	Rome	40	*
Francisella noatunensis	Rome	40	*
Klebsiella pneumoniae	Budapest	26	*
Streptococcus agalactiae	Budapest	26	*
Klebsiella pneumoniae	Rotterdam	39	**
Streptococcus pneumoniae	Rotterdam	78	**
Myroides odoratimimus	Rotterdam	78	*

Table 3.1: Fisher test for periodicity of species (* means a p-value < 0.05, while ** means a p-value < 0.01)

Overall, only 3 species show a periodicity in Rotterdam, 2 species in both Rome and Budapest, and no species result in a significant periodicity in Bologna.

It is worth pointing out that, especially in case of short time series, local minima or maxima occurring in the middle of a time series may lead to spurious significant periodicities. That may be the case of Rome, where the local minimum of many taxa occurs roughly in the middle of the time series, giving thus rise to a periodicity with a 40 weeks period (since 40 weeks is the time interval covered by the whole time series).

Figure 3.7 shows an example of Fourier spectrum, referred to the *Klebsiella pneumoniae* species in Rotterdam; it is worth noticing that a peak in the spectrum is found for the second harmonic. The *Klebsiella pneumoniae* species was chosen as

an example because the Fisher's test applied to the peak (occurring at the second harmonic) results in a highly significant periodicity, with period of 39 weeks, that is half of the total number of weeks covered by the Rotterdam time series.

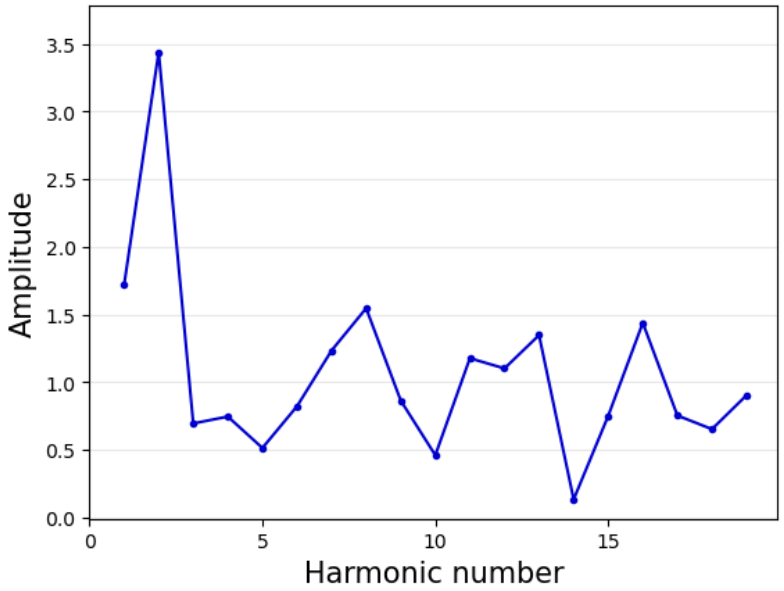


Figure 3.7: Example of Fourier Spectral Analysis (concerning the *Klebsiella pneumoniae* species in Rotterdam: the peak of the second harmonic, corresponding to a period of around 9 month, was evaluated by Fisher's test, $p < 0.01$)

Unfortunately, however, these results may be biased or even misleading, because of two concurrent reasons: first, the time interval between two consecutive data is not exactly uniform; second, the series taken into account are short and noisy (see for example Figure 3.5).

In the future, the hope will be to apply the Fourier spectral analysis to longer and less noisy series. For this purpose, in the VEO project, newer sewage's samples are being sampled and sequenced from the four cities; moreover, the bioinformatics pre-processing step is being optimised, in order to reduce noise.

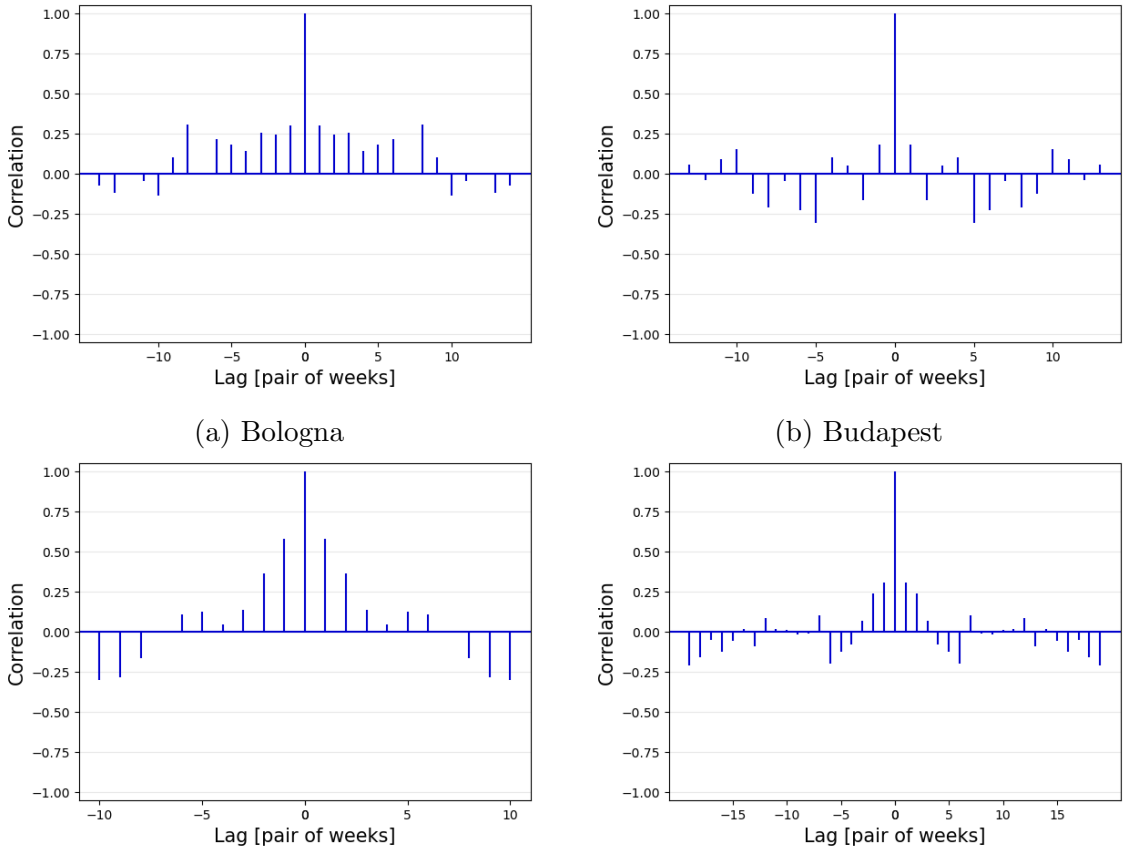
Auto-correlation and cross-correlation

The temporal stability of the bacterial relative abundances and of the pairwise relationships where investigated by means of the auto-correlation and cross-correlation techniques. Here, we focused our analysis on phyla rather than species, so that to consider the data at a more stable taxonomic level.

Here, the auto-correlogram of the most abundant phyla (*Proteobacteria*) and the cross-correlograms of the two most abundant phyla (*Proteobacteria* and *Firmicutes*) are reported (see Figures 3.8 and 3.9). The auto-correlograms of the 4 most

abundant phyla, and their cross-correlograms are reported in appendix D (D and D).

Notice that also in this case, the analyses were carried out starting from the data transformed via centred log-ratio.



(a) Bologna

(b) Budapest

(c) Rome

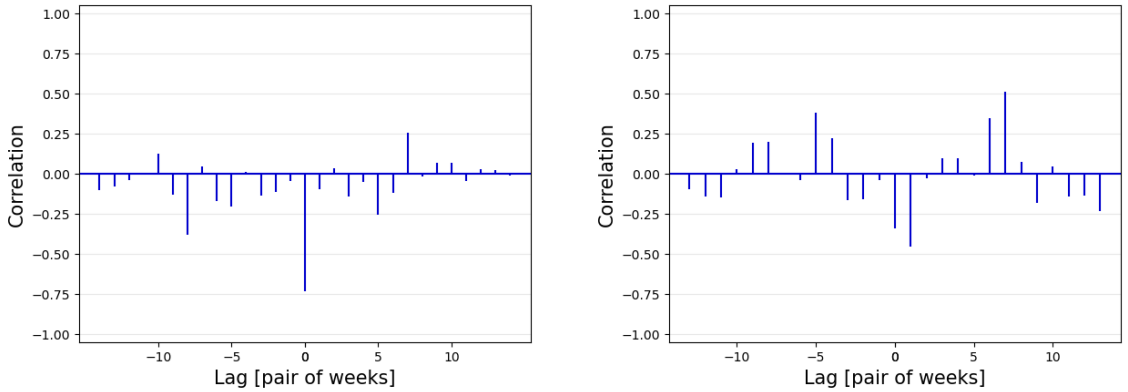
(d) Rotterdam

Figure 3.8: Example of auto-correlation for the *Proteobacteria* phylum in the four cities.

It is worth noticing that the auto-correlograms of all phyla have similar shape, which is conserved in all the four cities: positive auto-correlations are shown for small lags, while negative correlations appear for larger lags. An example is shown in Figure 3.8 for the *Proteobacteria* phylum in the four cities. It is difficult to properly interpret this behaviour, but it may be linked to the ability of bacteria to persist over time.

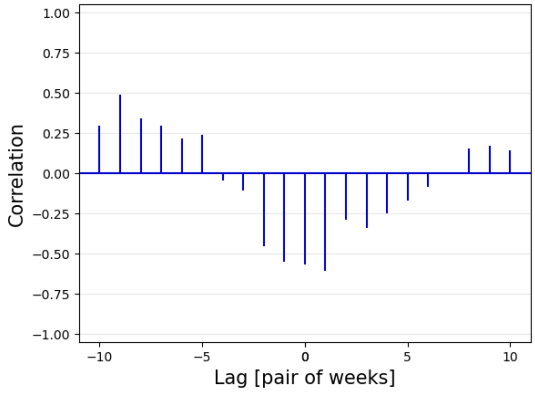
Also, it is worth mentioning that the above-described behaviour occurs also for many other taxa, and at all taxonomic levels, even if results are not shown in this thesis for brevity.

Figure 3.9 shows the cross-correlograms of the pair of phyla *Proteobacteria* and *Firmicutes* for all the cities.

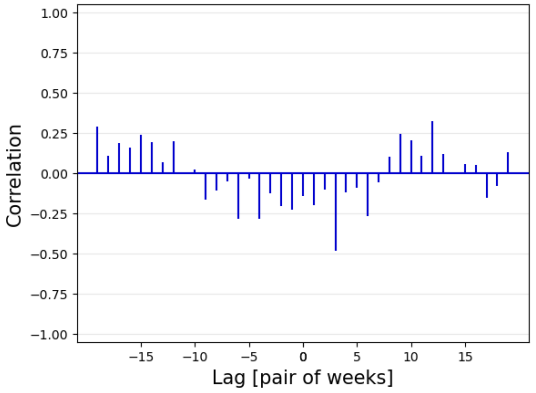


(a) Bologna

(b) Budapest



(c) Rome



(d) Rotterdam

Figure 3.9: Example of cross-correlation between *Proteobacteria* and *Firmicutes* phyla in the four cities.

For what concerns the cross-correlograms, no particular behaviour can be identified at first glance. In fact, the cross-correlation strongly depends not only on the chosen pair of taxa (as expected), but also on the considered city.

It is worth mentioning that the anti-correlation occurring between *Proteobacteria* and *Firmicutes* (Figure 3.9) finds some confirmations also in other very different contexts. For instance, Li *et al.* [30] have shown that the manure addition in an originally unproductive soil may considerably modify the abundances of microorganisms. In particular, the abundance of *Firmicutes* decreases when the manure is added, while *Proteobacteria* behaves in the opposite way. Hence, they results to be anti-correlated. Even more surprising is the anti-correlation between those two phyla, that emerges in a study which have analysed the human milk microbiota [31].

3.3 Temporal characterisation of the sewage α - diversity in the four cities

In order to provide an ecological characterisation of the available metagenomic samples, we used a set of nine indices to estimate their biodiversity in terms of $\alpha-$diversity. Notice that, depending on the index used, one or more of the following biodiversity factors will be captured: species richness, evenness and taxonomic distance.

First of all, we considered the Shannon index (Figure 3.10a), the Pielou index (Figure 3.10b), the Hill numbers of order 1 and 2 (^1D and ^2D, Figures 3.10c and 3.10d), and Gini-Simpson index (Figure 3.10e). Each index was computed for each sample, i.e. for each bacteria population corresponding to a certain date and a certain city, and it was than plotted as a function of time to facilitate the comparisons.

The results shown in Figure 3.10 exhibit some interesting features. First of all, Rotterdam's biodiversity, computed by means of all the above-mentioned indices, shows a much higher stability than that of the other cities. On the other hand, Bologna is characterised by very large fluctuations and by high noise. Finally, Rome and Budapest present an intermediate situation between those of Rotterdam and Bologna; however, in both cities there is a period of biodiversity collapse corresponding to the Summer of 2020, the significance of which deserved to be investigated.

Notice that all the indices considered up to now take into account both species richness and evenness.

In order to also take into account the taxonomic distance among taxa, we used two further indices: the Clarke and Warwick's Taxonomic Distinctness Index and Taxonomic Binary Distinctness Index.

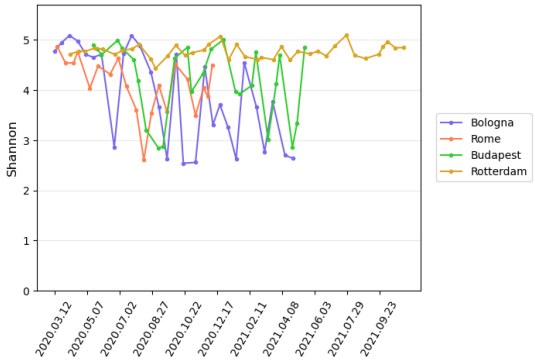
Both indices required an extra preprocessing step to be computed, that is the reconstruction of the taxonomic and phylogenetic classification of the taxa present in the metagenomes, an information that was not available in our data. To this aim, we exploited the NCBI database to establish the taxonomic distance between two different taxa (i.e. the distance to the closest common ancestor).

Due to the high time consuming processes required for the computation of these indices, they both have been calculated starting from the family level instead of the species level and for this reason they can not be strictly compared to the previously presented indices.

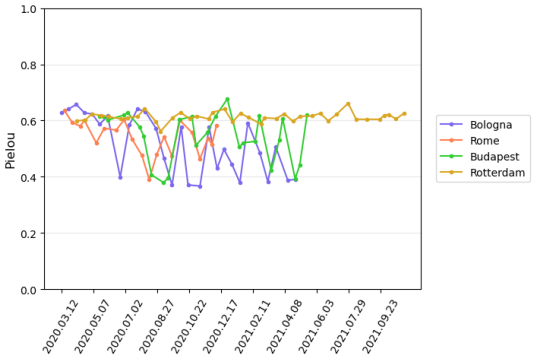
The plots of the two Clarke and Warwick's indices are shown in Figure 3.11

Both indices confirm the low variability of the Rotterdam time series and the presence of a decrease of biodiversity in the Rome and Budapest series during Summer of 2020.

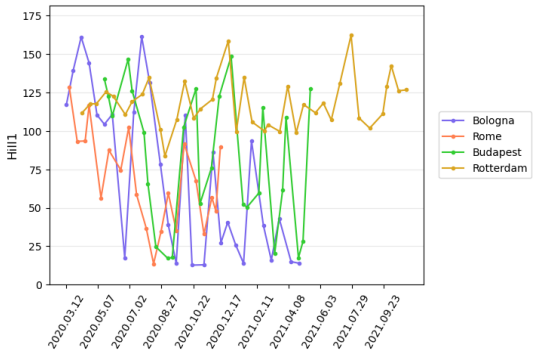
Although the Taxonomic Binary Distinctness index embeds less information than the other one (since, in this case, the evenness is not taken into account), it is worth



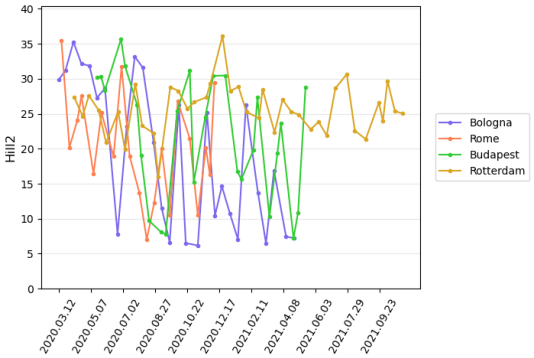
(a) Shannon index.



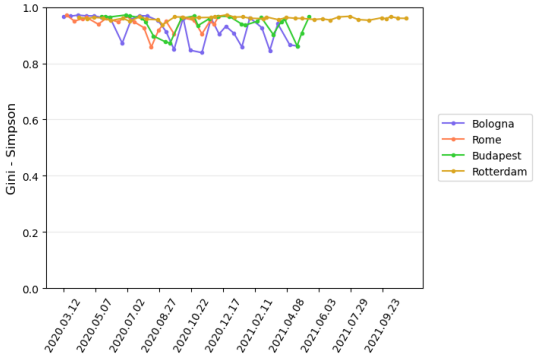
(b) Pielou index.



(c) Hill number of order 1.



(d) Hill number of order 2.



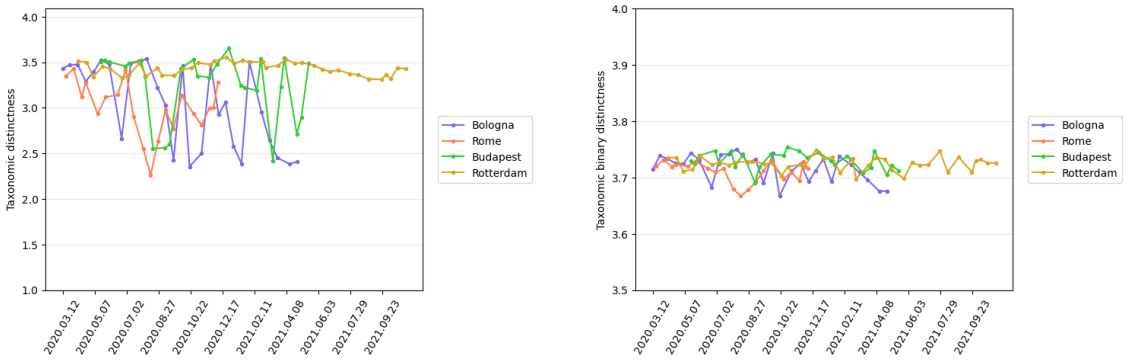
(e) Gini-Simpson index.

Figure 3.10: Time series of $\alpha-$diversity in the four cities.

of interest since it provides the trend in time of the average taxonomic distance between families. It can be noticed, in Figure 3.11b, that it is stable around the value of 3.7, meaning that, independently on time or city, the common ancestor is on average at the superkingdom or phylum level.

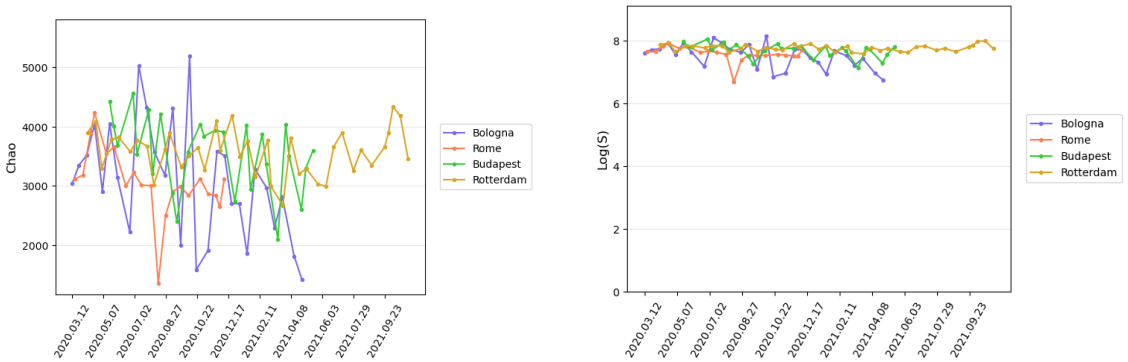
Lastly, the Chao index and the logarithm of the total number of species, $Log(S)$, have been computed and plotted (Figure 3.12).

These two indices are a measure of the species richness. They have been used to check and visualise how the number of species behaves over time in the different cities. On one hand, the Chao index gives an estimate of the real number of species



(a) Taxonomic distinctness index. (b) Taxonomic binary distinctness index.

Figure 3.11: Time series of the two Clarke and Warwick's indices in the four cities.



(a) Chao index. (b) Log(S) index.

Figure 3.12: Time series of species richness in the four cities.

belonging to a population. On the other hand, the species richness S is the actual number of species found in the population, and here is expressed in logarithmic form ($Log(S)$) so that it immediately reminds of the normalisation factor applied to the Shannon index to get the Pielou one.

It is worth noticing that also Chao and $Log(S)$ confirm the previously observed patterns.

As already highlighted above, there are two key points emerging from these analyses: the first is the stationarity of the time series referred to Rotterdam; the second is the decrease of biodiversity occurring in two specific periods of the series of Rome and Budapest. These features of the data can be observed in all the above-shown indices, which, indeed tends to be characterised by the same type of fluctuations, i.e. when a local minimum (or maximum) occurs in an index, for a given city, the other indices of the same city have a minimum (or maximum), too. This means that the studied samples are strongly affected by all the factors concerning the concept of biodiversity (species richness, evenness and taxonomic distinctness).

Stationarity and variability of the α - biodiversity indices

Finally, the stationarity of the time series around a constant value has been evaluated through the Augmented Dickey-Fuller test, applied to the Shannon, Pielou, Gini-Simpson, and Chao indices. The results are collected in table 3.2.

Index	City	p-value	Significance	Stationarity
Shannon	Bologna	0.910		No
	Budapest	0.007	**	Yes
	Rome	0.055		No
	Rotterdam	<0.001	**	Yes
Pielou	Bologna	0.859		No
	Budapest	0.008	**	Yes
	Rome	0.028	**	Yes
	Rotterdam	<0.001	**	Yes
Gini-Simpson	Bologna	0.866		No
	Budapest	0.272		No
	Rome	0.007	**	Yes
	Rotterdam	<0.001	**	Yes
Chao	Bologna	<0.001	**	Yes
	Budapest	<0.001	**	Yes
	Rome	0.148		No
	Rotterdam	<0.001	**	Yes

Table 3.2: Augmented Dickey-Fuller test for stationarity (* and ** mean $p < 0.05$ and $p < 0.01$, respectively)

What can be noticed is that Bologna has a significant stationarity only for the Chao index time series, while it is not stationary with respect to a constant value for all the other indices. In fact, the trend of the biodiversity in Bologna is decreasing, even if the species richness remains stationary in time with respect to a constant. This is an interesting behaviour that could be further explored in the future, when a larger amount of data will be available.

As expected, Rotterdam is stationary, no matter the index considered.

Then, the variability of the time series of the four cities were evaluated by comparing the variance of four biodiversity indices (Shannon, Pielou, Gini-Simpson, Chao)[2] between all pairs of cities using the Brown-Forsythe test.

Table 3.3 summarises the results.

		Bologna	**Budapest**	**Rome**
Shannon	**Budapest**	0.045 *		
	Rome	< 0.001 **	0.161	
	Rotterdam	< 0.001 **	< 0.001 **	< 0.001 **
Pielou	**Budapest**	0.093		
	Rome	< 0.001 **	0.094	
	Rotterdam	< 0.001 **	< 0.001 **	< 0.001 **
Gini-Simpson	**Budapest**	0.056		
	Rome	0.014 *	0.494	
	Rotterdam	< 0.001 **	< 0.001 **	< 0.001 **
Chao	**Budapest**	0.049 *		
	Rome	0.010 **	0.243	
	Rotterdam	< 0.001 **	0.011 *	0.524

Table 3.3: Brown-Forsythe test p-value for the homogeneity of variances of various biodiversity indices; * and ** mean significant heterogeneity, with $p < 0.05$ and $p < 0.01$, respectively

For the sake of clarity, it is worth highlighting that the Brown-Forsythe and the Augmented Dickey-Fuller test give different information on the time series:

- the Brown-Forsythe test compares different time series to check if they have the same variability; in this context it was used to verify if the Rotterdam time series has a significant stability[3];
- instead, the Augmented Dickey-Fuller test gives a significant result when a time series has a stationary trend around a constant value, no matter how large the fluctuations around it are.

[2]For the sake of clarity, only four indices have been chosen because the other ones (such as the Hill numbers) give similar information about the samples.

[3]Here, the term stability is intended as a synonymous of "low variability".

3.4 Understanding the decrease in biodiversity in Rome and Budapest during Summer of 2020

Proceeding with the analyses, one of the main aims was to investigate on the reasons that have led to the drop of biodiversity occurring in Rome and Budapest around August and September, in 2020.

Time evolution of the most abundant genera and species

To this aim, the evolution over time of the most representative genera and species has been examined for the two cities. In this context, the concept of "most representative" is intended as the most abundant bacteria in the period of biodiversity minimum.

Looking at the time series of the most abundant species, we noticed that many of the most representative species belong to the same genus and behaves in the same way. Hence, we decided to perform the analysis of the evolution of taxa over time not only at species level, but also at genus level.

For what concerns Rome, the observation of the time series of the individual genera has highlighted how the drops of biodiversity coincide with an increase in the relative abundances of 5 species; four of them belong to the genus *Pseudomonas* (*P. fluorescens*, *P. psychrophila*, *P. fragi*, *P. syringae*) and one to the genus *Klebsiella* (*K. pneumoniae*). The increase of these 5 species causes an abrupt decrease of the others species, especially the ones belonging to the *Acinetobacter* genus, as can be seen in Figure 3.13 and 3.14.

From the comparison of the time series of the biodiversity indices and the time series of the most abundant taxa, it was possible to notice some similarities and dissimilarities among the cities. For instance, it is noteworthy that both *K. pneumoniae* and the species of the genus *Pseudomonas* seem to have a significant role in the biodiversity fluctuation over time. Also, it is interesting to point out that no species of the genus *Pseudomonas* is among the most representative ones in Rotterdam. This behaviour is reflected also at the level of genus (*Pseudomonas*, *Klebsiella*), of family (respectively *Pseudomonadaceae*, *Enterobacteriaceae*), of order (respectively *Pseudomonadales*, *Enterobacteriales*); from the class level upwards, the two considered genera belong to the same taxa: *Gammaproteobacteria* for the class and *Proteobacteria* for the phylum.

The same behaviours and features described for the time series of Rome may be done also for Budapest. In this case it is worth noticing that the *Streptococcus* genus is much more present in Budapest than in Rome (see Figure 3.15); also the species *Streptococcus pneumoniae* and *Streptococcus agalactiae* seem to play the same role played by the *Acinetobacter johnsonii* in Rome, i.e. it decreases when the species of

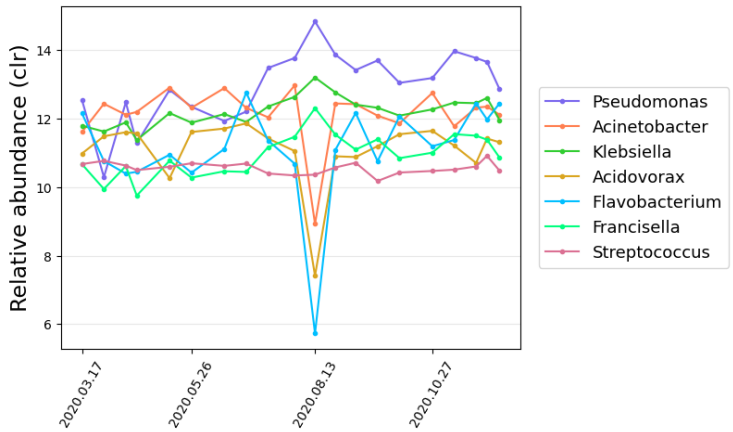


Figure 3.13: Time series of the 7 most representative genera in Rome, highlighting the collapse of three genera (*Flavobacterium*, *Acidovorax*, *Acinetobacter*) and the rise of three others (*Pseudomonas*, *Klebsiella*, *Francisella*)

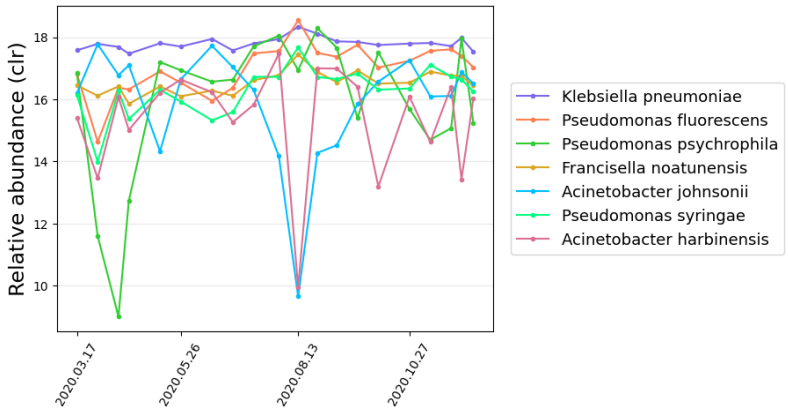


Figure 3.14: Time series of the 7 most representative species in Rome, highlighting the collapse of two species (*Acinetobacter johnsonii* and *Acinetobacter harbinensis*) and the rise of *Pseudomonas fluorescens*

Pseudomonas increases, and viceversa (see Figure 3.16).

Up to now, only qualitative observations have been shown to describe the different taxa compositions occurring during the time series. Now, a quantitative approach to explain these observations will be deepened.

In order to understand which differences occur among the period of minimum biodiversity and the other periods, in Rome and Budapest, we first divided the time series of each city into three periods: before, during and after the biodiversity minimum. The subdivision was based on the plot of the Shannon index, which is used to recognise the sub-period related to the above-mentioned minimum of biodiversity (see Figures 3.17a e 3.17b for the visualisation of the three periods in

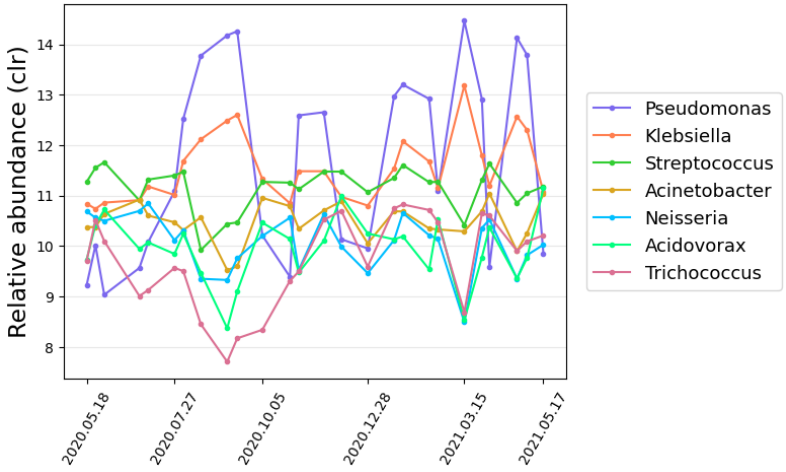


Figure 3.15: Time series of the 7 most representative genera in Budapest, highlighting the collapse of five genera (*Streptococcus*, *Trichococcus*, *Acidovorax*, *Acinetobacter*, *Neisseria*) and the rise of two others (*Pseudomonas*, *Klebsiella*)

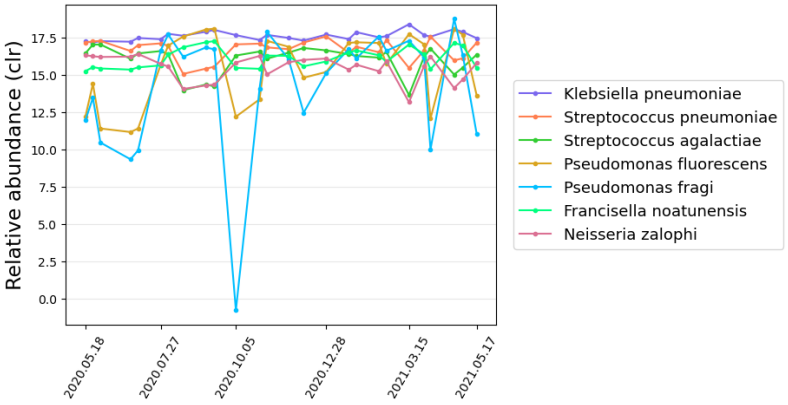


Figure 3.16: Time series of the 7 most representative species in Budapest

Rome and Budapest, respectively). Then, the 10 most abundant species[4] in the second time interval (i.e. inside the minimum) were considered.

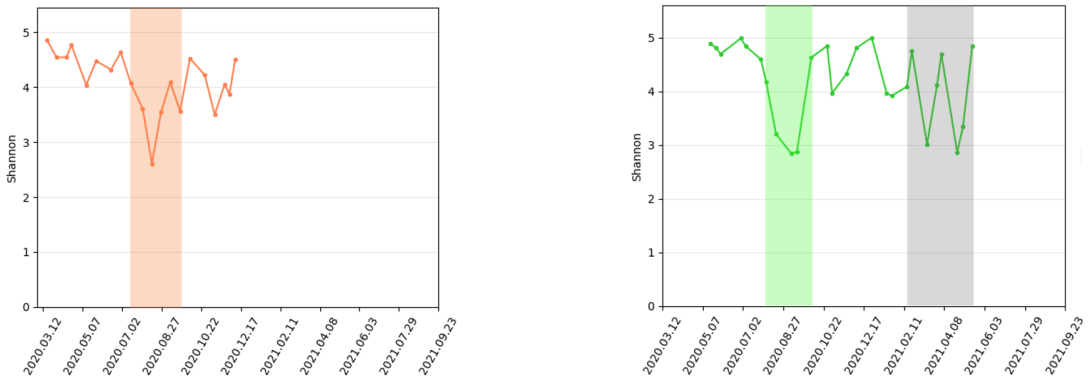
It is worth mentioning that some samples of the Budapest time series (those covered by the grey region in Figure 3.17b) are excluded in the analyses of the biodiversity minimum, because of the high fluctuations.

Thus, each series of the ten most abundant species is divided into three groups which were analysed by one-way ANOVA, followed by Honestly Significance Difference (HSD) Tukey's post-hoc test for multiple comparisons.

Tables 3.4 and 3.5 summarise the results for Rome and Budapest, respectively.

At least three important features can be noticed looking at the tables of both Rome and Budapest:

[4]In this context, the species have been ordered according to the median value of the abundances, expressed in centred log-ratio.



(a) Rome

(b) Budapest

Figure 3.17: Time series of Shannon index in Rome and Budapest; the time window corresponding to the collapse of biodiversity is in green, so that a subdivision in three periods may be done (before, inside, and after the minimum); the grey area indicates the fraction of time series excluded in the analyses of the biodiversity collapse, because of the high fluctuations

1. the most abundant species in the period of minimum biodiversity tend to change more their abundances between the first and the second periods (i.e. before and inside the minimum). Specifically, the abundances increase in the minimum;

2. no significant changes are detected while comparing the second and the third periods (i.e. inside and after the minimum);

3. as a consequence of both point 1 and 2, different significant results occur when comparing the first and the third periods (i.e. before and after the minimum), meaning that the fall in biodiversity may have caused a strong change in the bacterial composition of the sewage.

For what concerns the feature described in point 1: it is consistent with what we expected, since a decrease in biodiversity is related to the increases of the abundances of one or more species.

The same reasoning and observations may be done for the 10 most abundant genera of the second sub-period of Rome and Budapest (during the minimum), as can be seen in Tables 3.6 and 3.7.

It is worth noticing that the abundance of the *Streptococcus pneumoniae* species, in Budapest, significantly changes its behaviour inside the minimum, i.e. it reaches its minimum value at the minimum. Then, after the biodiversity fall, it increases again up to the abundances it had before the minimum (see Figure 3.18). This is in contrast with the above-described behaviour of the other species.

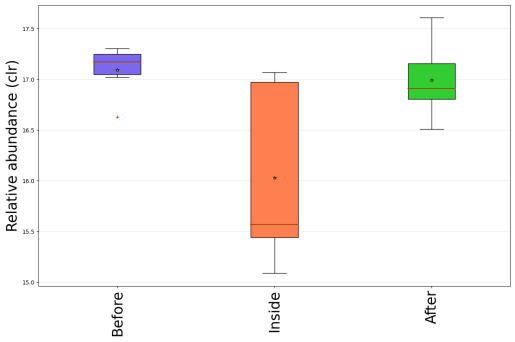


Figure 3.18: Behaviour of *Streptococcus pneumoniae* in Budapest, decreasing in the period of biodiversity collapse and then returning to previous level

Instead, an example of the typical behaviour of the most abundant species in the three sub-periods is depicted in Figure 3.19 by means of a box-plot.

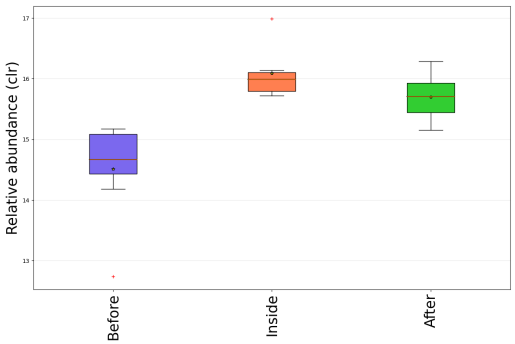


Figure 3.19: Example of characteristic behaviour of most species both in Rome and in Budapest, with species increasing in the period of biodiversity collapse and then remaining to the same level (*Pseudomonas filiscindens*, Rome)

Species	**F-test p-value**	**Tukey before vs inside (p-value, trend)**	**Tukey inside vs after (p-value, trend)**	**Tukey before vs after (p-value, trend)**
Klebsiella pneumoniae	0.015	0.013 ↑	0.086 =	0.717 =
Pseudomonas psychrophila	0.116	=	=	=
Pseudomonas fluorescens	< 0.001	< 0.001 ↑	0.434 =	0.005 ↑
Francisella noatunensis	< 0.001	< 0.001 ↑	0.282 =	0.010 ↑
Pseudomonas syringae	0.001	0.002 ↑	0.592 =	0.016 ↑
Pseudomonas fragi	< 0.001	< 0.001 ↑	0.117 =	0.007 ↑
Acinetobacter harbinensis	0.797	=	=	=
Pseudomonas filiscindens	< 0.001	< 0.001 ↑	0.522 =	0.006 ↑
Serratia marcescens	< 0.001	< 0.001 ↑	0.249 =	0.006 ↑
Pseudomonas putida	< 0.001	0.001 ↑	0.191 =	0.037 ↑

Table 3.4: ANOVA and HSD Tukey's test results between three time windows (before, inside, and after biodiversity collapse); the analysis concerns the 10 most abundant species in Rome, expressed in centred log-ratio

Species	**F-test p-value**	**Tukey before vs inside (p-value, trend)**	**Tukey inside vs after (p-value, trend)**	**Tukey before vs after (p-value, trend)**
Klebsiella pneumoniae	0.001	0.001 ↑	0.060 =	0.056 =
Pseudomonas fluorescens	0.008	0.013 ↑	0.862 =	0.022 ↑
Pseudomonas syringae	0.008	0.011 ↑	0.787 =	0.025 ↑
Francisella noatunensis	0.004	0.003 ↑	0.177 =	0.081 =
Pseudomonas psychrophila	0.211	=	=	=
Pseudomonas fragi	0.389	=	=	=
Pseudomonas filiscindens	0.005	0.006 ↑	0.624 =	0.024 ↑
Serratia marcescens	0.010	0.008 ↑	0.302 =	0.102 =
Pseudomonas putida	0.003	0.002 ↑	0.117 =	0.078 =
Streptococcus pneumoniae	0.011	0.015 ↓	0.022 ↑	0.944 =

Table 3.5: ANOVA and HSD Tukey's test results between three time windows (before, inside, and after biodiversity collapse); the analysis concerns the 10 most abundant species in Budapest, expressed in centred log-ratio

Genus	F-test p-value	Tukey before vs inside (p-value, trend)	Tukey inside vs after (p-value, trend)	Tukey before vs after (p-value, trend)
Pseudomonas	< 0.001	$< 0.001 \uparrow$	$0.496 =$	$0.002 \uparrow$
Klebsiella	< 0.001	$< 0.001 \uparrow$	$0.171 =$	$0.019 \uparrow$
Acinetobacter	0.519	$=$	$=$	$=$
Psychrobacter	0.006	$0.005 \uparrow$	$0.360 =$	$0.103 =$
Francisella	< 0.001	$< 0.001 \uparrow$	$0.355 =$	$0.002 \uparrow$
Acidovorax	0.169	$=$	$=$	$=$
Flavobacterium	0.164	$=$	$=$	$=$
Streptococcus	0.049	$0.039 \downarrow$	$0.268 =$	$0.610 =$
Myroides	0.164	$=$	$=$	$=$
Janthinobacterium	0.320	$=$	$=$	$=$

Table 3.6: ANOVA and HSD Tukey's test results between three time windows (before, inside, and after biodiversity collapse); the analysis concerns the 10 most representative genera in Rome, in term of centred logratio transformed counts

Genus	F-test p-value	Tukey before vs inside (p-value, trend)	Tukey inside vs after (p-value, trend)	Tukey before vs after (p-value, trend)
Pseudomonas	0.008	0.006 ↑	0.232 =	0.110 =
Klebsiella	0.002	0.001 ↑	0.022 ↓	0.236 =
Francisella	0.007	0.005 ↑	0.104 =	0.206 =
Streptococcus	0.025	0.039 ↓	0.037 ↑	0.998 =
Acinetobacter	0.199	=	=	=
Neisseria	0.027	0.022 ↓	0.373 =	0.192 =
Myroides	0.803	=	=	=
Acidovorax	0.142	=	=	=
Comamonas	0.183	=	=	=
Endomicrobium	0.048	0.039 ↑	0.391 =	0.298 =

Table 3.7: ANOVA and HSD Tukey's test results between three time windows (before, inside, and after biodiversity collapse); the analysis concerns the 10 most representative genera in Budapest, in term of centred logratio transformed counts

Correlation networks before, during, and after the biodiversity decrease

Finally, in order to investigate the minimum of biodiversity, correlation networks have been built for the cities of Rome and Budapest, separately for three periods, already mentioned: before, during, and after the minimum.

The nodes of these networks are all those taxa having at least 10 counts in at least 75% of the samples[5]. This filtering procedure is used in order to avoid spurious correlations which may occur when dealing with vectors with a lot of null components.

The correlation networks are shown in Figure 3.20 for Rome and in Figure 3.21 for Budapest (in both cases, before, inside and after the minimum).

It is worth mentioning that the nodes are linked together when the Pearson's correlation coefficient between the two considered taxa is greater than a certain threshold (in the images the threshold was chosen equal to 0.5) and are weighted based on the value of the correlation (in the images, higher weights are represented by links with darker colours).

It is immediate to observe that, during the time periods of the biodiversity collapse, there is a greater correlation between species, which can be better visualised by looking at the degree rank plots.

Each series of the ten most central species is divided into three groups which were analysed by one-way ANOVA, followed by Honestly Significance Difference (HSD) Tukey's post-hoc test for multiple comparisons, following similar steps with respect to the one-way ANOVA analysis performed for the most abundant species.

Two different centrality measures have been used to evaluate the centrality of nodes, so that they could be ranked: degree centrality and betweenness centrality, which have been already described in chapter 2, section 2.3.

Once the nodes have been ranked, the one-way ANOVA procedure is performed to compare the mean abundances (in centred log-ratio) of the 10 most central nodes in the three sub-periods, for both Rome and Budapest.

Tables 3.8, 3.9, 3.10 and 3.11 summarise the results, for Rome and Budapest, and for both the centrality measures used to rank the nodes.

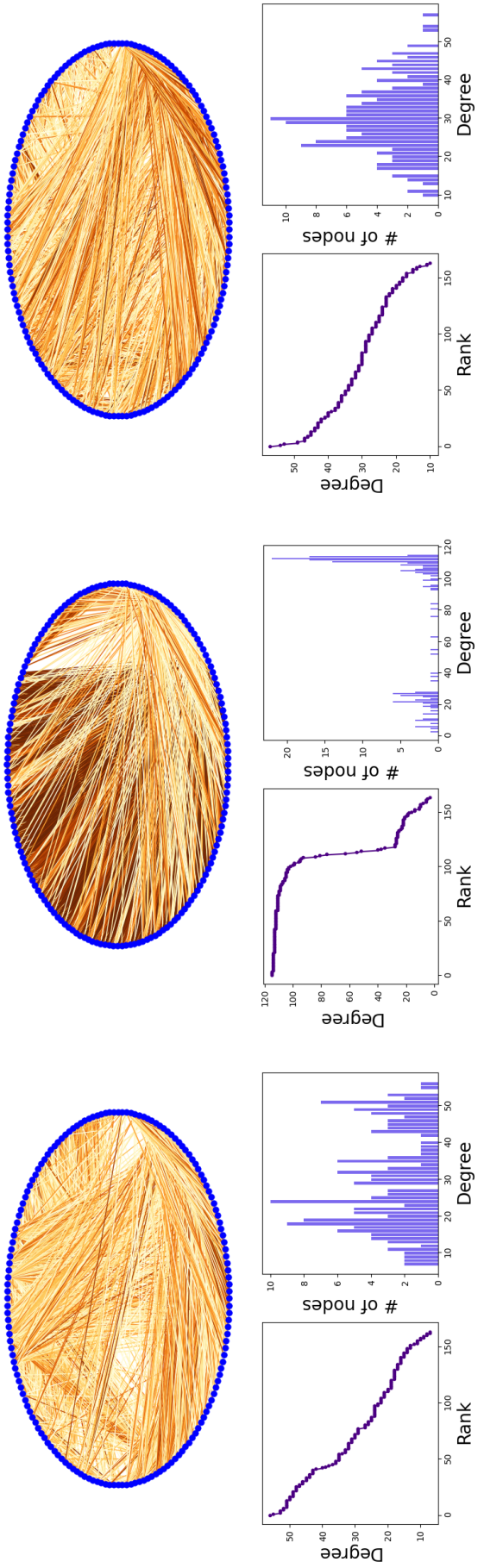
When considering the most central taxa, no significant feature or trend is recognisable.

In conclusion, the analysis performed on the most abundant species allows to propose a possible explanation for the biodiversity drop: in fact, the decreasing bio-

[5]Also other criteria (i.e. other pairs of parameter) have been used to select the taxa for the networks, for instance, considering those taxa having 100 counts in at least 75% of the samples, and so on. Here, it is preferred to consider the criterion written in the text since it retains information from more taxa, allowing more precise descriptions. The correlation networks resulting from the some of the other criteria may be found in the GitHub repository [29]

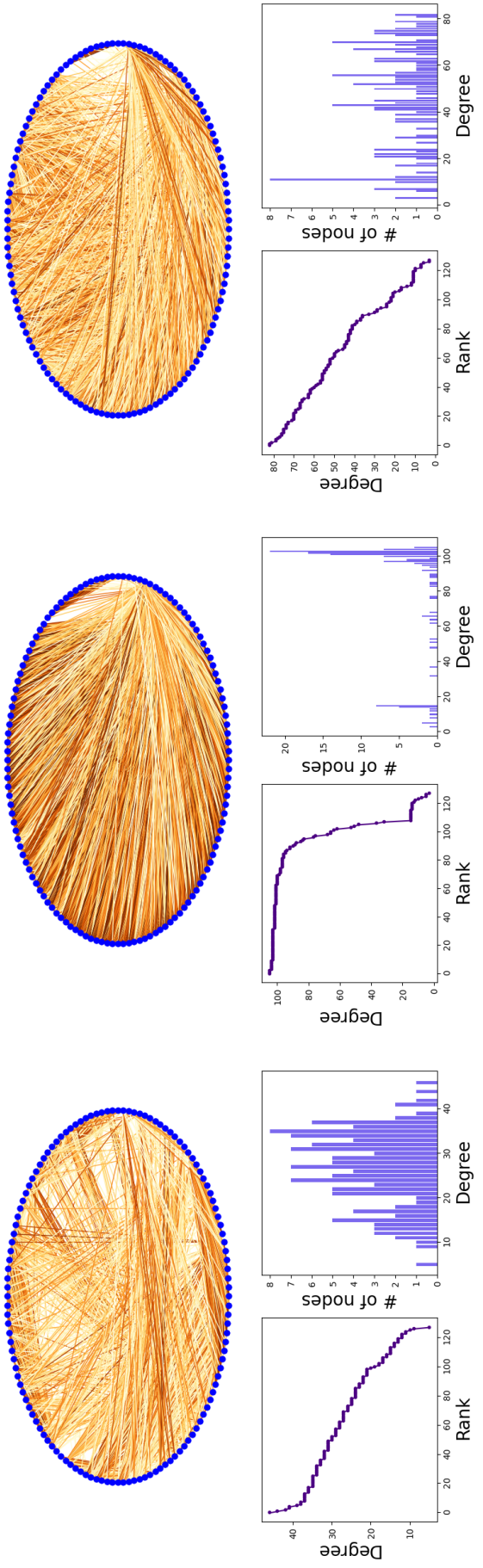
diversity correspond to the increase of the abundances of some species, as expected, in particular those which are the most abundant ones in the period of minimum biodiversity.

On the other hand, the biodiversity minimum seems to affect also the relationship among species: in fact, inside the minimum there is a richness of correlation among species; moreover, some of the most central ones change their abundances in the three periods, reaching their minimum when the biodiversity falls.



(a) Before (b) Inside (c) After

Figure 3.20: Correlation networks, along with their degree rank plots and degree histograms, of the three time periods (before, inside and after) related to the minimum in Rome



(a) Before (b) Inside (c) After

Figure 3.21: Correlation networks, along with their degree rank plots and degree histograms, of the three time windows (before, inside and after) related to the minimum in Budapest

Species	F-test p-value	Tukey before vs inside (p-value, trend)	Tukey inside vs after (p-value, trend)	Tukey before vs after (p-value, trend)
Psychrobacter glacincola	0.141	=	=	=
Pseudomonas baetica	0.411	=	=	=
Pseudomonas sp. E2.2	0.305	=	=	=
Pseudomonas taetrolens	0.245	=	=	=
Akkermansia muciniphila	0.003	0.003 ↓	0.216 =	0.114 =
Pseudoxanthomonas suwonensis	0.714	=	=	=
Streptococcus pneumoniae	< 0.001	< 0.001 ↓	0.126 =	0.016 ↓
Pseudomonas chlororaphis	0.074	=	=	=
Bacillus toyonensis	0.185	=	=	=
Stenotrophomonas koreensis	0.631	=	=	=

Table 3.8: ANOVA and HSD Tukey's test results between three time windows (before, inside, and after biodiversity collapse); the analysis concerns the 10 most representative species in Rome, ranked according their betwenness centrality

Species	**F-test p-value**	**Tukey before vs inside (p-value, trend)**	**Tukey inside vs after (p-value, trend)**	**Tukey before vs after (p-value, trend)**
Chryseobacterium solincola	0.192	=	=	=
Flavobacterium cucumis	0.109	=	=	=
Acidovorax sp. Root267	0.210	=	=	=
Delftia tsuruhatensis	0.041	0.033 ↓	0.433 =	0.351 =
Flavobacterium sp. LM5	0.182	=	=	=
Flavobacterium tegetincola	0.344	=	=	=
Bacteroides vulgatus	0.114	=	=	=
Mycoplasma arthritidis	0.207	=	=	=
Chryseobacterium chaponense	0.207	=	=	=
Flavobacterium aquatile	0.141	=	=	=

Table 3.9: ANOVA and HSD Tukey's test results between three time windows (before, inside, and after biodiversity collapse); the analysis concerns the 10 most representative species in Rome, ranked according their degree centrality

Species	**F-test p-value**	**Tukey before vs inside (p-value, trend)**	**Tukey inside vs after (p-value, trend)**	**Tukey before vs after (p-value, trend)**
Mycobacterium simulans	0.005	0.006 ↓	0.014 ↑	0.837 =
Neisseria sp. JXZ-10	0.013	0.020 ↓	0.891 =	0.030 ↓
Klebsiella pneumoniae	0.001	0.001 ↑	0.060 =	0.056 =
Stenotrophomonas sp. 92mfcol6.1	0.023	0.018 ↓	0.199 =	0.324 =
Desulfovibrio desulfuricans	0.062	=	=	=
Trichococcus sp. ES5	0.001	0.005 ↓	0.002 ↑	0.928 =
Cloacibacterium normanense	0.411	=	=	=
Trichococcus pasteurii	0.031	0.058 =	0.039 ↑	0.991 =
Streptobacillus moniliformis	0.040	0.058 =	0.060 =	0.993 =
Paenibacillus lactis	0.420	=	=	=

Table 3.10: ANOVA and HSD Tukey's test results between three time windows (before, inside, and after biodiversity collapse); the analysis concerns the 10 most representative species in Budapest, ranked according their betwenness centrality

Species	**F-test p-value**	**Tukey before vs inside (p-value, trend)**	**Tukey inside vs after (p-value, trend)**	**Tukey before vs after (p-value, trend)**
Dechloromonas sp. HZ	0.117	=	=	=
Bergeyella sp. AF14	0.234	=	=	=
Neisseria zalophi	0.002	0.002 ↓	0.020 ↑	0.317 =
Arcobacter cryaerophilus	0.717	=	=	=
Candidatus Arsenophonus triatominarum	0.396	=	=	=
Fusobacterium sp. CSL-7530	0.008	0.010 ↓	0.021 ↑	0.871 =
Sebaldella termitidis	0.048	0.042 ↓	0.540 =	0.207=
Comamonas aquatica	0.056	=	=	=
Uruburuella suis	< 0.001	< 0.001 ↓	< 0.001 ↑	0.944 =
Bacillus phocaeensis	0.011	0.011 ↓	0.038 ↑	0.723 =

Table 3.11: ANOVA and HSD Tukey's test results between three time windows (before, inside, and after biodiversity collapse); the analysis concerns the 10 most representative species in Budapest, ranked according their degree centrality

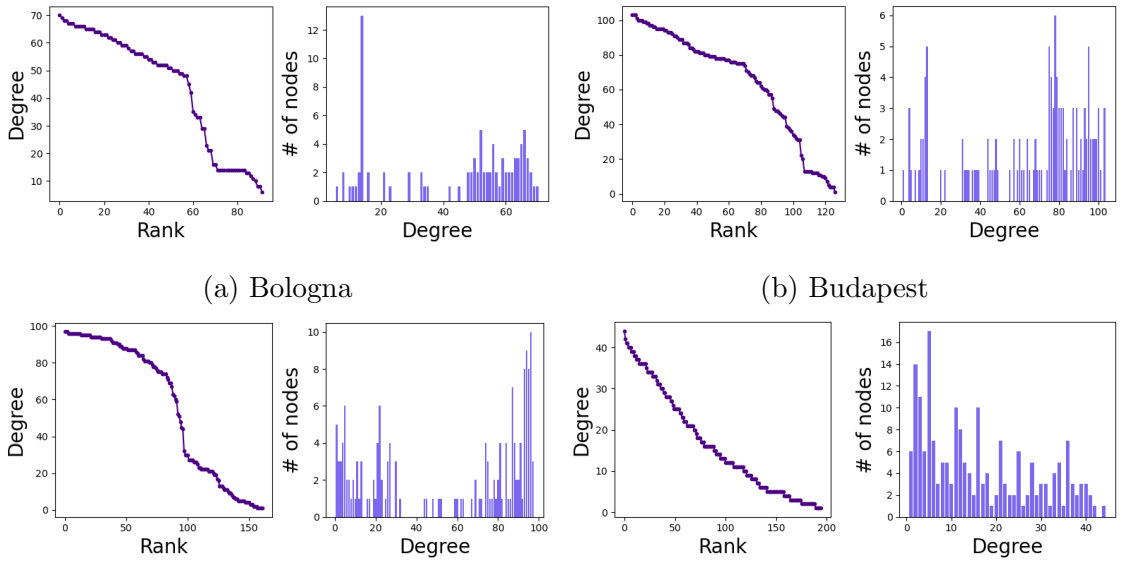
3.5 Species Correlation networks in the four cities

Given the interesting results obtained from the network analysis of the three sub-periods considered for Rome and Budapest, we decided to perform a similar analysis to the whole time series of the four cities.

Again, the nodes of the correlation networks represent all of those species having at least 10 counts in at least 75% of the samples.

In Figures 3.22 a-d, the degree rank plot and the degree histogram derived from the correlation networks of the four time series are shown.

It is interesting to point out that the time series of Rotterdam seems to be the only one with a degree distribution among the taxa that smoothly decreases in the degree rank plot, while, due to the presence of an abrupt decrease in the degree rank plot, the nodes of the other cities may be clustered into two subgroups, based on their degree.



(a) Bologna

(b) Budapest

(c) Rome

(d) Rotterdam

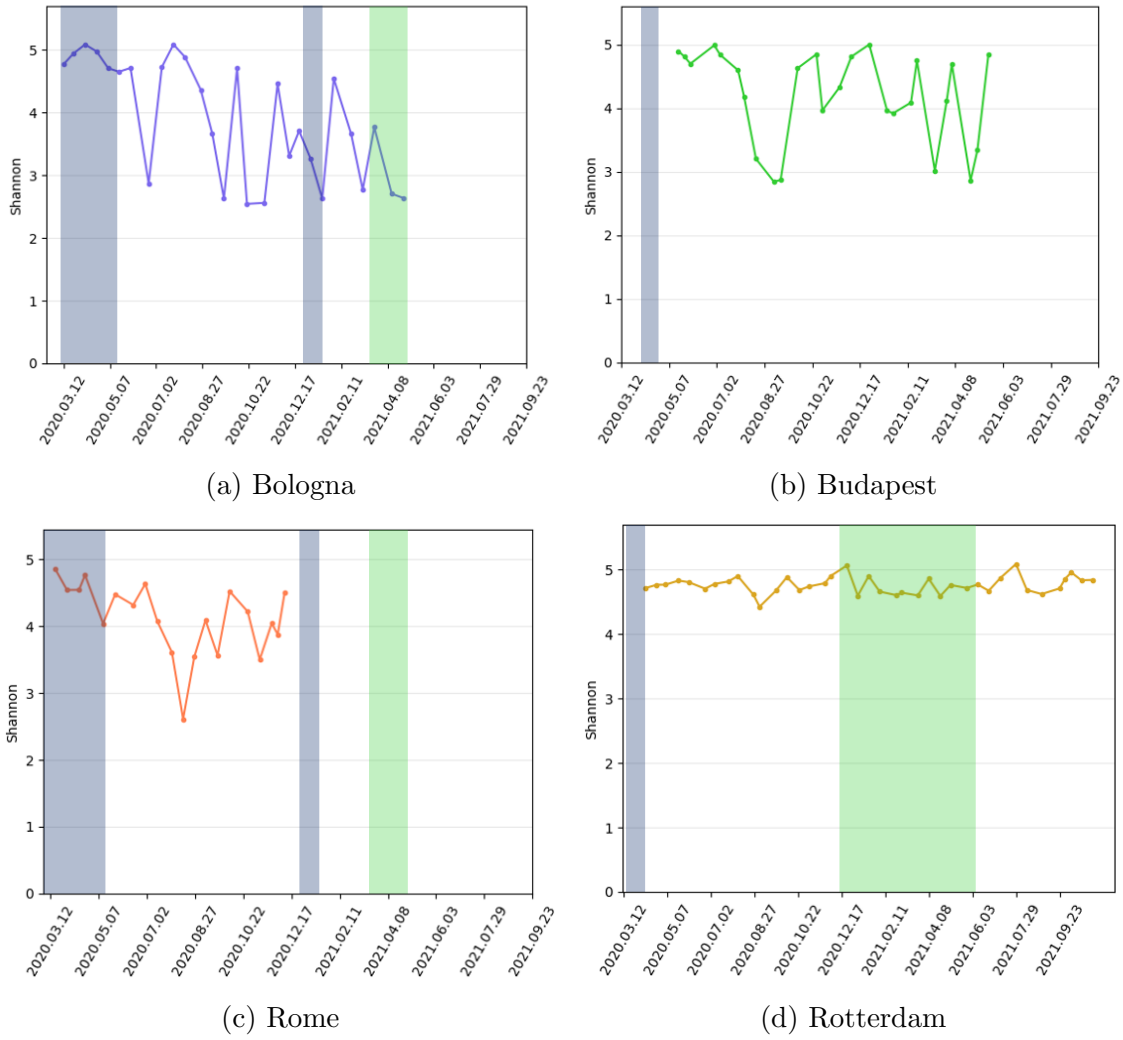
Figure 3.22: Degree rank plot and degree histogram derived from the correlation networks of the time series of the four cities

The behaviour of the degree distribution of the Rotterdam correlation network may be related to the above-mentioned interesting stability characterising the time series of the Dutch city.

3.6 Impact of COVID-19 lockdowns

Finally, a brief discussion on how the lockdowns caused by the COVID-19 pandemic may be affected the time series and the biodiversity of the bacterial content of the

sewage. Figures 3.23 a-d show the Shannon index time series for Bologna, Budapest, Rome, and Rotterdam, also highlighting, with different colours, the lockdown periods (in blue) and those periods characterised by lighter restrictions (in green).



(a) Bologna

(b) Budapest

(c) Rome

(d) Rotterdam

Figure 3.23: Time series of Shannon index in the four cities with highlighted the periods of lockdowns (in blue) and minor restrictions (in green)

It is worth pointing out that no lockdowns effects on biodiversity may be detected from the plots. This is due to two reasons, at least: first, in most cases, the time series cover a short period of time, so that some period of restrictions are not covered by the samples; second, the sampling rate is too large to appreciate any possibly significant effect due to lockdowns and changes in the lifestyle.

Within the VEO project, some other samples have been already collected but have not been sequenced yet. Once the sequenced data will be available, the analysis will be performed to these samples, too. In this way, possible relationships between the sewage metagenome compositions and the lockdowns (or even other parameters linked to the pandemic, such as the number of cases) will be investigated.

Chapter 4

Conclusions

The results of this work largely depend on the nature of the data, which comes from a major project at European level in progress.

Of course, the main strength of this work is therefore that it is a large-scale project having a great potential and vast room for improvement. However, the fact of being a project still in progress causes some tricky issues to deal with: for example, the short period of the time series does not allow the optimal use of several techniques, which instead may be applied when the studied time interval will be expanded.

Another problem is the noise characterising the data, which are absolute abundances of DNA fragments belonging to different microorganisms. The data were treated and normalised as counts, taking into account their compositional nature. However, it is clear that a normalisation based on metagenomic features (such as coverage size and length of the scaffolds) could clean up the data from the intrinsic noise that appears to be, at least partially, linked to the methodology used.

Despite these issues, some interesting results emerged from the performed analyses.

The results have highlighted different patterns for the four cities. In particular, Rotterdam has a vary stable composition, which can be noticed by looking at both the taxa compositions of the samples and the α-biodiversity indices.

On the other hand, Rome and Budapest exhibit a period with a drop in biodiversity. It is worth pointing out that this drop is linked to two concurrent aspects: the change in abundance of some of the most abundant species (such as the *Pseudomonas* spp.) and the correlation among species which is enriched during the period of minimum biodiversity.

Moreover, Bologna is characterised by noisy data and, thus, high fluctuations in the time series of both taxa and biodiversity indices.

No relevant seasonality effects are detected, as well as no consequences of the periods of lockdown due to the COVID-19 pandemic. However, it is worth highlighting that the time series does not cover all the lockdown periods and there is not

a sufficient number of samples from those periods to obtain significant results.

4.1 Future developments of the study

Of course, it is possible to imagine further and interesting developments: first of all, the possible increase in the size of the time series will allow both to apply the same techniques again on an enlarged database, so as to validate or correct some of the results found so far, and it will allow the application of other techniques, which could not be used up to now.

Among them, for instance, the Hurst exponent [32] will allow to study the long-term memory of time series, while the Lyapunov exponent [33] will allow to measure the speed of increase of small perturbations applied to the sample's composition, hence, giving a quantitative information on the predictability of the bacterial community dynamics. Both those exponents are sensitive to the sampling frequency and need the time series to be sufficiently long and uniformly sampled [34, 35].

The hope is also to expand this study integrating the data with other features that can be retrieved from metagenomic samples. In particular, it will be interesting to investigate the samples antimicrobial resistance, considering that almost all the bacterial species that have been highlighted as particularly relevant are implicated in antibiotic resistance processes [36].

Taking a look at the VEO European project with an eye to the future, it would be of great help to analyse time series data covering larger time periods, so that the results presented here may be validated, and additional ones may be revealed, in a perspective of further enhancing the power of these biological studies for an improved quality of life.

Appendix A

Data features

In this appendix, some features of the data, cited in the text, are presented summed up in tables.

Unknowns in the samples

One inevitable feature of a metagenomic sample is the presence of *unknown* taxa, i.e. of taxa which have not been identified through the binning procedure.

Table A.1 displays the average amount (in percentage) of *unknowns* for each taxonomic level.

Taxonomic level	**Average unknowns %**
Superkingdom	0.01
Phylum	0.94
Class	2.23
Order	3.34
Family	8.64
Genus	12.43
Species	33.57

Table A.1: Average percentage of unknowns in the metagenomic dataset

Going down with the ranks, the (relative) amounts of DNA fragments not identified increases, as expected.

It is important to underline that the percentage abundances shown in the table are referred to the complete and not-yet-filtered sample, i.e. the taxonomic reconstruction used to maintain only bacteria has not been performed yet.

Lost taxa

During the preprocessing step, as well as during some successive analyses, a taxonomic reconstruction of taxa is required, especially in order to filter the population taking only bacteria.

During this reconstruction, some taxa have not been found in the NCBI database, so they have been considered as lost.

In Table A.2, a list of taxa lost during the taxonomic reconstruction through the NCBI database. Obviously, the superkingdom level is omitted since no reconstruction is needed.

Taxonomic level	**Lost taxa %**
Phylum	0.5
Class	0
Order	0.92
Family	0.1
Genus	0.14
Species	0.12

Table A.2: Percentage of lost taxa in the metagenomic dataset

It is worth pointing out that some of the lost taxa, belonging to the taxonomic rank of species, genus and family and order, have been found in the Silva database [37]. However, they all belong to the Eukaryota superkingdom, thus they are removed from the samples during the filtering phase.

Appendix B

Simplex and Aitchison simplex

In order to properly describe the differences between a general simplex and the Aitchison simplex, some mathematical concepts should be defined.

First of all, a polytope is the generalisation of three-dimensional polyhedra to any number of dimensions. Its precise definition strongly depends on the field of study; this results in many not-equivalent nor consistent definitions. The most general one allows a polytope to be open, close, self-intersecting, bounded or unbounded.

A simplex is the simplest possible polytope made with line segments and it results in a generalisation to any dimensions of triangles. In particular, a k-simplex is a k-dimensional polytope, which is the convex hull[1] of its $k+1$ vertices.

Suppose to have $k+1$ points $x_0, \dots, x_k \in \mathbb{R}^N$ which are affinely independent, i.e. that $x_1 - x_0, \dots, x_k - x_0$ are linearly independent. The mathematical definition of a simplex is given by:

$$\mathcal{C} = \mathbf{conv}\{x_0, \dots, x_k\} = \left\{\theta_0 x_0 + \cdots + \theta_k x_k \middle|\, \theta_i \geq 0,\ i = 0, \dots, k;\ \mathbf{1}^T\theta = 1\right\} \quad \text{(B.1)}$$

where $\mathbf{1}$ is the k-dimensional vector with all entries equal to 1.

The Aitchison simplex is the particular case of a simplex where the $k+1$ vectors above described are the unit vectors of $\mathbb{R}^{k+1}$. This leads to the definition of probability simplex [38], described below in $\mathbb{R}^k$, for the sake of simplicity.

The probability simplex is the $(k-1)$-dimensional simplex determined by the k unit vectors $e_1, \dots, e_k \in \mathbb{R}^k$, i.e.:

$$\mathcal{S} = \left\{x_1 + \cdots + x_k \middle|\, x_i \geq 0,\ i = 1, \dots, k;\ \mathbf{1}^T\mathbf{x} = 1\right\} \quad \text{(B.2)}$$

Vectors in the probability simplex correspond to probability distributions on a set with k elements, in which x_i is the probability of the i^{th} element.

The probability simplex is not the Aitchison simplex yet. In fact, the name Aitchison simplex is used in this dissertation to refer to a probability simplex with the following properties and operations:

[1] The *convex hull* of a geometrical shape is the (unique) minimal convex set containing it.

- Any operation on compositional data (points on the simplex) must be expressed by scale invariant functions of the components. These functions must be real and 0-degree homogeneous, i.e. the condition

$$f(\lambda \mathbf{x}) = f(\mathbf{x}) \tag{B.3}$$

is satisfied for any positive constant λ and for any point $\mathbf{x}$ on the simplex. This is the scale invariant analysis principle [39].

- Given $\mathbf{x}$ and $\mathbf{y}$, two data points in the $(k-1)$-dimensional simplex, a perturbation operator $\circ$ is defined:

$$\mathbf{x} \circ \mathbf{y} = \mathscr{C}(x_1 y_1, \ldots, x_k y_k) = \left(\frac{x_1 y_1}{\sum_{i=1}^{k} x_i y_i}, \ \ldots \ , \frac{x_k y_k}{\sum_{i=1}^{k} x_i y_i} \right) \ . \tag{B.4}$$

- An isomorphic transform that map a simplex $\mathcal{S}$ into the real space $\mathbb{R}^k$ called centred log-ratio (or centre log-ratio) is defined:

$$\begin{gathered} \mathrm{clr}: \ S \longrightarrow U, \ \ U \subset \mathbb{R}^k \\ \mathrm{clr}(\mathbf{x}) = \left(\log \frac{x_1}{g(\mathbf{x})}, \ \ldots \ , \log \frac{x_k}{g(\mathbf{x})} \right) \end{gathered} \tag{B.5}$$

where $g(\mathbf{x})$ is the geometric mean of the k-dimensional vector $\mathbf{x}$.

- The boundary of the simplex B.2 is excluded in the Aitchison simplex, i.e.:

$$\mathcal{S} = \left\{ x_1 + \cdots + x_k |\, x_i > 0, \, i = 1, \ldots, k; \ \mathbf{1}^T \mathbf{x} = 1 \right\} \tag{B.6}$$

As a consequence of the scale invariant analysis principle, simplex of the form[2]:

$$\mathcal{S} = \left\{ x_1 + \cdots + x_k |\, x_i > 0, \, i = 1, \ldots, k; \ \mathbf{1}^T \mathbf{x} = \alpha \right\} \tag{B.7}$$

may be transformed in probability simplex as in B.2. The operation to use is called closure and is defined as follows:

$$\mathscr{C}(x_1, \ldots, x_k) = \left(\frac{x_1}{\sum_{i=1}^{k} x_i}, \ \ldots \ , \frac{x_k}{\sum_{i=1}^{k} x_i} \right) \ . \tag{B.8}$$

Other operations could be defined on a simplex but are not included in the appendix so that it will not be weighed down.

[2]Notice that a simplex like this may be found in many practical problems. The counts of DNA fragments in a sample is a case example.

Appendix C

Proof: Hill1 coincides with the exponential of the Shannon index

Let us consider S positive numbers p_i, for $i = 1, \ldots, S$, such that $\sum_{i=1}^{S} p_i = 1$ and

$$^1D = \left(\sum_{i=1}^{S} p_i^q \right)^{\frac{1}{1-q}}$$

The aim is to prove that ^1D is continuous with derivatives of all orders at $q = 1$ and that

$$^1D = \lim_{q \to 1} \left(\sum_{i=1}^{S} p_i^q \right)^{\frac{1}{1-q}} = \exp\left(-\sum_{i=1}^{S} p_i \ln p_i \right) \tag{C.1}$$

Setting $q = 1 + b$ and applying the logarithm on both sides, it will be sufficient to prove that

$$\lim_{b \to 0} \frac{1}{b} \ln \sum_{i=1}^{S} p_i^{1+b} = \sum_{i=1}^{S} p_i \ln p_i$$

Let us consider only the left hand side of this equation, that can be re-written as

$$\lim_{b \to 0} \frac{1}{b} \ln \sum_{i=1}^{S} p_i p_i^b = \lim_{b \to 0} \frac{1}{b} \ln \sum_{i=1}^{S} p_i \exp(b \ln p_i) \ .$$

Now, remembering that, for small values of x, $\exp x \simeq 1 + x$ holds, the following expression may be obtained:

$$\lim_{b \to 0} \frac{1}{b} \ln \sum_{i=1}^{S} p_i \exp(b \ln p_i) = \lim_{b \to 0} \frac{1}{b} \ln \left(\sum_{i=1}^{S} p_i + b \sum_{i=1}^{S} p_i \ln p_i \right)$$

Now, observing that $\sum_{i=1}^{S} p_i = 1$ and using the approximation, for small values of x, $\ln(1 + x) \simeq x$, last equation may be re-written:

$$\lim_{b\to 0}\frac{1}{b}\ln\left(\sum_{i=1}^{S}p_i+b\sum_{i=1}^{S}p_i\ln p_i\right)=\lim_{b\to 0}\frac{1}{b}\cdot b\cdot\sum_{i=1}^{S}p_i\ln p_i=\sum_{i=1}^{S}p_i\ln p_i$$

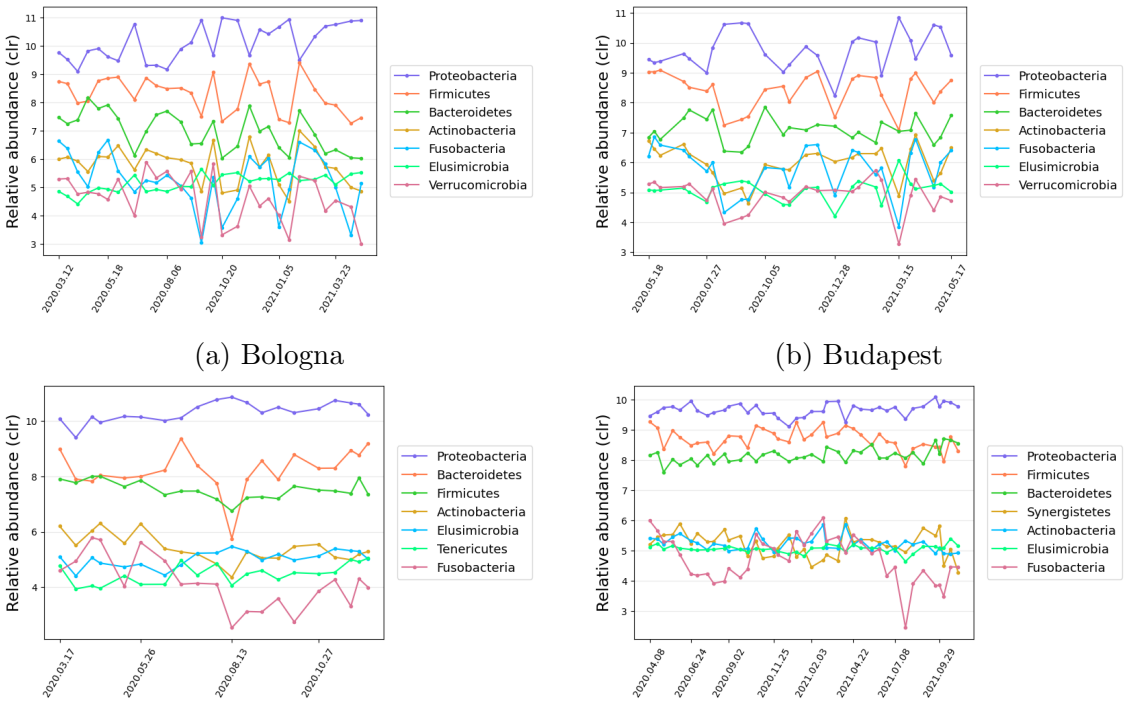
In order to prove the continuity of the derivatives of ${}^1D$, Let us start from the following consideration: suppose $f(x), g(x)$ to be two functions that can be expanded as power series in a neighbourhood of $x = 0$ and let $f(0) = 0$ and $g(0) = 0$; then, if $f(x)/g(x)$ is continuous in $x = 0$ follows that it also has derivatives of all orders, and they can be expanded as a power series. In the considered case, $g(x)$ is b and $f(x)$ is $\sum_{i=1}^{S} p_i \ln p_i$. Thus, ${}^1D$ has continuous derivatives.

Appendix D

Supplementary material

In this appendix, the plots and outcomes mentioned in the text are reported. In particular, the time series of the most abundant taxa, as well as the auto- and cross-correlations of the most relevant phyla are shown. In the final section, a complete list of the significant outcomes of the Fisher test for the evaluation of the significance of periodicities are reported, summarised in a table.

Time series of most abundant taxa



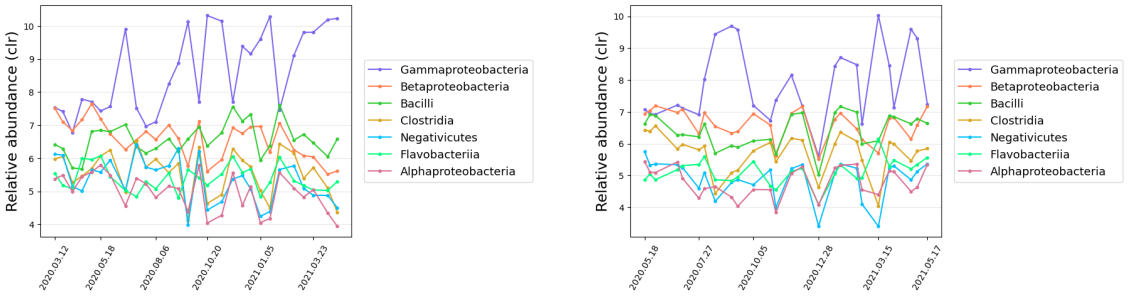
(a) Bologna

(b) Budapest

(c) Rome

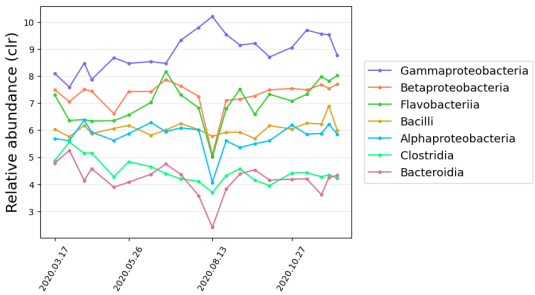
(d) Rotterdam

Figure D.1: Time series of the 7 most abundant phyla in the four cities; abundances expressed in centred log-ratio

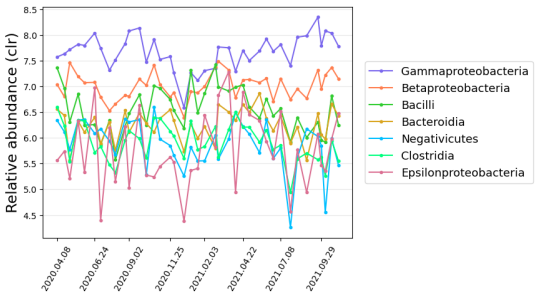


(a) Bologna

(b) Budapest

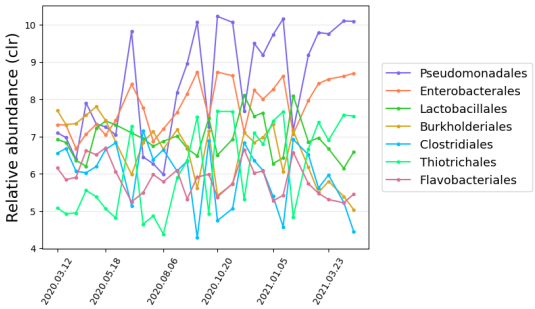


(c) Rome

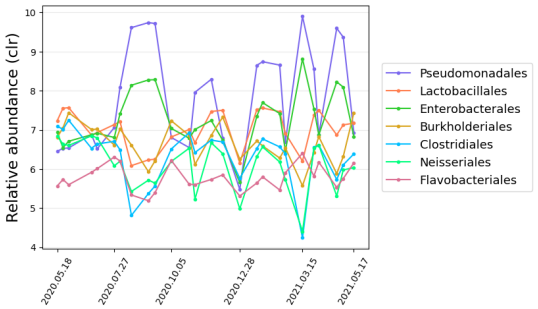


(d) Rotterdam

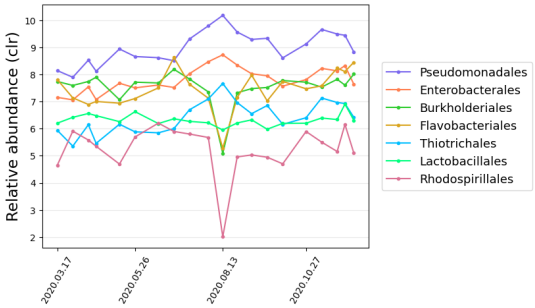
Figure D.2: Time series of the 7 most abundant classes in the four cities; abundances expressed in centred log-ratio



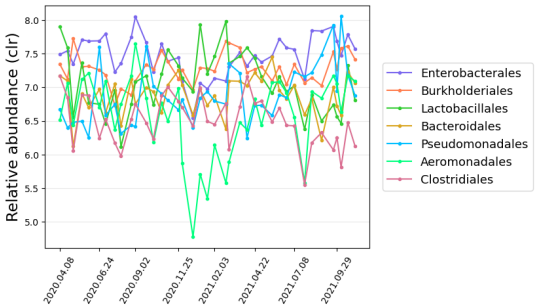
(a) Bologna



(b) Budapest

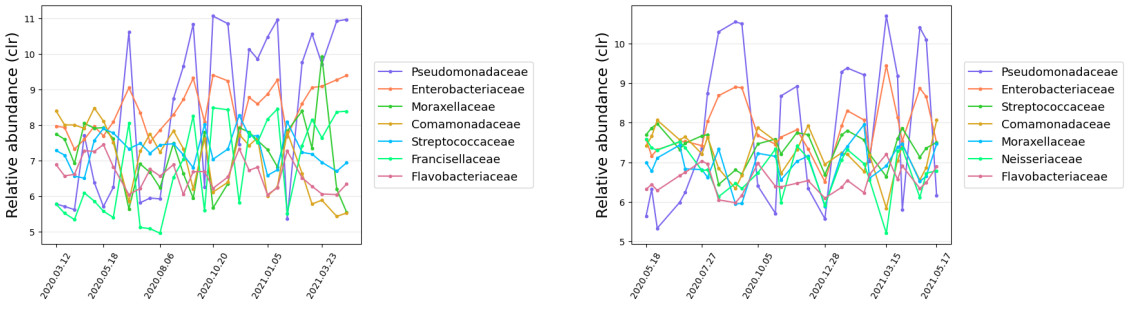


(c) Rome



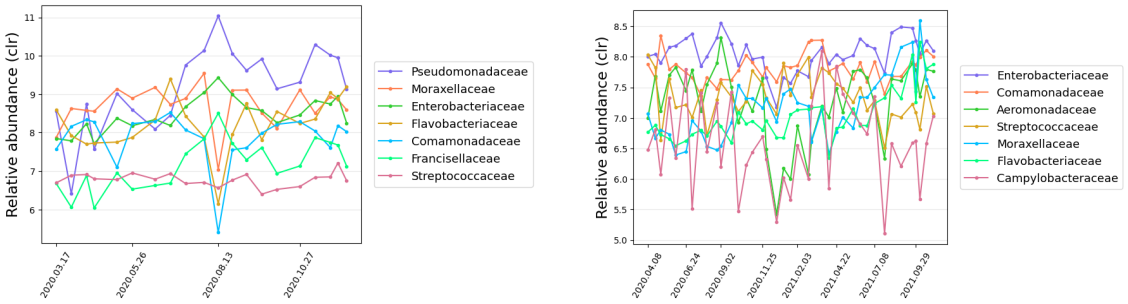
(d) Rotterdam

Figure D.3: Time series of the 7 most abundant orders in the four cities; abundances expressed in centred log-ratio



(a) Bologna

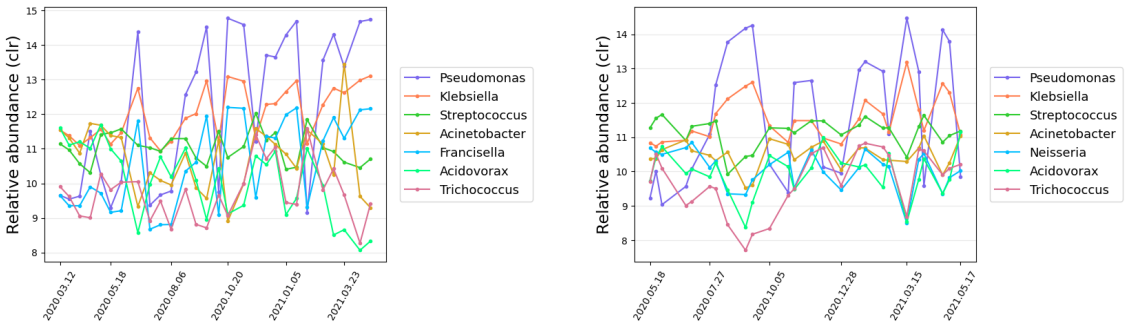
(b) Budapest



(c) Rome

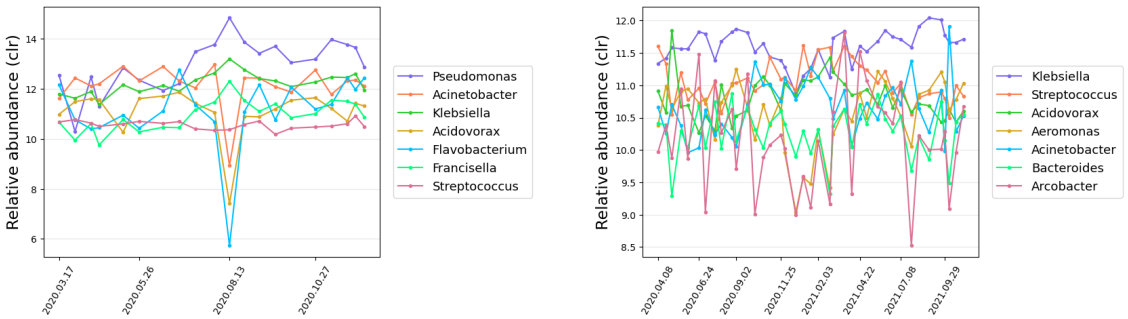
(d) Rotterdam

Figure D.4: Time series of the 7 most abundant families in the four cities; abundances expressed in centred log-ratio



(a) Bologna

(b) Budapest

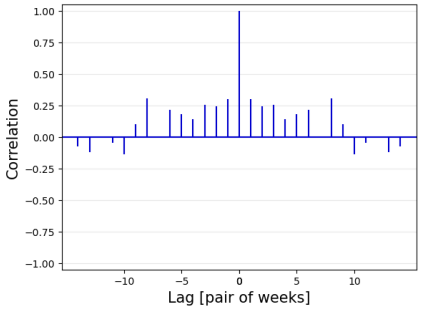


(c) Rome

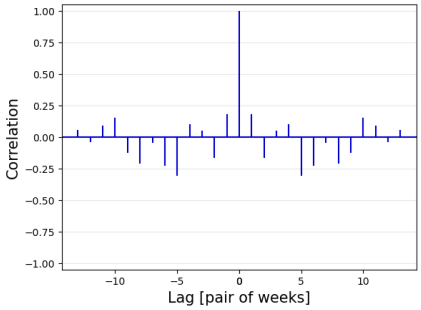
(d) Rotterdam

Figure D.5: Time series of the 7 most abundant genera in the four cities; abundances expressed in centred log-ratio

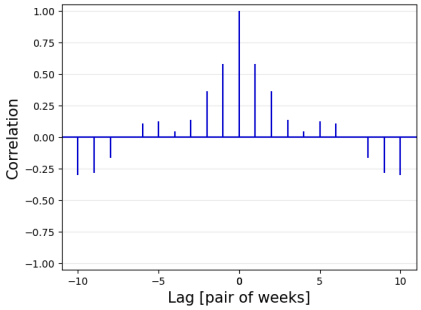
Auto-correlations of phyla



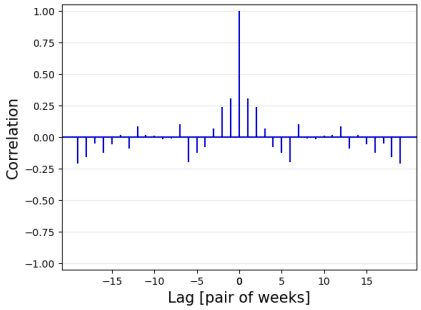
(a) Bologna



(b) Budapest

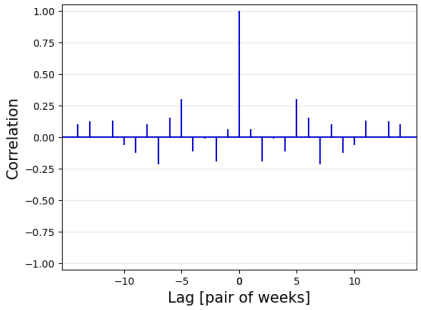


(c) Rome

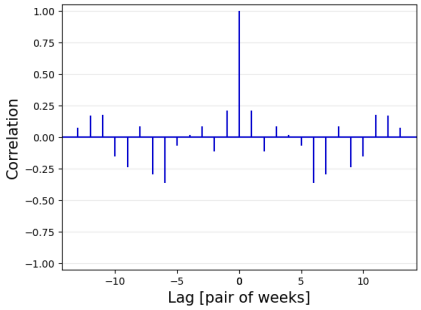


(d) Rotterdam

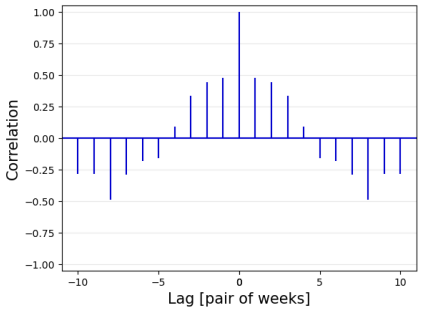
Figure D.6: Auto-correlations of the the *Proteobacteria* phylum in the four cities



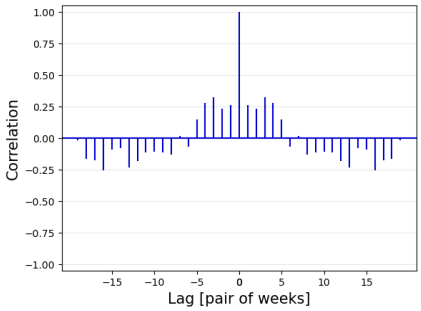
(a) Bologna



(b) Budapest

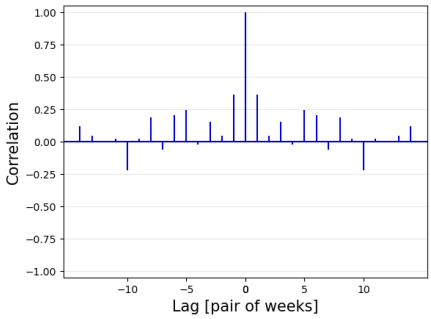


(c) Rome

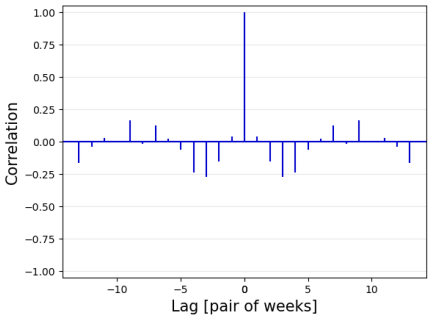


(d) Rotterdam

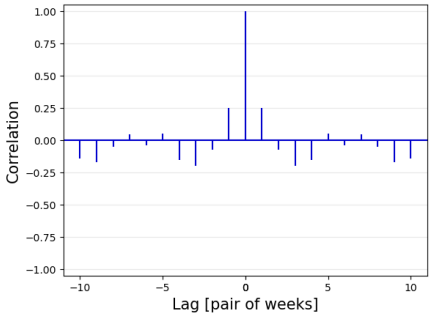
Figure D.7: Auto-correlations of the *Firmicutes* phylum in the four cities



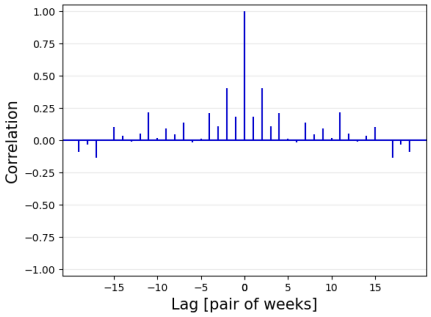
(a) Bologna



(b) Budapest

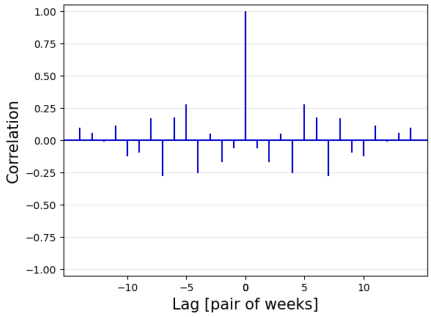


(c) Rome

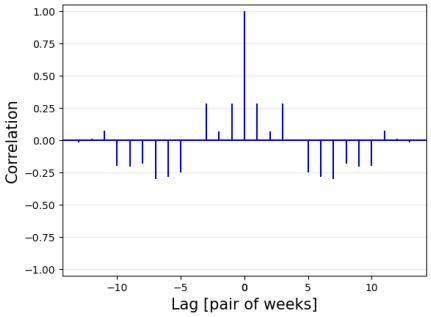


(d) Rotterdam

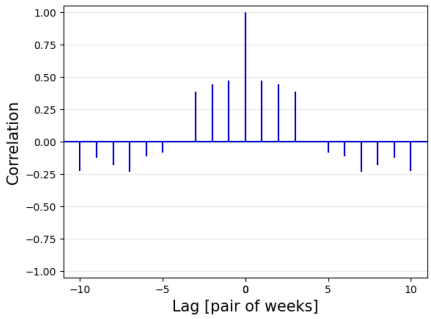
Figure D.8: Auto-correlations of the *Bacteroidetes* phylum in the four cities



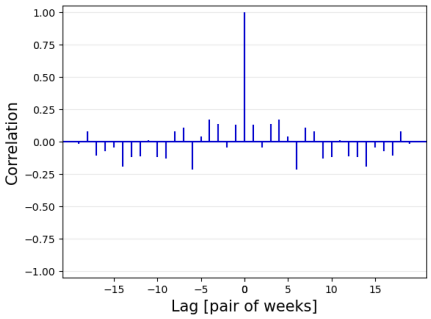
(a) Bologna



(b) Budapest



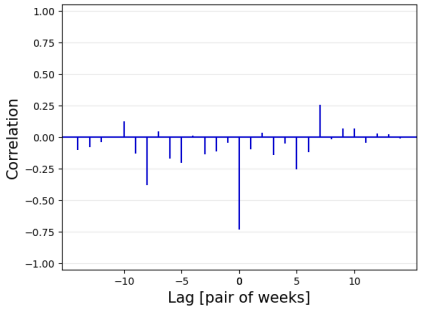
(c) Rome



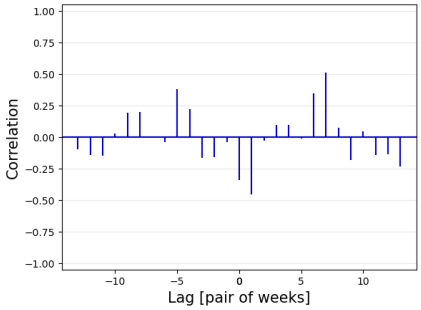
(d) Rotterdam

Figure D.9: Auto-correlations of the *Actinobacteria* phylum in the four cities

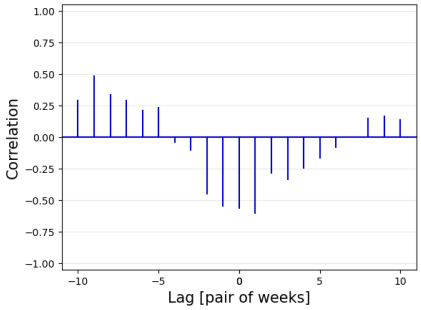
Cross-correlations of phyla



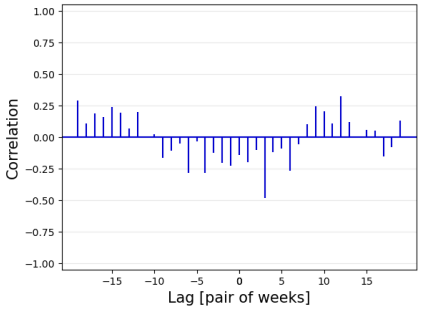
(a) Bologna



(b) Budapest

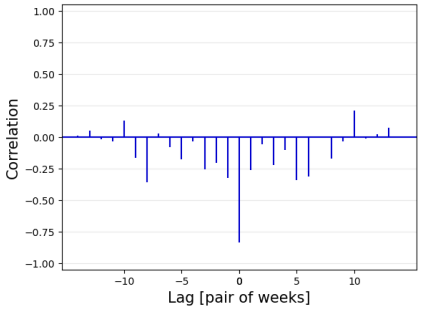


(c) Rome

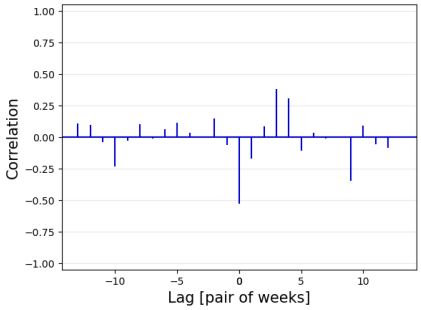


(d) Rotterdam

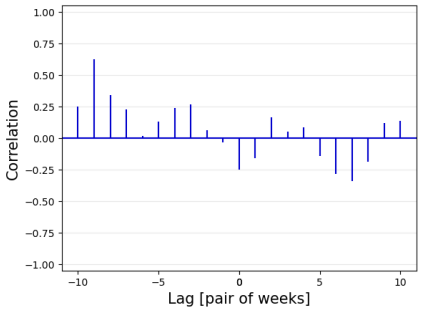
Figure D.10: Cross-correlations of *Proteobacteria* and *Firmicutes* in the four cities



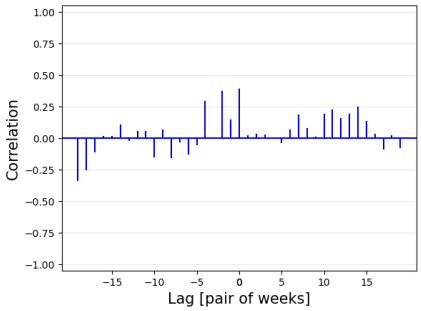
(a) Bologna



(b) Budapest

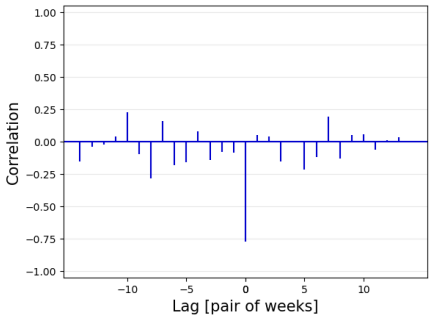


(c) Rome

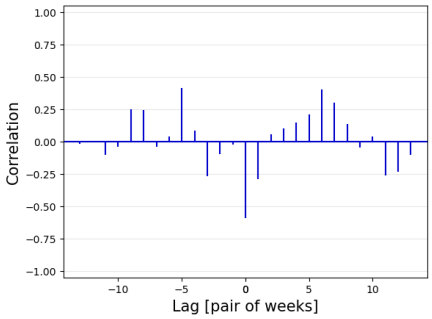


(d) Rotterdam

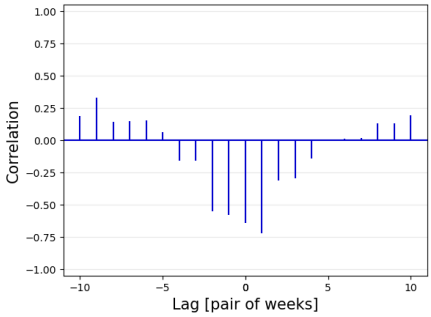
Figure D.11: Cross-correlations of *Proteobacteria* and *Bacteroidetes* in the four cities



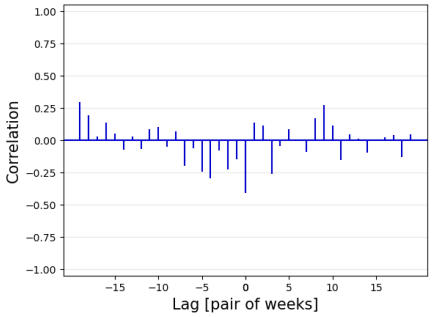
(a) Bologna



(b) Budapest

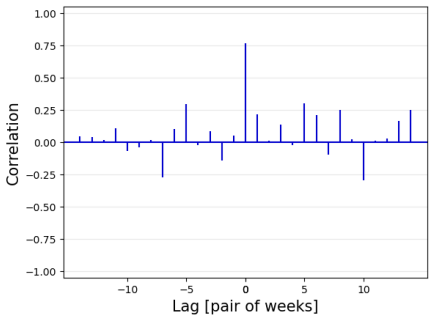


(c) Rome

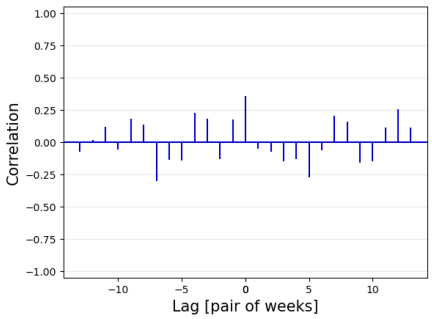


(d) Rotterdam

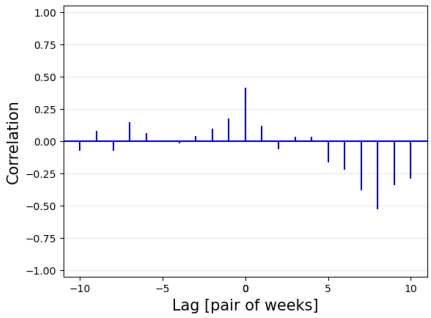
Figure D.12: Cross-correlations of *Proteobacteria* and *Actinobacteria* in the four cities



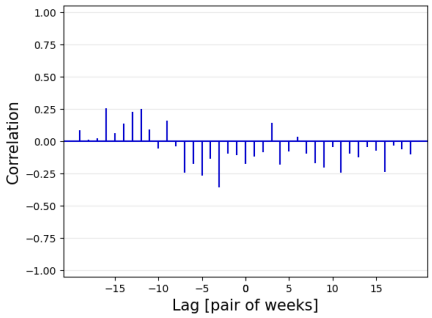
(a) Bologna



(b) Budapest

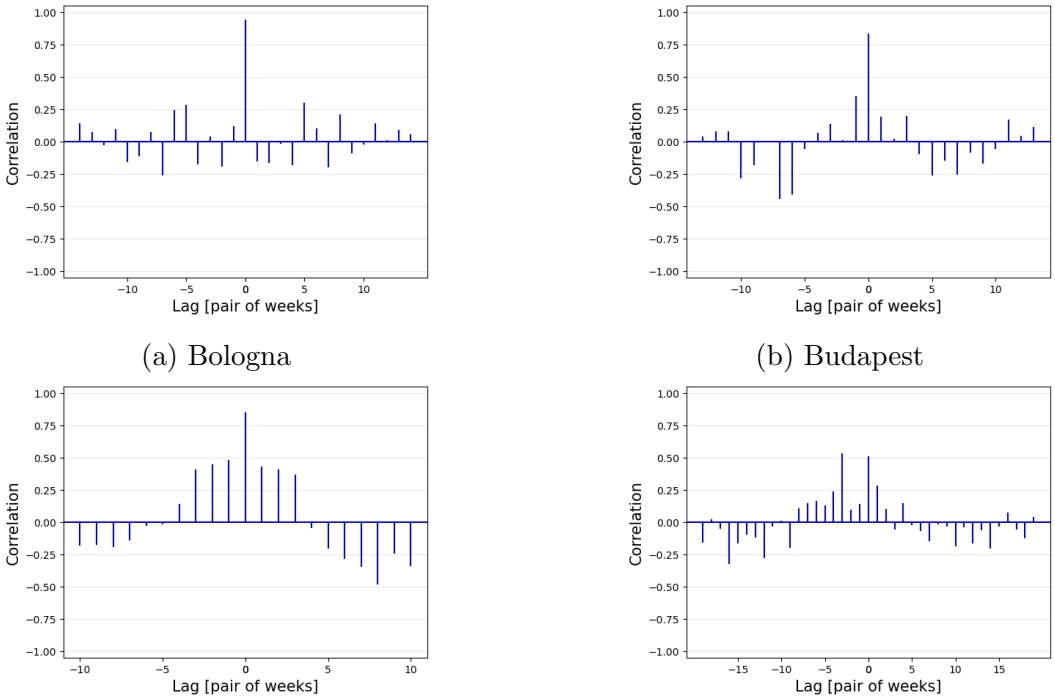


(c) Rome



(d) Rotterdam

Figure D.13: Cross-correlations of *Firmicutes* and *Bacteroidetes* in the four cities



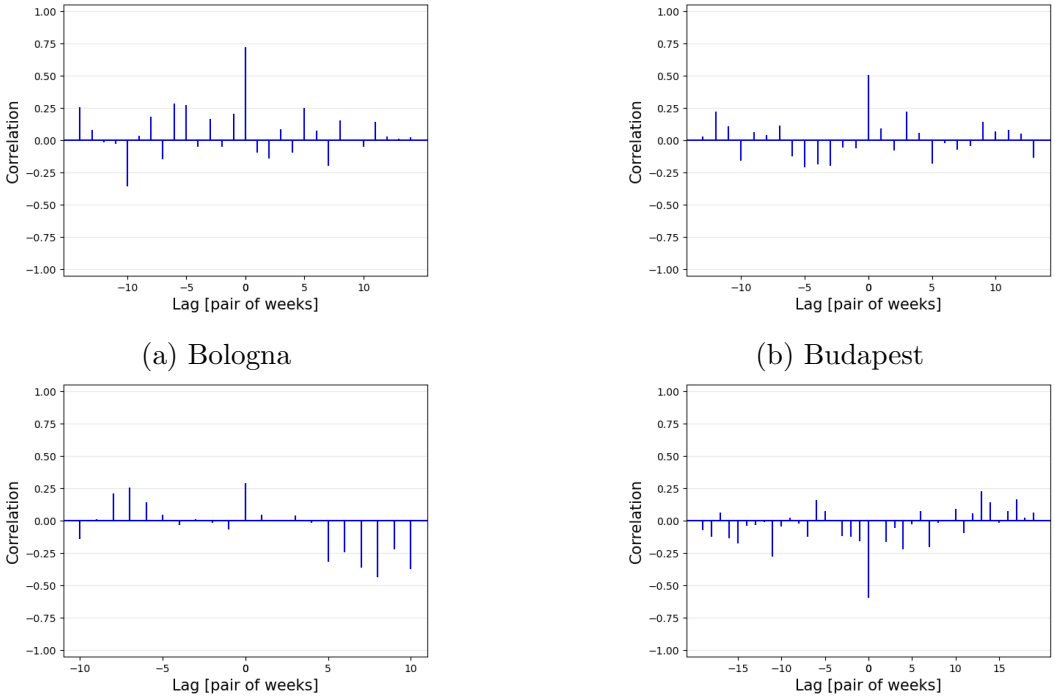
(a) Bologna

(b) Budapest

(c) Rome

(d) Rotterdam

Figure D.14: Cross-correlations of *Firmicutes* and *Actinobacteria* in the four cities



(a) Bologna

(b) Budapest

(c) Rome

(d) Rotterdam

Figure D.15: Cross-correlations of *Bacteroidetes* and *Actinobacteria* in the four cities

Significant peaks in the Fourier spectrum

Here, a table collecting the outcomes of the Fisher's test for the significant peaks of the time series is shown for the following taxonomic levels: phylum, class, order, family, genus.

Taxon. level	Taxon	City	Period	Significance
Phylum	Elusimicrobia	Bologna	56	*
	Proteobacteria	Rome	40	**
	Firmicutes	Rome	40	**
	Actinobacteria	Rome	40	*
	Fusobacteria	Rome	40	**
	Fusobacteria	Rotterdam	39	**
Class	Gammaproteobacteria	Rome	40	*
	Clostridia	Rome	40	*
	Gammaproteobacteria	Rotterdam	78	*
Order	Pseudomonadales	Rome	40	*
	Enterobacterales	Rome	40	*
	Thiotrichales	Rome	40	*
	Enterobacterales	Budapest	26	*
	Enterobacterales	Rotterdam	78	*
	Aeromonadales	Rotterdam	78	*
Family	Pseudomonadaceae	Rome	40	*
	Enterobacteriaceae	Rome	40	*
	Francisellaceae	Rome	40	*
	Enterobacteriaceae	Budapest	26	*
	Moraxellaceae	Budapest	7	*
	Enterobacteriaceae	Rotterdam	78	*
	Comamonadaceae	Rotterdam	39	*
	Aeromonadaceae	Rotterdam	78	*
	Moraxellaceae	Rotterdam	39	**
	Flavobacteriaceae	Rotterdam	78	*

Table D.1: [1/2] Fisher test for periodicity (* means a p-value < 0.05, while ** means a p-value < 0.01); the period corresponding to the peak is expressed in weeks

Taxon. level	**Taxon**	**City**	**Period**	**Significance**
Genus	Pseudomonas	Rome	40	*
	Klebsiella	Rome	40	*
	Francisella	Rome	40	*
	Klebsiella	Budapest	26	*
	Acinetobacter	Budapest	7	*
	Trichococcus	Budapest	52	**
	Streptococcus	Rotterdam	78	*
	Aeromonas	Rotterdam	78	*

Table D.2: [2/2] Fisher test for periodicity (* means a p-value < 0.05, while ** means a p-value < 0.01); the period corresponding to the peak is expressed in weeks

Bibliography

[1] H. Winkler. *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche*. Jena: Verlag Fischer, 1920.

[2] C.R. Woese and G.E. Fox. "Phylogenetic structure of the prokaryotic domain: the primary kingdoms". In: *Proceedings of the National Academy of Sciences of the United States of America* 74.11 (1977), pp. 5088–5090. DOI: 10.1073/pnas.74.11.5088.

[3] D.J. Lane et al. "Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses". In: *Proceedings of the National Academy of Sciences of the United States of America* 82.20 (1985), pp. 6955–6959. DOI: 10.1073/pnas.82.20.6955.

[4] N.R. Pace et al. "The analysis of natural microbial populations by ribosomal RNA sequences". In: *Advances in Microbial Ecology* 9 (1986), pp. 1–55. DOI: 10.1007/978-1-4757-0611-6_1.

[5] World Health Organisation. *Tripartite and UNEP support OHHLEP's definition of One Health*. Last access: 2022-11-25. URL: www.who.int/news/item/01-12-2021-tripartite-and-unep-support-ohhlep-s-definition-of-one-health.

[6] K. Faust et al. "Metagenomics meets time series analysis: unraveling microbial community dynamics". In: *Current Opinion in Microbiology* 25 (2015), pp. 56–66. DOI: 10.1016/j.mib.2015.04.004.

[7] *National Center for Biotechnology Information*. Last access: 2022-11-25. URL: www.ncbi.nlm.nih.gov.

[8] *GitHub repository: taxonomy_ranks*. Get taxonomy ranks information with ETE3 Python3 module, Last access: 2022-11-25. URL: www.github.com/linzhi2013/taxonomy_ranks.

[9] J. Aitchison. "The Statistical Analysis of Compositional Data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2 (1982), pp. 139–160. DOI: 10.1111/j.2517-6161.1982.tb01195.x.

[10] J.A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn. "Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation". In: *Mathematical Geology* 35.3 (2003), pp. 253–278. DOI: 10.1023/A:1023866030544.

[11] M.S. Bartlett. "On the theoretical specification of sampling properties of autocorrelated time series". In: *Journal of Royal Statistical Society, Series B* 8.1 (1946), pp. 27–41. DOI: 10.2307/2983611.

[12] G.E.P. Box and G.M. Jenkins. *Time series analysis: Forecasting and control*. San Francisco: Holden-Day, 1976.

[13] James W. Cooley and John W. Tukey. "An algorithm for the machine calculation of complex Fourier series". In: *Mathematics of Computation* 19 (1965), pp. 297–301.

[14] R.A. Fisher. "Tests of significance in harmonic analysis". In: *Proceedings of the Royal Society A* 125.796 (1929), pp. 54–59. DOI: 10.1098/rspa.1929.0151.

[15] E.O. Wilson. *Biodiversity*. Washington: National Academy Press, 1988.

[16] C.E. Shannon. "A mathematical theory of communication". In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[17] E.C. Pielou. "The Measurement of Diversity in Different Types of Biological Collections". In: *Journal of Theoretical Biology* 13 (1966), pp. 131–144. DOI: 10.1016/0022-5193(66)90013-0.

[18] E.H. Simpson. "Measurement of Diversity". In: *Nature* 163 (1949), p. 688. DOI: 10.1038/163688a0.

[19] C.W. Gini. "Variabilità e mutabilità". In: *Studi Economico-Giuridici della R. Università di Cagliari* 3 (1912), pp. 3–159.

[20] M. O. Hill. "Diversity and evenness: a unifying notation and its consequences". In: *Ecology* 54.2 (1973), pp. 427–432. DOI: 10.2307/1934352.

[21] A. Chao. "Non-parametric estimation of the number of classes in a population". In: *Scandinavian Journal of Statistics* 11.4 (1984), pp. 265–270. DOI: 10.2307/4615964.

[22] K.R. Clarke and R.M. Warwick. "A taxonomic distinctness index and its statistical properties". In: *Journal of Applied Ecology* 35.4 (1998), pp. 523–531. DOI: 10.1046/j.1365-2664.1998.3540523.x.

[23] J.R. Bray and J.T. Curtis. "An Ordination of the Upland Forest Communities of Southern Wisconsin". In: *Ecological Monographs* 27.4 (1957), pp. 325–349. DOI: 10.2307/1942268.

[24] J.C. Gower. "Some distance properties of latent root and vector methods used in multivariate analysis". In: *Biometrika* 53.3-4 (1966), pp. 325–338. DOI: 10.1093/biomet/53.3-4.325.

[25] D.A. Dickey and Fuller W.A. "Distribution of the Estimators for Autoregressive Time Series with a Unit Root". In: *Journal of the American Statistical Association* 74.366a (1979), pp. 427–431. DOI: 10.1080/01621459.1979.10482531.

[26] W.A. Fuller. *Introduction to Statistical Time Series*. Wiley, New York, 1995.

[27] H. Levene. "Robust Tests for Equality of Variances". In: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Ed. by I. Olkin et al. Vol. 2. Stanford studies in mathematics and statistics. Palo Alto: Stanford University Press, 1960, pp. 278–292.

[28] M.B. Brown and A.B. Forsythe. "Robust tests for the equality of variances". In: *Journal of the American Statistical Association* 69.346 (1974), pp. 364–367. DOI: 10.1080/01621459.1974.10482955.

[29] *GitHub repository - MD_Thesis-suppl_mat*. Supplementary material of "Characterisation of the sewage microbiome in four European cities throughout the COVID-19 pandemic by means of ecological and network modelling". URL: www.github.com/Ettore1024/MD_Thesis-suppl_mat.

[30] W. Li et al. "Functional potential differences between Firmicutes and Proteobacteria in response to manure amendment in a reclaimed soil". In: *Canadian Journal of Microbiology* 66.12 (2020), pp. 689–697. DOI: 10.1139/cjm-2020-0143.

[31] S. Moossavi et al. "Composition and Variation of the Human Milk Microbiota Are Influenced by Maternal and Early-Life Factors". In: *Cell Host Microbe* 25.2 (2019), pp. 324–335. DOI: 10.1016/j.chom.2019.01.011.

[32] H.E. Hurst. "Long-term storage capacity of reservoirs". In: *Transactions of the American Society of Civil Engineers* 116 (1951), pp. 770–799.

[33] R. Brown, P. Bryant, and H. Abarbanel. "Computing the Lyapunov spectrum of a dynamical system from an observed time series". In: *Physical Review A* 43.6 (1991), pp. 2787–2806. DOI: 10.1103/PhysRevA.43.2787.

[34] M. Resta. "Hurst exponent and its applications in time-series analysis". In: *Recent Patents on Computer Science* 5.3 (2012), pp. 211–219. DOI: 10.2174/2213275911205030211.

[35] M. Dlask and J. Kukal. "Hurst exponent estimation from short time series". In: *Signal, Image and Video Processing* 13 (2019), pp. 263–269. DOI: 10.1007/s11760-018-1353-2.

[36] Istituto Superiore di Sanità. *AR-ISS: sorveglianza nazionale dell'Antibiotico-Resistenza - Dati 2021*. Last access: 2022-11-25. URL: www.epicentro.iss.it/antibiotico-resistenza/ar-iss/RIS-1_2021.pdf.

[37] *Silva - high quality ribosomal RNA databases*. Last access: 2022-11-25. URL: www.arb-silva.de.

[38] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[39] B. S. Daya Sagar, Q. Cheng, and F. Agterberg. *Handbook of Mathematical Geosciences*. Springer Open, 2018.