COVID-19 epidemiological data in Croatia and central Europe

Stilinović, Roč

UNIVERSITY OF ZAGREB

FACULTY OF SCIENCE

DEPARTMENT OF MATHEMATICS

Roč Stilinović

COVID-19 EPIDEMIOLOGICAL DATA IN CROATIA AND CENTRAL EUROPE

Diploma thesis

Thesis mentor:
doc. dr. sc. Pavle Goldstein

Zagreb, July 10, 2024

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Kome sve reć fala, koga da izdvojim
Ajde za početak, roditeljima mojim
Koji od početka, beskrajno me vole
Podršku mi daju, pamet mi ne sole

Celoj familiji, od sestre do baka
S njima i teška, vremena su laka
Kolegama svima, od Vanne do Franje
Što smo međusobno, širili si znanje

Frendovima mnogim, od Marka do Ane
Jer u dane kišne, s njima sunce svane
Vaterpolo klapi, uz njih čovjek shvati
Da se iz poraza, svatko jači vrati

Profačima brojnim, mentoru najviše
Što neznanja mene, oni časkom liše
I za kraj sam sebi, kaj uporan sam bio
Ostvarivši sve kaj, osam let sam htio

Contents

Contents **iv**

Introduction **2**

1 Data and terminology **3**
1.1 Main data 3
1.2 Additional data 3
1.3 Terminology 4

2 Mathematical methods **5**
2.1 Probability Theory 5
2.2 Descriptive statistics 9
2.3 Moving average 12

3 Results **13**
3.1 EMR 13
3.2 DIR-DDR 15
3.3 DDR-EMR 20

4 Comments and discussion **25**
4.1 DIR-DDR analysis 25
4.2 DDR-EMR analysis 27

Bibliography **29**

Introduction

At the end of 2019, a new virus named SARS-CoV-2 emerged, infecting humans and quickly spreading from China to other parts of the world. This rapid transmission led to a serious situation that soon escalated into a global pandemic known as Covid-19. In this paper, we conduct a comparative analysis using statistical data to examine the behavior of this virus in five countries: Croatia, Germany, Italy, Slovenia, and Austria, with a focus on the second and fourth waves of infection. Special emphasis is placed on the relationship between the number of infections and the number of deaths, as well as the impact of Covid-19 fatalities on overall mortality rates in each of these countries. Motivation for this study stems from the features present in Figure 1.

The figure shows the number of infections per million from January 27, 2020, to September 18, 2021. The starting point corresponds to the date of the first confirmed Covid-19 case in Germany, while the ending point is set 600 days later. This time frame includes multiple waves of infection, with particular focus on the second and fourth waves. For subsequent analysis, the end point will be adjusted to include data for the entire fourth wave of the infection.

We have observed one very interesting phenomenon. We can see that all the waves start and end at almost the same time point while the heights differ significantly from each other.

Different heights might suggest the fact that each country had different approach to deal with the epidemic (e.g., more or less strict lockdowns) or that the people followed recommendations to a different degree depending on the country.

As for the same wave beginnings and ends, we know that during the second infection wave there existed some travel restrictions (e.g., valid Covid certificates) between countries. So, the only reasonable explanation for this phenomenon is that those travel restrictions were inadequate. Also, it suggests that the infection rate in each country was boosted by infections from some common, mostly European pool. Higher influx corresponded with increased infections, whereas decline in influx resulted in reductions in infection numbers.

For our analysis we have chosen three main sets of data; DIR, DDR and MDR. Our objective was to address several questions arising naturally from the data. What was the impact of restrictions on the number of infections and deaths? Did vaccination help? What

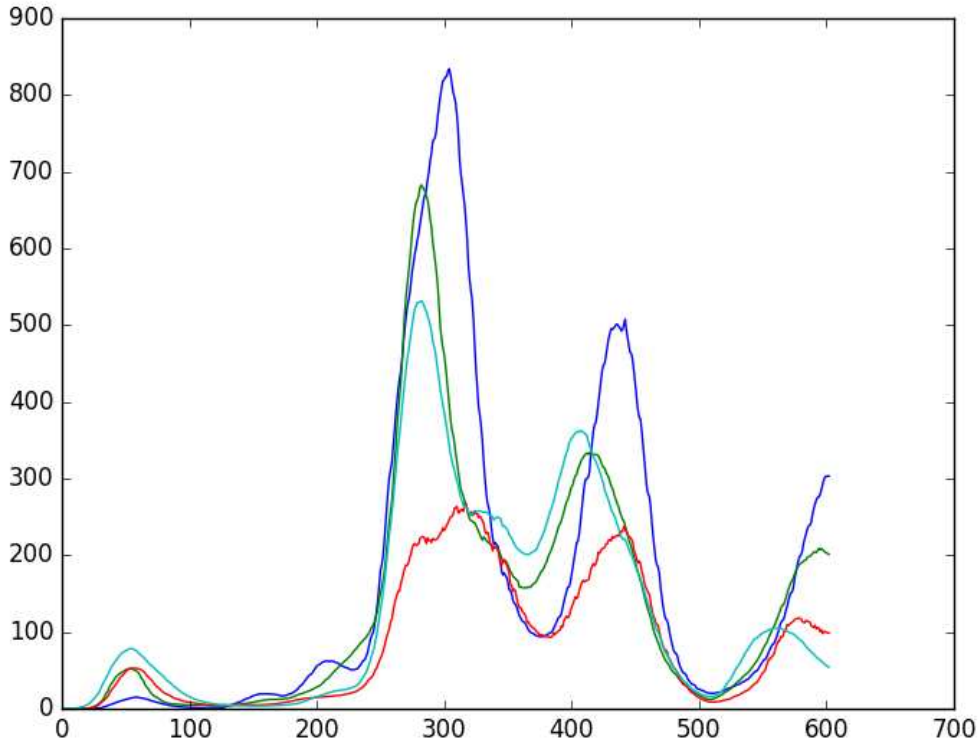

Figure 1: Number of infected per million. Blue: Cro, Green: Aut, Turquoise: Ita, Red: Ger. x axis: time period in days, y axis: number of infected per million.

is the relationship between infections and deaths? Does it vary between the second and fourth infection waves? Is it the same inside and outside the infection waves? Was overall mortality higher during the pandemic? Was Covid-19 the leading cause of death? We will try to answer some of these questions. Our primary focus will be on interpreting the statistical results and parameters derived, with particular emphasis on DIR-DDR and DDR-EMR relationship.

Chapter 1

Data and terminology

In this chapter, we will start by introducing the primary datasets used in our analysis, followed by the introduction of additional data. The three main datasets we utilized are DIR, DDR, and MDR. While these datasets were analyzed across all five countries, for clarity, we will illustrate each dataset using one country as an example.

1.1 Main data

DIR stands for Daily Infection Rate. It is a sequence whose elements tell us how many people were reported to be infected by SARS-CoV-2 on a daily basis.

DDR stands for Daily Death Rate. It is a sequence whose elements tell us how many people have died from Covid-19 infection on a daily basis.

Both datasets were sourced from the Johns Hopkins University and Medicine website. The data were reported directly to Johns Hopkins by local authorities in accordance with WHO guidelines (note: these guidelines underwent several revisions during the pandemic).

MDR stands for Monthly Death Rate. It consists of sequences indicating the total number of deaths in a given month within a specific country over multiple years. Our data covers the period from January 2010 to August 2022. This dataset was retrieved from the European Center for Disease Prevention and Control (ECDC).

1.2 Additional data

For our analysis, we also incorporated additional data, including the overall weekly vaccination rate and the weekly vaccination rate within specific age groups.

The overall weekly vaccination rate is the percentage of the population that has received at least one dose of the Covid-19 vaccine up to a specific date. For our analysis, the ending

date is 1 December 2021, while the starting date varies across countries based on their initial vaccine uptake.

The weekly vaccination rate within specific age groups is the number of people receiving the first dose of vaccine during that week, divided into various age groups. In our analysis, we will focus on two age groups: those over 60 years of age and those under 60 years of age. This data sets were retrieved from the ECDC.

1.3 Terminology

Here we will briefly explain the terms “infection wave” and “intensity.”

An infection wave is a period characterized by the rise and subsequent fall in the number of infections. It can be visualized as a hill-shaped segment in each graph in Figure 1.

Intensity refers to the peak number of infections during the infection wave. It can be visualized as the local maximum of the infection wave in Figure 1.

Chapter 2

Mathematical methods

Now that we have introduced the data we use, we are ready to present the mathematical methods applied in our analysis. Before definitions from descriptive statistics we will present some basic results from probability theory using [2].

2.1 Probability Theory

Probability Space

Definition 2.1.1. *A **random experiment**, or a **random trial**, is an experiment whose outcomes, i.e., results, are not uniquely determined by the conditions under which we conduct the experiment.*

Definition 2.1.2. *The **sample space** Ω is a non-empty set that represents the set of all outcomes of a random experiment. The elements ω of the set Ω are called **elementary events**.*

Definition 2.1.3. *A family $\mathcal{F}$ of subsets of Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) is a **$\sigma$-algebra of sets** on Ω if:*

1. $\emptyset \in \mathcal{F}$;
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$;
3. $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Definition 2.1.4. *Let $\mathcal{F}$ be a σ-algebra on the set Ω. The ordered pair $(\Omega, \mathcal{F})$ is called a **measurable space**.*

Definition 2.1.5. *Let $(\Omega, \mathcal{F})$ be a measurable space. A function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is a **probability** (on $\mathcal{F}$, on Ω) if it satisfies:*

1. $\mathbb{P}(A) \geq 0, \forall A \in \mathcal{F}$;

2. $\mathbb{P}(\Omega) = 1$;

3. $A_i \in \mathcal{F}, i \in \mathbb{N}$ *and* $A_i \cap A_j = \emptyset$ *for* $i \neq j \Longrightarrow \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Definition 2.1.6. *An ordered triple* $(\Omega, \mathcal{F}, \mathbb{P})$*, where* $\mathcal{F}$ *is a* σ*-algebra on* Ω *and* $\mathbb{P}$ *is a probability on* $\mathcal{F}$*, is called a* ***probability space****.*

Random Variable

Definition 2.1.7. *Let* S *be an arbitrary non-empty set and* $\mathcal{A}$ *be a family of subsets of* S *(*$\mathcal{A} \subset \mathcal{P}(S)$*). Denote by* $\sigma(\mathcal{A})$ *the smallest* σ*-algebra of subsets of* S *containing* $\mathcal{A}$*. We call it the* ***σ-algebra generated by*** $\mathcal{A}$ *.*

Definition 2.1.8. *Let* $\mathcal{B}$ *denote the* σ*-algebra generated by the family of all open sets on* $\mathbb{R}$*.* $\mathcal{B}$ *is called the* ***Borel σ-algebra*** *on* $\mathbb{R}$*, and the elements of the* σ*-algebra* $\mathcal{B}$ *are called* ***Borel sets*** *.*

Definition 2.1.9. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space. A function* $X : \Omega \to \mathbb{R}$ *is a* ***random variable*** *(on* Ω*) if* $X^{-1}(B) \in \mathcal{F}$ *for arbitrary* $B \in \mathcal{B}$*, i.e.,* $X^{-1}(B) \subset \mathcal{F}$*.*

Definition 2.1.10. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space, and let* $X : \Omega \to \mathbb{R}^n$*. We say that* X *is an* ***n-dimensional random vector*** *(or simply a* ***random vector****) (on* Ω*) if* $X^{-1}(B) \in \mathcal{F}$ *for every* $B \in \mathcal{B}^n$*, i.e.,* $X^{-1}(\mathcal{B}^n) \subset \mathcal{F}$*.*

Definition 2.1.11. *Let* X *be a random variable on* $(\Omega, \mathcal{F}, \mathbb{P})$*.* X *is a* ***simple random variable*** *if its range is a finite set.*

X is a simple random variable if and only if

$$X = \sum_{k=1}^{n} x_k \mathcal{K}_{A_k}$$

where $x_1, x_2, \ldots, x_n$ *are real numbers, and* $A_1, A_2, \ldots, A_n$ *are pairwise disjoint events with* $\bigcup_{k=1}^{n} A_k = \Omega$*.* $\mathcal{K}_{A_k}$ *denotes the characteristic function of the set* A_k*.*

Let $X_1, X_2 : \Omega \to \mathbb{R}$*. Then we define the functions* $X_1 \vee X_2$ *and* $X_1 \wedge X_2$ *on* Ω *by:*

$$(X_1 \vee X_2)(\omega) = \max\{X_1(\omega), X_2(\omega)\}, \quad \omega \in \Omega,$$

and

$$(X_1 \wedge X_2)(\omega) = \min\{X_1(\omega), X_2(\omega)\}, \quad \omega \in \Omega.$$

Using first of the two functions, we define the positive and negative parts of the real function X on Ω*:*

$$X^+ = X \vee 0, X^- = (-X) \vee 0.$$

X^+ *and* X^- *are non-negative real functions, and we have:*

$$X = X^+ - X^-,$$

$$|X| = X^+ + X^-.$$

Corollary 2.1.12. *X is a random variable if and only if* X^+ *and* X^- *are random variables.*

Theorem 2.1.13. *Let X be a non-negative random variable on* Ω*. Then there exists an increasing sequence* $(X_n, n \in \mathbb{N})$ *of non-negative simple random variables such that* $X = \lim_{n\to\infty} X_n$ *(on* Ω*).*

Mathematical Expectation and Variance

The definition of mathematical expectation is conducted in three steps. First, the mathematical expectation of a simple random variable is defined, then of a non-negative random variable, and finally of a general random variable.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{K}$ be the set of all simple random variables defined on Ω, and $\mathcal{K}_+$ the set of all non-negative functions in $\mathcal{K}$.

Let $X \in \mathcal{K}$, $X = \sum_{k=1}^{n} x_k \mathcal{K}_{A_k}$, where $A_1, A_2, \ldots, A_n \in \mathcal{F}$ are mutually disjoint.

Definition 2.1.14. ***Mathematical expectation*** *of X, or simply the* ***expectation*** *of X, is denoted by* $\mathbb{E}[X]$ *and defined as:*

$$\mathbb{E}[X] = \sum_{k=1}^{n} x_k \mathbb{P}(A_k).$$

Now let X be a **non-negative random variable** defined on Ω. According to Theorem 1.2.13, there exists an increasing sequence $(X_n)n \in \mathbb{N}$ of non-negative simple random variables such that $X = \lim n \to \infty X_n$. The sequence $(\mathbb{E}[X_n])n \in \mathbb{N}$ is an increasing sequence in $\mathbb{R}+$, so $\lim_{n\to\infty} E[X_n]$ exists and may be equal to $+\infty$.

Definition 2.1.15. ***Mathematical expectation*** *of X, or simply the* ***expectation*** *of X, is defined as*

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n].$$

Now let X be an **arbitrary random variable** on Ω. It holds that $X = X^+ - X^-$, where X^+ and X^- are non-negative random variables and $X^+, X^- \geq 0$.

Definition 2.1.16. *We say that the **mathematical expectation** of X, or simply the **expectation** of X, **exists** (or is defined) if at least one of the quantities $\mathbb{E}[X^+]$, $\mathbb{E}[X^-]$ is finite, i.e., if $\min\{\mathbb{E}[X^+], \mathbb{E}[X^-]\} < +\infty$. Then by definition, we set*

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

We list basic properties of mathematical expectation:

Theorem 2.1.17. *We have:*

1. If $\mathbb{E}[X]$ exists and $c \in \mathbb{R}$, then $\mathbb{E}[cX]$ exists and
$$\mathbb{E}[cX] = c\mathbb{E}[X].$$
2. If $X \leq Y$, then
$$\mathbb{E}[X] \leq \mathbb{E}[Y].$$
In the sense that
$$\text{if } -\infty < \mathbb{E}[X], \text{ then } -\infty < \mathbb{E}[Y] \text{ and } \mathbb{E}[X] \leq \mathbb{E}[Y],$$
or
$$\text{if } \mathbb{E}[Y] < \infty, \text{ then } \mathbb{E}[X] < \infty \text{ and } \mathbb{E}[X] \leq \mathbb{E}[Y].$$
3. If $\mathbb{E}[X]$ exists, then
$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|].$$
4. If $\mathbb{E}[X]$ exists, then $\mathbb{E}[X\mathcal{K}_A]$ exists for every $A \in \mathcal{F}$. If $\mathbb{E}[X]$ is finite, then $\mathbb{E}[X\mathcal{K}_A]$ is finite for every $A \in \mathcal{F}$.
5. Let X and Y be non-negative random variables. Then
$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Definition 2.1.18. *Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathbb{E}[X]$ be finite. Then we define the **variance** of X, denoted by $\mathrm{Var}(X)$ or σ_X^2, as follows:*

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Remark 2.1.19. *The positive square root of the variance is called the **standard deviation** of X and is denoted by σ_X .*

2.2 Descriptive statistics

For this section we use [1], [3] and [4]. We will explain the linear regression model, Pearson correlation coefficient, p-value, and moving average. To do so, we will first introduce some additional terms used in their definitions or descriptions.

During experiments and research, a numerical or non-numerical variable X is measured or observed. Variable consists of n observed values $x_1, x_2, ..., x_n$. In our analysis, we worked exclusively with numerical variables, so we will assume that all variables are numerical from this point forward. Numerical variables can be further divided into discrete (typically the result of counting) and continuous (such as physical measurements like weight and height). We worked only with discrete variables, so we will also assume that all variables are discrete from this point onward.

Pearson correlation coefficient

Before defining the Pearson correlation coefficient, we need to say what is mean.

Let X be variable with the following observed values:

$$x_1, x_2, ..., x_n \tag{1}$$

Definition 2.2.1. ***Mean*** *of* (1) *is the number* $\overline{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$

We will also introduce the terms sample variance and sample standard deviation.

Definition 2.2.2. ***Sample variance*** *of* (1) *is the number* $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Definition 2.2.3. ***Sample standard deviation*** *of* (1) *is the number* $s := +\sqrt{s^2}$

We can now introduce some additional notation that will be used in the Pearson correlation coefficient definition. Let us assume that we have two variables, X and Y with paired observed values:

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n) \tag{2}$$

$$S_{XX} := \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2 \tag{3}$$

$$S_{XY} := \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{xy} \tag{4}$$

$$S_{YY} := \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n \cdot \overline{y}^2 \tag{5}$$

Definition 2.2.4. *The quantity* $cov(X, Y) := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ *is called the* ***sample covariance*** *of X and Y*

So, $\frac{1}{n-1}S_{XX}$ and $\frac{1}{n-1}S_{YY}$ are sample variances of X and Y respectively, and $\frac{1}{n-1}S_{XY}$ is a sample covariance of X and Y.

Now we can define Pearson correlation coefficient, which is used to measure linear correlation between two variables X and Y.

Definition 2.2.5. ***Pearson correlation coefficient*** *is the number* $r_{XY} := \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}}$

The following holds:

$$-1 \leq r_{XY} \leq 1$$

We say that if:

- $r_{XY} < 0$, then X and Y are negatively correlated
- $r_{XY} > 0$, then X and Y are positively correlated
- $r_{XY} = 0$, there is no correlation between X and Y

Furthermore, we can divide strength of linear correlation in a following way:

- $0.8 \leq |r_{XY}|$, X and Y have strong correlation
- $0.4 \leq |r_{XY}| < 0.8$, X and Y have moderate correlation
- $0 < |r_{XY}| < 0.4$, X and Y have weak correlation

Now that we have defined strength of linear correlation, we are ready to describe linear regression model.

Linear regression model

The main idea is to adjust the line with the equation $y = \alpha + \beta x$ to the points (2). It can be done by using least square method. In other words, we would like to minimize the function $L(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n}({y_i}^2 + \alpha^2 + \beta^2 x_i^2 - 2\alpha y_i - 2\beta x_i y_i + 2\alpha\beta x_i)$. We do that by solving the following system of equations:

$$\frac{\partial L}{\partial \alpha}(\alpha, \beta) = 0 \tag{6}$$

$$\frac{\partial L}{\partial \beta}(\alpha, \beta) = 0 \tag{7}$$

Let us first solve (6).

$$\begin{aligned} 0 = \frac{\partial L}{\partial \alpha}(\alpha, \beta) &= \sum_{i=1}^{n} (2 \cdot \alpha - 2 \cdot y_i + 2 \cdot \beta x_i) \\ \sum_{i=1}^{n} \alpha &= \sum_{i=1}^{n} (y_i - \beta x_i) \\ n \cdot \alpha &= n \cdot \overline{y} - n \cdot \beta \cdot \overline{x} \\ \alpha &= \overline{y} - \beta \overline{x} \end{aligned}$$

We get an estimated value:

$$\hat{\alpha} = \overline{y} - \beta \overline{x} \tag{8}$$

Now we can solve (7).

$$0 = \frac{\partial L}{\partial \beta}(\alpha, \beta) = \sum_{i=1}^{n} (2 \cdot \beta \cdot {x_i}^2 - 2 \cdot x_i \cdot y_i + 2 \cdot \alpha x_i)$$

We use α obtained in (8).

$$\begin{aligned} 0 &= \sum_{i=1}^{n} (2 \cdot \beta \cdot {x_i}^2 - 2 \cdot x_i \cdot y_i + 2 \cdot \overline{y} \cdot x_i - 2 \cdot \beta \cdot \overline{x}) \\ \beta \cdot \sum_{i=1}^{n} ({x_i}^2 - x_i \cdot \overline{x}) &= \sum_{i=1}^{n} (x_i \cdot y_i - x_i \cdot \overline{y}) \\ \beta \cdot (\sum_{i=1}^{n} {x_i}^2 - \overline{x} \cdot \sum_{i=1}^{n} x_i) &= \sum_{i=1}^{n} x_i \cdot y_i - \overline{y} \cdot \sum_{i=1}^{n} x_i \\ \beta &= \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} {x_i}^2 - n \cdot \overline{x}^2} \end{aligned}$$

Using (3) and (5), we get the estimated value:

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} \tag{9}$$

So, the resulting line has the equation given in slope-intercept form:

$$\hat{y} = \hat{\alpha} + \hat{\beta} \cdot x \tag{10}$$

$\hat{\alpha}$ is called the intercept and $\hat{\beta}$ is called the slope of linear regression.

Residuals

Of course, in real life situations it is not reasonable to expect all the points (2) to lie on a single line. We can calculate for each point how far it is from the regression line (10) simply by subtracting their respective values on the y-axis. More formally:

Definition 2.2.6. *Linear regression* ***residuals*** *are values $y_i - \hat{y}_i$ where $\hat{y}_i$ is the predicted value from* (10)*.*

Role of p-value

In linear regression, the p-value helps determine whether the relationship between the dependent variable Y and independent variable X is statistically significant. It quantifies the probability of obtaining results that are at least as extreme as those observed, assuming that the null hypothesis (H_0) is true. H_0 in this case states that there is no correlation between X and Y (i.e., $\hat{\beta} = 0$).

The p-value is important because it helps us to understand whether the observed correlation or regression results are likely due to chance or if there is a significant relationship between the variables. A low p-value (typically ≤ 0.05) indicates that we can reject the null hypothesis, suggesting that the relationship between X and Y is statistically significant. Conversely, a high p-value suggests that we fail to reject the null hypothesis, indicating that any observed relationship is likely due to random variation.

2.3 Moving average

In our analysis, we will deal only with simple moving averages.

A Simple Moving Average (SMA) is a statistical measure that is used to smooth out short-term fluctuations and highlight longer-term trends or cycles in data. Let us suppose that we have variable X with data (1).

Definition 2.3.1. ***Simple moving average*** *is the number $SMA_k(t) = \frac{1}{k} \sum_{i=0}^{k-1} x_{t-i}, k \leq t \leq n$.*

Chapter 3

Results

We will begin this chapter by explaining how we calculated EMR from the MDR dataset. Next, we will present the results obtained from the DIR-DDR and DDR-EMR analyses conducted during the second and fourth waves of infection. Finally, we will show the relationship between DDR and EMR over a 12-month period.

3.1 EMR

To explain EMR we will first introduce the term ExMR which stands for Expected Mortality Rate. We have calculated ExMR using the MDR data set for a ten-year period lasting from January 2010 to December 2019. So, we have a total of 12 sequences, each containing the number of deceased individuals during a specific month. Each sequence consists of 10 elements, one for each year. Calculation was performed using the simple method which we explain here. For simplicity, the first two steps are explained for one sequence (i.e. deaths in the same month during a 10 year period).

1. Elimination of extreme values.

   Let $M_{j0} = (m_{j,1}, m_{j,2}, ..., m_{j,10})$ be a sequence representing the number of deaths in a $j-th$ month over a 10 year period. We ordered the elements in M_{j0} from the lowest to the highest value and get $M_j = (m_{j,(1)}, m_{j,(2)}, ..., m_{j,(10)}), m_{j,(1)} \leq m_{j,(2)} \leq ... \leq m_{j,(10)}$. We remove the first and the last element of M_j and get $M'_j = (m_{j,(2)}, m_{j,(3)}, ..., m_{j,(9)})$.

2. Calculation of monthly averages.

   We calculated the mean of the remaining 8 values in M'_j: $\overline{m_j} = \frac{1}{8} \sum_{i=2}^{9} m_{j,(i)}$

3. Formation of the average mortality rate.

We take the 12 calculated means $\overline{m_j}$, $j = 1, 2, ..., 12$ and put them in a new sequence $AMR = (\overline{m_1}, \overline{m_2}, ..., \overline{m_{12}})$ containing 12 average numbers of deaths, one for each month.

4. Application of 3-sliding averages.

   a) For months from February to November, we calculated the 3-sliding average as follows: $ExMR_j = \frac{1}{3}(\overline{m_{j-1}} + \overline{m_j} + \overline{m_{j+1}})$, $j = 2, 3, ..., 11$

   b) For January, we calculated the 3-sliding average using December, January, and February as follows: $ExMR_1 = \frac{1}{3}(\overline{m_{12}} + \overline{m_1} + \overline{m_2})$

   c) For December, we calculated the 3-sliding average using November, December, and January as follows: $ExMR_{12} = \frac{1}{3}(\overline{m_{11}} + \overline{m_{12}} + \overline{m_1})$

5. Creating ExMR.

   We create ExMR by putting the 12 values in a new sequence, $ExMR = (ExMR_1, ExMR_2, ..., ExMR_{12})$

It should be noted that the fourth step has also been done for the 5-sliding average, but it has not shown any significant differences in results.

We can now show how ExMR looks like for each country. We will look at Figure 3.1.

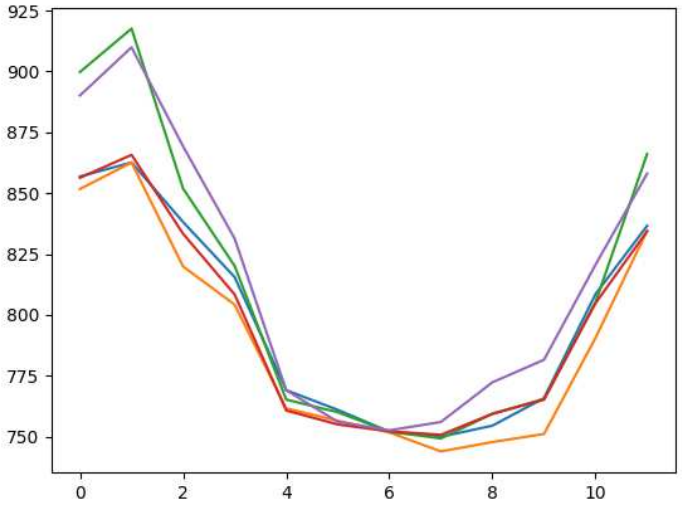

Figure 3.1: ExMR throughout a year. Green: Ita, Purple: Slo, Blue: Ger, Red: Aut, Orange: Cro. x axis: months, y axis: ExMR.

The graphs are not drawn to scale but the idea is to see that the overall trend is consistent: more deaths are expected during the winter months compared to the summer months.

Now we will explain the calculation of EMR. EMR stands for Excess Mortality Rate. It is calculated for the period from January 2020 to August 2022 by subtracting the initial number of deaths obtained from ECDC for each month from the corresponding number obtained in ExMR. For comparison with DDR, we also calculated the average daily EMR. This was achieved by dividing each EMR value by the exact number of days in the corresponding month.

3.2 DIR-DDR

DIR-DDR in the second infection wave

As can be seen in Figure 1, the second infection wave started in July 2020 and continued until February 2021. During this period, the total number of COVID-19 infections was 5,510,279 (from July 1, 2020, to February 28, 2021), and the total number of deaths was 145,354. It should be noted that no vaccine was available during this wave.

Let us now look at Figure 3.2.

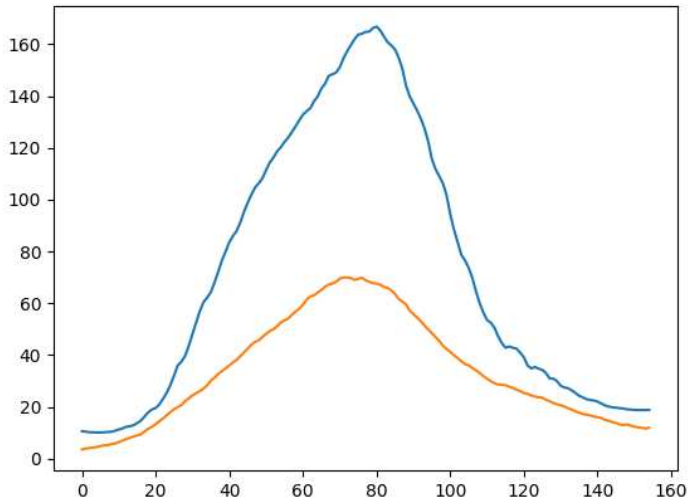

Figure 3.2: DIR-DDR Cro, second infection wave. Blue: DIR, Orange: DDR. x axis: time period in days, y axis: values.

Figure 3.2 shows the DIR and DDR trends during the second infection wave in Croatia. For better visualization, DIR data were divided by 20, and they were shifted by 17 days to achieve the best fit. The graph indicates a strong correlation between DIR and DDR.

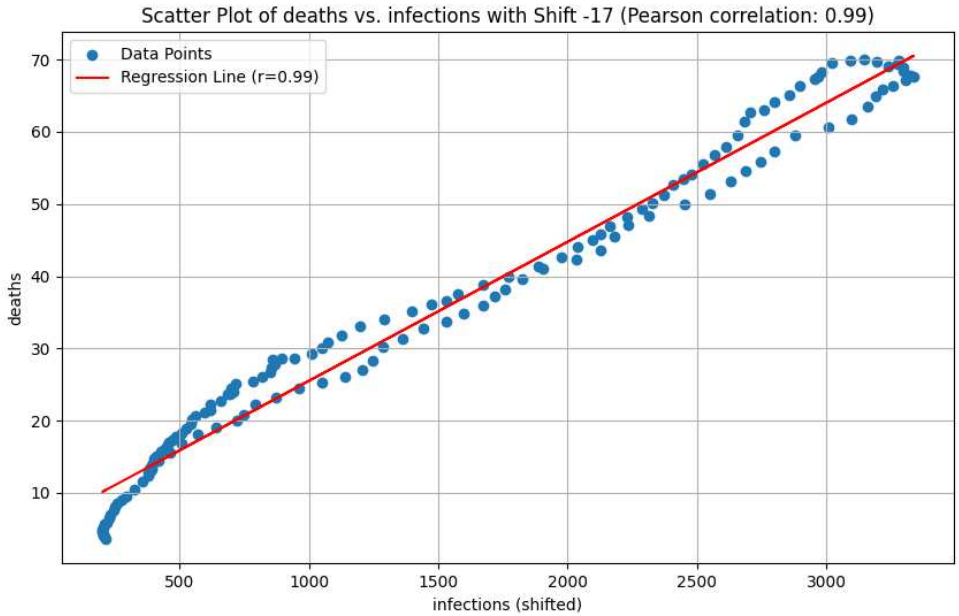

Figure 3.3: DIR-DDR scatter plot Cro, second infection wave. Blue: Data points, Red: Linear regression line. x axis: independent variable (X from linear regression model) values, y axis: dependent variable (Y from linear regression model) values.

As can be seen in Figure 3.3, it is evident that the regression line closely aligns with the data points, indicating a linear increase in the number of deaths as infections rise. Next, we will examine the residuals.

In Figure 3.4, the residuals display a discernible pattern rather than being randomly distributed. Nonetheless, it is worth noting that these deviations are very small and can be considered negligible.

We can now show Table 3.1 with calculated linear regression parameters, Pearson correlation coefficients and p-values for all countries.

According to Table 3.1, several observations can be made. All the p-values are very small. Pearson correlation coefficients are all greater than 0.8. The intercepts are negligible, as they are relatively small for each country. The slopes vary among countries, predominantly clustering around value 0.02. Instead of the term “slope”, we will use the term eF, which stands for “expected fatality” and use it during DIR-DDR analysis. Confi-

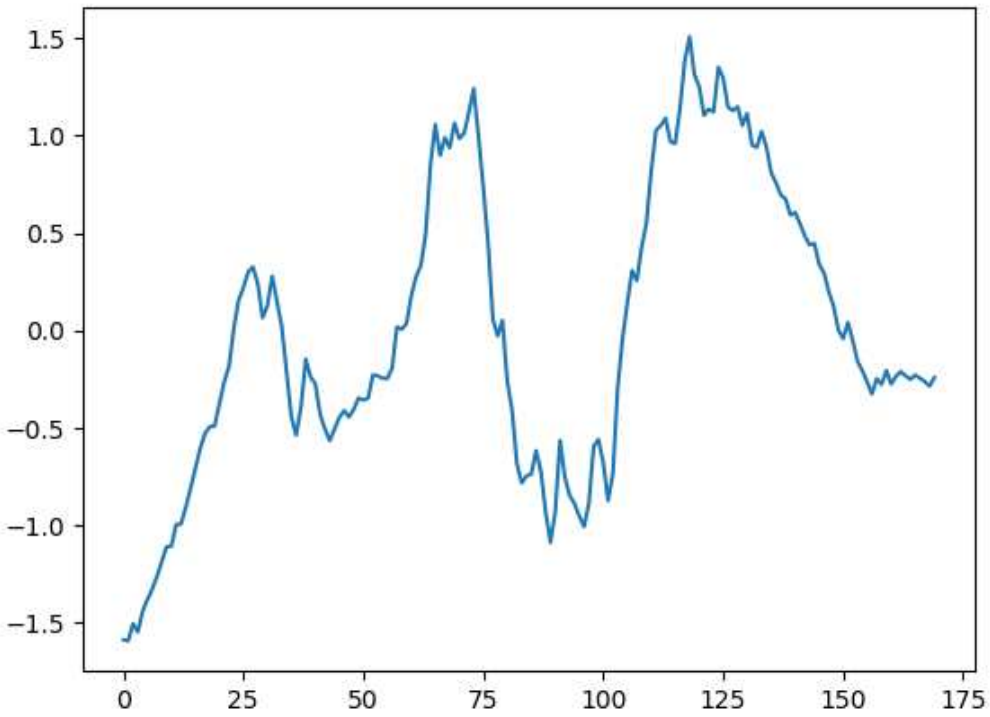

Figure 3.4: DIR-DDR residuals Cro, second infection wave. x axis: time period in days, y axis: residual values.

Country	intercept	slope	confidence interval	Pearson corr. coeff.	p-value
Austria	9.9721	0.0205	$\pm 2 \cdot 0.0004$	0.9726	$1.5650 \cdot 10^{-92}$
Croatia	6.1269	0.0193	$\pm 2 \cdot 0.0002$	0.9896	$1.0716 \cdot 10^{-130}$
Germany	-62.9888	0.0355	$\pm 2 \cdot 0.0007$	0.9661	$8.8508 \cdot 10^{-95}$
Italy	79.2767	0.0211	$\pm 2 \cdot 0.0004$	0.9725	$3.2589 \cdot 10^{-89}$
Slovenia	-4.7512	0.0244	$\pm 2 \cdot 0.0008$	0.8897	$9.2157 \cdot 10^{-73}$

Table 3.1: Intercept, slope with confidence interval, Pearson correlation coefficient and p-value.

dence intervals are very narrow showing possible overlap only for Austria and Italy and in extreme case for Austria and Croatia.

Now we show Table 3.2.

Country	eF	intensity
Austria	0.0205	690
Croatia	0.0193	830
Germany	0.0355	250
Italy	0.0211	520
Slovenia	0.0244	800

Table 3.2: eF and intensity.

Table 3.2 presents the relationship between the calculated eF and the observed intensities. We can see that there are big differences between intensities for different countries.

DIR-DDR in the fourth infection wave

The fourth infection wave began in August 2021 and lasted until January 2022. During that period the total number of Covid-19 infections was 11 661 227 (from 1 August 2021 until 31 January 2022) and the total number of deaths was 39 187. During the fourth wave of infection, a vaccine was available and accessible to everyone. Hence, here we look at vaccination rates as additional factor.

Let us now look at Figure 3.5.

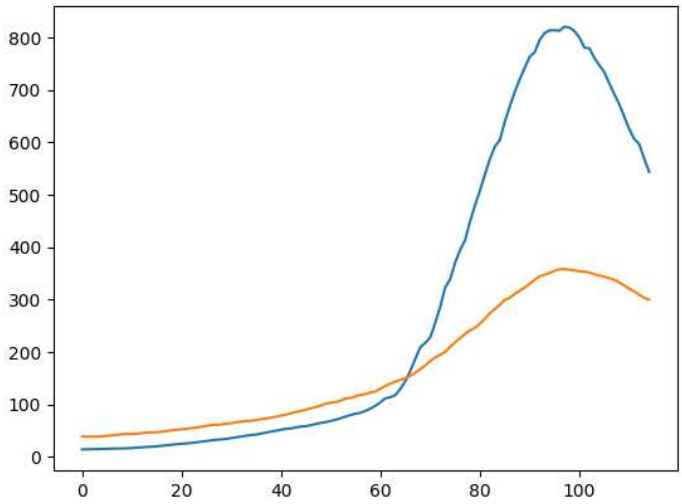

Figure 3.5: DIR-DDR Ita, fourth infection wave. Blue: DIR, Orange: DDR. x axis: time period in days, y axis: values.

Figure 3.5 shows the DIR and DDR trends during the fourth infection wave in Italy. For better visualization, DIR data were divided by 200, and they were shifted by 12 days to achieve the best fit. We can see that DIR and DDR are well correlated.

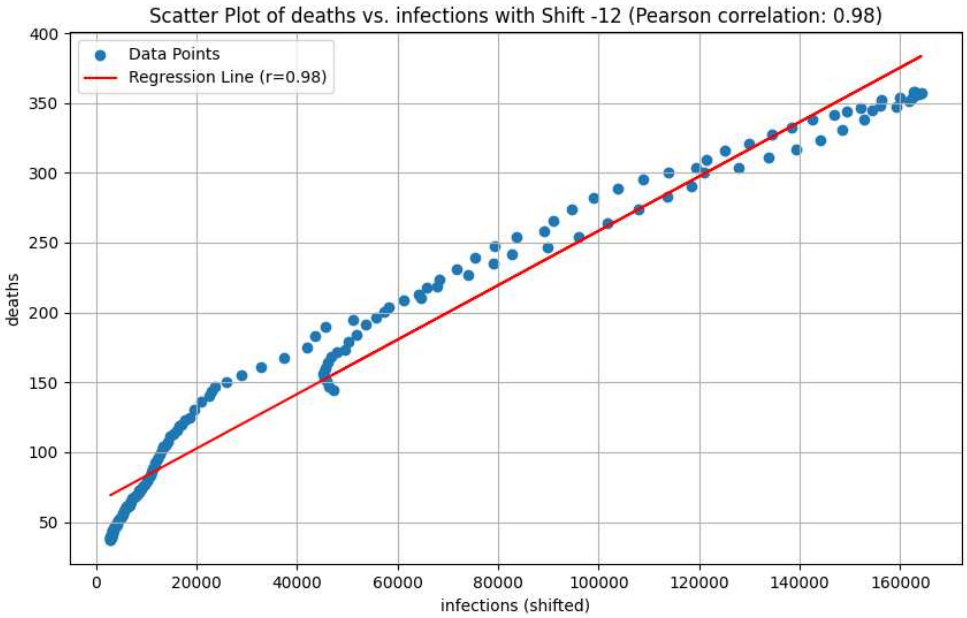

Figure 3.6: DIR-DDR scatter plot Ita, fourth infection wave. Blue: Data points, Red: Linear regression line. x axis: independent variable (X from linear regression model) values, y axis: dependent variable (Y from linear regression model) values.

As can be seen in Figure 3.6 it is evident that the regression line closely aligns with the data points, indicating a linear increase in the number of deaths as infections rise.

We can now show Table 3.3 with the calculated linear regression parameters, Pearson's correlation coefficients, and p-values for all countries.

According to Table 3.3, we can see similar results as for second infection wave. Again it holds that all the p-values are very small, Pearson correlation coefficients all greater than 0.8 and the intercepts negligible all of them being even smaller in absolute value.

The slopes vary among countries, predominantly clustering around value 0.006. Note: all slopes, except Croatia's, have decreased by an order of magnitude. Confidence intervals are again very narrow, this time showing possible overlap only for Austria and Germany.

Let us now look at Table 3.4

Table 3.4 presents the relationship between the calculated eF and the intensities as well as their relationship with the overall vaccination rate data and age group specific vaccina-

Country	intercept	eF	confidence interval	Pearson corr. coeff.	p-value
Austria	2.0923	0.0055	$\pm 2 \cdot 0.0001$	0.9596	$7.0763 \cdot 10^{-78}$
Croatia	2.1978	0.0125	$\pm 2 \cdot 0.0002$	0.9779	$2.9021 \cdot 10^{-109}$
Germany	26.6874	0.0055	$\pm 2 \cdot 0.0002$	0.9043	$7.2182 \cdot 10^{-66}$
Italy	60.9377	0.0019	$\pm 2 \cdot 3.0447 \cdot 10^{-5}$	0.9859	$1.2479 \cdot 10^{-89}$
Slovenia	0.0071	0.0061	$\pm 2 \cdot 5.8619 \cdot 10^{-5}$	0.9932	$5.7884 \cdot 10^{-140}$

Table 3.3: Intercept, eF with confidence interval, Pearson correlation coefficient and p-value.

Country	eF	intensity	overall vaccination %	vaccination % 60+	vaccination % u60
Croatia	0.0125	450	54.9	75.1	46.6
Germany	0.0055	110	69.8	89.6	61.6
Italy	0.0019	100	75.2	92.8	67.5
Slovenia	0.0061	320	54.5	76.9	46.1
Austria	0.0055	200	72.9	89.9	66.9

Table 3.4: eF, intensity and vaccination rates.

tion data calculated from last week in 2020 (when first person was vaccinated) until week 49 in 2021.

3.3 DDR-EMR

DDR-EMR in the second infection wave

Let us look at Figure 3.7.

Figure 3.7 shows the DDR and EMR trends during the second infection wave in Germany. We can see that DDR and EMR are well correlated.

We can now show Table 3.5 that contains the calculated slopes with confidence intervals, Pearson correlation coefficients, and p-values for all countries.

According to Table 3.5, we can see similarities with DIR-DDR case with all p-values being again very small, and Pearson correlation coefficients all greater than 0.8.

In this case slopes vary between approximately 0.70 for Croatia and 0.95 for Austria. Confidence intervals are relatively narrow showing possible overlapping only for Germany and Italy.

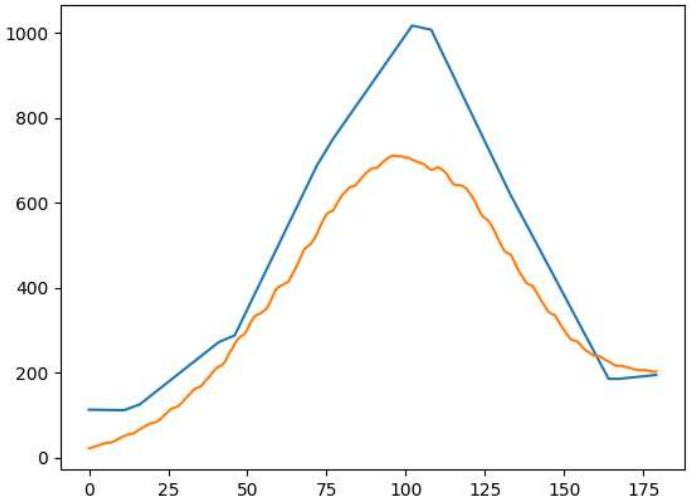

Figure 3.7: DDR-EMR Ger, second infection wave. Blue: EMR, Orange: DDR. x axis: time period in days, y axis: values.

Country	slope	confidence interval	Pearson correlation coeff.	p-value
Austria	0.9478	$\pm 2 \cdot 0.0245$	0.9481	$1.6887 \cdot 10^{-85}$
Croatia	0.7031	$\pm 2 \cdot 0.0123$	0.9753	$3.7178 \cdot 10^{-112}$
Germany	0.7909	$\pm 2 \cdot 0.0192$	0.9512	$7.8984 \cdot 10^{-93}$
Italy	0.7654	$\pm 2 \cdot 0.0197$	0.9486	$7.7601 \cdot 10^{-86}$
Slovenia	0.8572	$\pm 2 \cdot 0.0139$	0.9786	$2.6184 \cdot 10^{-117}$

Table 3.5: Slopes with confidence intervals, Pearson correlation coefficients and p-values.

DDR-EMR in the fourth infection wave

Let us look at Figure 3.8.

Figure 3.8 shows the DDR and EMR trends during the fourth infection wave in Germany. It appears that there could be a correlation between DDR and EMR, although a clear common trend is no longer apparent.

We can now show Table 3.6 that contains the calculated slopes with confidence intervals, Pearson correlation coefficients, and p-values for all countries.

According to Table 3.6 we can see that there is distinction between Germany and other countries regarding the parameters. The p-values, despite some differing greatly from each other, are again all very small (i.e., a lot smaller than 0.05). For other four countries Pearson correlation coefficients are again greater than 0.8 while for Germany it has value

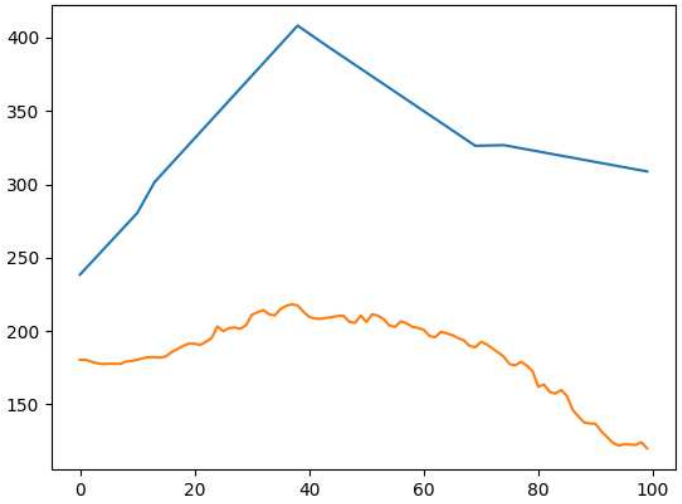

Figure 3.8: DDR-EMR Ger, fourth infection wave. Blue: EMR, Orange: DDR. x axis: time period in days, y axis: values.

Country	Slope	confidence interval	Pearson correlation coeff.	p-value
Austria	0.8016	$\pm 2 \cdot 0.0344$	0.9202	$1.0381 \cdot 10^{-41}$
Croatia	0.8025	$\pm 2 \cdot 0.0122$	0.9779	$1.4490 \cdot 10^{-136}$
Germany	0.4048	$\pm 2 \cdot 0.0806$	0.4196	$1.8386 \cdot 10^{-6}$
Italy	0.6937	$\pm 2 \cdot 0.0129$	0.9786	$1.0734 \cdot 10^{-89}$
Slovenia	0.7316	$\pm 2 \cdot 0.0080$	0.9942	$1.4100 \cdot 10^{-96}$

Table 3.6: Slopes with confidence intervals, Pearson correlation coefficients and p-values.

a little bit greater than 0.4. Confidence intervals show that there are multiple possible overlaps between all countries except Germany whose slope interval is very broad.

DDR-EMR during 12-month period

In this section, we will see the behavior of DDR and EMR during 12-month period starting with March 2020 and ending with March 2021 to identify similarities or differences in behavior within and outside the infection waves. First, let us explain how we calculated cumulative numbers, all of which are calculated on monthly basis.

- For DDR. Let i be the month of interest ($i = 1$ for March 2020, $i = 2$ for April 2020 etc.). Let $d_{i,j}$ be the number of reported deaths on $j - th$ day of the month i and let n_i

be the number of days of month i.

1. We first calculate Monthly Death Rate (MDR). We denote monthly number of deaths during month i with m_i. $m_i = \sum_{j=1}^{n_i} d_{i,j}$. In that way we have obtained MDR = $(m_1, m_2, ..., m_{12})$.
2. Now we calculate cumulative Monthly Death Rate (cMDR). We denote cumulative monthly number of deaths up to month i with cm_i. So, $cm_i = \sum_{k=1}^{i} m_k$. In that way we have obtained cMDR = $(cm_1, cm_2, ..., cm_{12})$.

- For EMR. Let emr_i be the excess mortality for the $i - th$ month. We calculate cumulative EMR (cEMR). We denote cumulative EMR up to month i with $cemr_i$. So, $cemr_i = \sum_{k=1}^{i} emr_k$. In that way we have obtained cEMR = $(cemr_1, cemr_2, ..., cemr_{12})$.

For better visualization, we will show the figure using logarithmic scale. In addition, we use EMR data calculated with a 3-month average and a 6-month average. Let us now look at Figure 3.9.

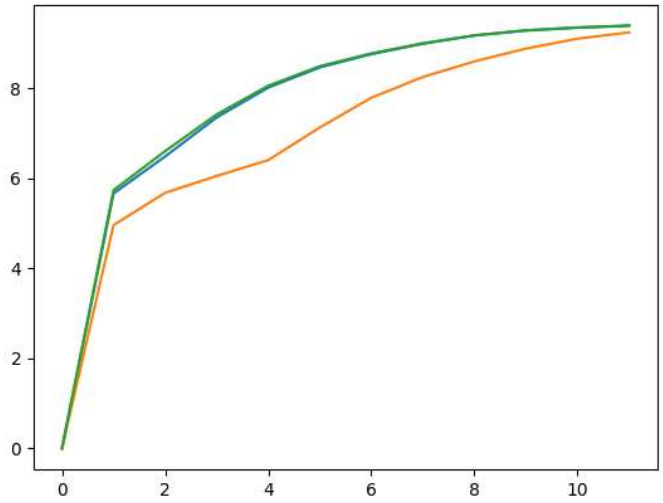

Figure 3.9: DDR-EMR cumulative logarithmic Aut, March 2020 - March 2021. Blue: EMR with 3-month moving average, Green: EMR with 6-month moving average, Orange: DDR. x axis: time period in months, y axis: logarithmic values.

Figure 3.9 shows that there is difference between DDR and EMR correlation throughout the 12-month period and that the two EMR lines overlap almost perfectly.

Chapter 4

Comments and discussion

4.1 DIR-DDR analysis

DIR-DDR analysis of the second infection wave

Our study utilizes a substantial dataset, encompassing nearly 6 million infected and nearly 150,000 deceased individuals, ensuring a robust sample size.

As shown in Figure 3.2, there is a clear correlation between DIR and DDR, with both metrics increasing and decreasing in tandem, suggesting a linear relationship. This correlation was further validated through linear regression, as depicted in Figure 3.3. Additionally, both metrics peak at approximately the same time. Residual analysis in Figure 3.4 revealed a non-random distribution, although the residuals are very small and therefore considered negligible.

In Table 3.1, Pearson correlation coefficients are presented, all of which are positive and greater than 0.8. This signifies a very strong positive linear correlation between DIR and DDR across all countries. **Now we should point out that these two variables are taken independently from one another. In other words, there was no tracking to confirm whether the individuals who were infected are the same individuals who died**.

Regarding eF, the combination of low p-values and strong correlation between DIR and DDR supports the assertion that the slope of the linear regression directly corresponds to the percentage of infected individuals expected to die, justifying the notation eF. Furthermore, the shifts observed in DIR suggest that the average period between infection and death is approximately two weeks. Therefore, during the second infection wave, the expected fatality rate of Covid-19 was approximately 2 %.

From Table 3.2, it is evident that there is a nearly perfect inverse correlation between eF and intensity. For all countries except Slovenia, lower eF values correspond to higher intensity values.

DIR-DDR analysis of the fourth infection wave

Similar to the second infection wave, the p-values in Table 3.3 are very low, indicating statistical significance. The dataset for infections was even larger, exceeding 11.5 million cases. Although the number of deceased was slightly smaller, just under 40,000, it remains substantial.

Figure 3.4, like in the case of Figure 3.2 indicates strong linear correlation between DIR and DDR. This correlation was once more confirmed through linear regression, as depicted in Figure 3.5.

From Table 3.3 we can see that Pearson correlation coefficients are again all higher than 0.8 which confirms linearity for all countries.

Regarding eF, the combination of low p-values and strong linear correlation between DIR and DDR again supports the assertion that the slope of the linear regression directly indicates the percentage of infected individuals expected to die. DIR shifts again suggest that the average period between infection and death is approximately two weeks. During the fourth infection wave, the expected fatality rate of Covid-19 was approximately 0.5 %.

From Table 3.4, it is evident that there is a perfect positive rank correlation between eF and intensities. In all countries, higher eF values correspond to higher intensity values. Additionally, there is an almost perfect inverse rank correlation between eF and overall vaccination rates, and consequently between intensities and overall vaccination rates. It is positive only for Slovenia and Croatia, while Germany and Austria have the same slope value despite differences in overall vaccination rates. To explain this phenomenon, we examine vaccination rates among specific age groups.

The vaccination rate was nearly 2 % higher in Slovenia for the most at-risk population, while in the less at-risk population, Croatia had a 0.5 % higher vaccination rate. Austria and Germany had practically the same vaccination rate in the most at-risk group, but there was more than a 5 % difference in the less at-risk group. This suggests that vaccinating the elderly population had a much greater impact on the expected fatality rate than overall vaccination rate. The case of Italy further reinforces this conclusion.

DIR-DDR analysis comparison

There are many similarities between the DIR-DDR comparisons for the second and fourth infection waves. Both cases demonstrate statistical significance and strong linear correlations, validating eF as a significant measure for observation. However, there are also notable differences between them.

During the second infection wave, the approximate eF is four times greater than during the fourth wave. Additionally, while the second infection wave shows an almost perfect inverse rank correlation between eF and intensity, the fourth infection wave exhibits an almost perfect positive rank correlation.

All these similarities and differences lead to important conclusions.

Conclusions

1. Strong linearity suggests that if the number of infected people doubles, the expected number of deaths also doubles within approximately two weeks.

2. Non-randomness of residuals is likely due to fewer tests during holidays or festive periods, leading to fewer reported infections, while increased social interactions contribute to higher virus spread and consequently more deaths.

3. Significant differences in eF values between the second and fourth waves, coupled with vaccination rates, indicate that vaccination had a substantial impact on mortality. Moreover, the shift in the correlation between eF and intensity from inverse during the second wave to nearly perfectly positive during the fourth wave suggests that vaccination also influenced susceptibility. This is further supported by the observation that intensities are much lower for each country during the fourth infection wave than during the second.

4. By examining vaccination rates among different age groups, we can reasonably conclude that expected fatality was primarily influenced by the vaccination of the elderly population.

4.2 DDR-EMR analysis

DDR-EMR analysis of the second infection wave

Again, as in DIR-DDR case, we obtain statistically significant results, as indicated by the very low p-values shown in Table 3.5.

As shown in Figure 3.7, there is a strong indication of a positive linear correlation between DDR and EMR, which is corroborated by the Pearson correlation coefficients in Table 3.5, all of which are greater than 0.8.

In this context, the slope indicates the proportion of Covid-19 deaths within EMR. We observe that this proportion ranges from approximately 70 % in Croatia to approximately 95 % in Austria.

DDR-EMR analysis of the fourth infection wave

Again as in the second wave case, we have very low p-values as shown in Table 3.6.

As depicted in Figure 3.8, there is an indication of a positive linear correlation between DDR and EMR, although it appears weaker compared to the second wave. The Pearson correlation coefficients in Table 3.6 are all greater than 0.8, except for Germany, indicating strong correlations between DDR and EMR for the other four countries, while for Germany, the correlation is moderate.

The slopes range from approximately 40 % (with a very wide confidence interval) for Germany to approximately 80 % for Croatia and Austria.

DDR-EMR analysis comparison

When comparing slopes between the second and fourth infection waves, we find that slope values were significantly lower for all countries except Croatia, where the slope is notably higher during the fourth infection wave.

DDR-EMR cumulative analysis

By looking at Figure 3.9 we note that the difference between the two EMR lines is barely visible and thus negligible. Therefore, there is no significant difference in calculating EMR using 3-month and 6-month sliding averages. Additionally, there is a noticeable difference in the share of Covid-19 deaths in EMR during periods of high intensity compared to periods of low intensity.

Conclusions

It can be concluded with high certainty that the share of Covid-19 deaths in EMR can be accurately calculated. During infection waves, Covid-19 deaths predominantly contribute to excess mortality. Conversely, during low-intensity periods, the proportion of Covid-19 deaths in excess mortality is lower. However, for more detailed study of DDR-EMR relationships further analysis is required.

Bibliography

[1] Miljenko Huzak, *Matematička statistika*, https://web.math.pmf.unizg.hr/nastava/ms/index.php?sadrzaj=predavanja.php, Accessed on Jun 27, 2024.

[2] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, 2002.

[3] Newcastle University, *Strength of correlation*, https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/strength-of-correlation.html, Accessed on Jun 29, 2024.

[4] Wikipedia, *p-value*, https://en.wikipedia.org/wiki/P-value, Accessed on Jun 29, 2024.

Summary

In this thesis, we analyze the impact of coronavirus disease Covid-19 on infections and deaths across five Central European countries, including Croatia. Our datasets include Daily Infection Rate (DIR), Daily Death Rate (DDR), Excess Mortality Rate (EMR), and vaccination rates. Statistical methods employed include simple linear regression and moving average analysis. We interpret our results using p-values and Pearson correlation coefficients, focusing on correlations between DIR-DDR and DDR-EMR datasets. Our findings are statistically significant and provide crucial insights on the disease's impact, especially regarding eF, vaccination rates and the influence of Covid-19 deaths on overall mortality.

CV

Roč Stilinović was born in Zagreb on 1 May, 1997. He attended the first four grades of elementary school in ”OŠ Tina Ujevića” in Zagreb and the second four grades in ”OŠ Cvjetno Naselje” in Zagreb. He graduated from ”XI. Gimnazija” in Zagreb in 2016. In 2022 he received his bachelor’s degree in Mathematics and enrolled in BioMedMath course at Faculty of Science in Zagreb as one of the first four students.