Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche

Relazione Finale

Unveiling the Digital Frontier: An In-Depth Analysis of the Digital Economy and Society Index

Relatore Dr. Manuela Scioni
Dipartimento di Scienze Statistiche, Università degli Studi di Padova

Correlatore Dr. Paola Annoni
Directorate-General for Communications Networks, Content and Technology,
European Commission

Laureanda: Anna Gotti
Matricola N 2023128

Anno Accademico 2022/2023

Preface

This master thesis is the result of a fruitful period spent in Belgium, first studying for a full academic year at KU Leuven as part of an exchange program, followed by a traineeship at the European Commission under the Directorate-General for Communications Networks, Content, and Technology. At the European Commission, I had the opportunity to delve into the composite indicators realm and their application in policy-related decision-making. This is how the idea of an in-depth analysis of the Digital Economy and Society Index based on robust methods came about.

I would like to express my gratitude to those who made all of this possible.
First, I would like to thank Paola Annoni and Manuela Scioni, for their trust and for guiding me trough all of this research experience.
Then, I would like to thank the Department of Statistical Sciences of the University of Padova, which financially supported this project through the Departments of Excellence 2018-2022 action for 'Scholarships for internships at international organizations abroad.'
Moreover, I would like to thank all the colleagues at DG CONNECT and the DESI team, who have shown interest in my work since the beginning and contributed invaluable insights.
And finally, thanks to my friends and my family for the unlimited support they gave me.

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence.
BAP	Budget Allocation Process.
CB	Cross Border.
CI	Composite Indicator.
CN	Connectivity.
CPCA	Classical Principal Components Analysis.
DESI	Digital Economy and Society Index.
DPS	Digital Public Services.
EW	Equal Weighting.
FA	Factor Analysis.
FTTP	Fiber To The Premises.
HC	Human Capital.
IA	Impact Analysis.
IDI	ICT Development Index.
IDT	Integration of Digital Technology.

KPIs	Key Principal Indicators.
MCD	Minimum Covariance Determinant.
MCDM	Multi-Criteria Decision Making.
NRI	Networked Readiness Index.
OD	Orthogonal Distance.
OFM	Orthogonal Factor Model.
PCA	Principal Components Analysis.
PCs	Principal Component.
ROBPCA	Robust Principal Components Analysis.
SA	Sensitivity Analysis.
SD	Score Distance.
SVD	Singular Value Decomposition.
TAI	Technology Achievement Index.
VHCN	Very High Capacity Networks.

Summary

The present work is the result of a traineeship experience, which took place at the European Commission under the Directorate-General for Communications Networks, Content, and Technology (DG CONNECT). The study has been conducted on the Digital Economy and Society Index (DESI) - 2022 edition, using the most recent version available of the Index (published in July 2022, European Commission 2022). The results obtained through the statistical assessment have been used for the formulation of the DESI 2023 edition.

Chapter 1 lays the groundwork by delving into the realm of composite indicators, clarifying their linkage to the concept of homogeneity. The chapter further underscores the significance of digitalization as a latent measure and explores the role of DESI in capturing this multifaceted phenomenon.

Then Chapter 2 describes the methodological choices related to the computation of DESI 2022 scores while Chapter 3 introduces the tools instrumental in evaluating the internal consistency of DESI 2022. Notably, Cronbach's alpha, Principal Component Analysis (PCA), the robustified variant ROBPCA, and the Pearson Correlation Coefficient are unveiled as the key statistical methodologies that play a pivotal role in observing the index coherence.

Chapter 4 contains the internal consistency analysis conducted across the various dimensions of DESI. This analysis is a precursor to justifying the aggregation rules employed in score computation. Moreover, a series of proposed adjustments to the dimensional frameworks is presented, with the

overarching aim of optimizing the internal consistency of the index. Chapter 5 aims at showcasing the subtle impact of the proposed adjustments on the dimensional scores of DESI. Additionally, this chapter undertakes the task of addressing the weight allocation challenge posed by the introduction of novel sub-dimensions within existing dimensions. To tackle this, sensitivity analysis over sub-dimensional weights is employed.

Chapter 1

Introduction

Composite Indicators (CIs) are synthetic indices of individual indicators often used to rank countries in various performance and policy areas (Freudenberg 2003b). A CI is a summary measure that combines multiple indicators or variables to provide a single measure of a concept or phenomenon. To achieve this, aggregation is involved, which refers to the process of combining the individual measures into the *aggregate* measure. In essence, a composite indicator can represent a *complex system* comprised of multiple *components*, which can be easier to comprehend as a whole rather than decomposing into its individual parts (Greco et al. 2019). The indicators or variables used to construct a composite indicator should be relevant to the concept being measured. Thus, the selection of indicators or variables and the chosen method for aggregation are of primary importance in a composite indicator and can have a significant impact on the final scores.

1.1 Building a CI

In recent times, the number of existing CIs has risen exponentially, due to their increased popularity in many research fields and to the high availability and complexity that data have assumed in the last decade. This big success

did not come without any criticism. Since the most relevant attribute of a CI is its ability to integrate large amounts of information into a single and easy-to-interpret score, the main concerns arising from the aggregation process are intuitive:

1. How to select and correctly aggregate together the individual indicators involved?
2. How is the final score obtained?
3. How well will the final score represent the underlying latent concept that we are interested in capturing?

All three listed questions are somehow interrelated.

In the following sections, different methodologies for building composite indicators will be discussed, related to both the aggregation approaches adopted and the CI structure. The ***Digital Economy and Society Index*** (DESI) ***- 2022 edition*** (European Commission 2022) will be taken as our test case. The reason DESI 2022 is chosen as the reference CI is twofold: it includes the lack of a previous methodological assessment of the index and DESI relevance in multiple policy areas. In fact, the main purpose of the following analysis is to conduct a statistical assessment of DESI, by addressing its structure through consistency proofs within its components, which justify the methodological aggregation choices. By using performance measures, i.e. impact analysis, the influence of framework updates suggested by the internal consistency analysis will be measured. Instead, a sensitivity analysis will be used in some cases to evaluate the uncertainty derived from testing different sets of weights on the newly defined framework components.

1.1.1 Conceptual framework and selection of individual indicators

Starting from the first concern, individual indicators should be organized under a *conceptual framework* which defines the concept being measured and

helps in identifying the indicators or variables used to capture it. For these reasons, it should be based on a clear understanding of the assessed phenomenon and the context in which it occurs. Thus, a CI should be developed in consultation with experts in the field, stakeholders, and potential users of it, as well as the collection of the individual indicators that are organized under it (Freudenberg 2003a). To better visualize the conceptual framework (see Figure 1.1), first, a hierarchical structure must be defined where the key dimensions of the main concept are identified and then individual indicators are selected and arranged below those. More levels of complexity might exist in a conceptual framework, where higher levels (dimensions) relate to fairly general areas and provide a basis for the specification of lower levels (sub-dimensions), from where it is then easier to identify the indicators. Given that the conceptual framework should specify the relationships between the dimensions and indicators, and should identify any interactions between them, the final score is obtained from this structure.

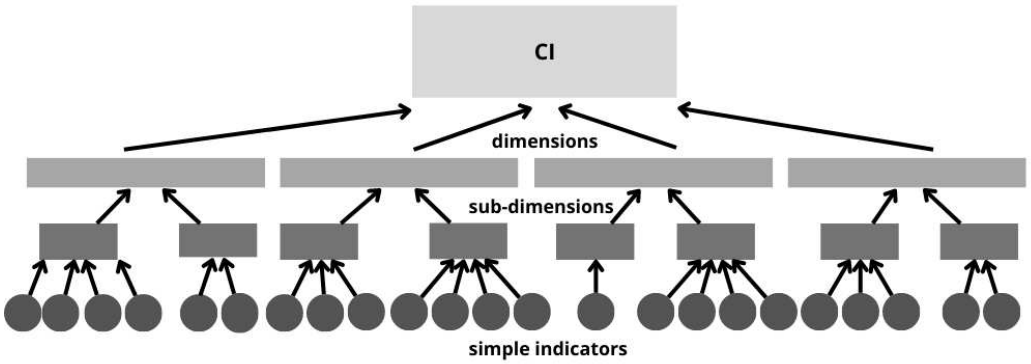


Figure 1.1: Example for the hierarchical structure of the conceptual framework of a CI

1.1.2 Weigthing and aggregation

Once the conceptual framework has been defined and individual indicators are brought to the same measurement scale through a normalization step, aggregation under the same construct is possible. The second concern

on CIs can then be answered. In fact, there are multiple ways of *aggregating individual indicators and weight* for them to obtain the final score.

The *selection of weights* can significantly affect the rankings of units in the CI, presenting a challenge in the construction of composite indicators known as the *index problem* (Rawls 1971). Moreover, the assigned weight has a twofold meaning (Nardo et al. 2008): an explicit one, which refers to the weight assigned to each criterion in the composite index; an implicit one, which instead relates to the trade-off between pairs of criteria during aggregation. Thus, selecting a weighting scheme always turns into a subjective choice, even when *equal weighting* (EW) is used. Accordingly to EW, no weights are assigned to the indicators thus it is also called an 'attributes-based weighting system' in Slottje (1991). This approach suffers from some problems, i.e. double counting, which occurs when combining indicators that are highly correlated. (Freudenberg 2003b; Nardo et al. 2008). Additionally, since EW can be applied hierarchically, in case the indicators are grouped into a higher order, such as a dimension, and weights are distributed equally dimension-wise, it does not imply that the individual indicators will have equal weights (Nardo et al. 2008). Thus, the actual weight will depend on the number of individual indicators in each dimension. The literature offers a variety of methods alternatively to the EW, some of which are *participatory* methods, i.e. budget allocation process (BAP), or *data-driven* ones, i.e. factor analysis (FA) (Greco et al. 2019). The BAP method is based on a group of experts selection: a specific group of decision-makers is provided with a certain number of points to distribute among the indicators or groups of indicators, such as dimensions. The final weighting scheme of the composite indicator is based on the average of the allocation choices made by the members of the expert panel (Custance and Hillier 1998). The BAP is for instance used by the European Commission for the creation of the *e-Business Readingness Index* (Pennoni et al. 2006) and the *Internal Market Index* (Tarantola et al. 2004). Instead, the FA method is reported in the context of CIs as an example of a data-driven technique (Decancq and Lugo 2013). Once the conceptual

model has been set, FA is used for weight elicitation, by grouping individual indicators according to their degree of correlation (Nardo et al. 2008).

Finally, when the weighting system has been decided, the *aggregation method* allows us to obtain the final score. According to Nardo et al. (2008), aggregation methods can be divided into three categories: linear, geometric, and multicriteria. Each aggregation leads to different compensation and allows the weights to be a measure of importance, in the case of non-compensatory aggregation methods. *Compensability* refers to the existence of trade-offs and the use of weights with the intensity of preference originates compensatory aggregation methods and gives the meaning of trade-offs to weights (Munda 2012). For instance, strong compensability allows for outstanding performance in some aspects to balance the weaknesses in others and vice-versa. Thus, the compensatory nature between indicators varies depending on the aggregation used, according to Munda (2008). In this way, the previous categorization translates in compensation terms as *full* (corresponding to linear aggregation), *partial* (in the case of geometric methods and partially-compensatory multicriteria methods), and *zero* (in the case of non-compensatory multicriteria decision making (MCDM) techniques). A fully compensatory aggregation function has to be justified by a satisfactory internal consistency level within and across sub-dimensions and across dimensions (Decancq and Lugo 2010; Seth 2009; El Gibari et al. 2018). When using standard linear composite aggregation rules, compensability among the different individual indicators is always assumed which implies complete substitutability among the different components of the CI. The MCDM literature instead advocate non-compensatory and non-linear aggregation rules as an alternative to this approach. In the context of multi-criteria, different methods are available providing with fully or partial non-compensatory techniques, namely the counting method proposed by Alkire and Foster (2011) or purely multi-criteria approaches based on partial order (Annoni 2007; Annoni and Brüggemann 2009; Brüggemann and Carlsen 2012).

1.1.3 Measures of performance

While we can select the individual indicators that best correlate to the underlying latent variable capturing the phenomenon of interest, it is hard to say how well a CI approximates that latent variable. This point refers to the third concern about CIs. In fact, by definition, a latent factor is a latent variable that cannot be observed, and thus can't be directly measured. Instead, what is possible to test for, and also helps in answering the question about the performance of a CI, is *homogeneity* among the observable individual indicators. In this context, many statistical methods for testing homogeneity and thus multivariate correlation have been developed, namely *Cronbach's alpha* (Cronbach 1951), *Principal Component Analysis* (PCA -Jolliffe (2002)), and *Factor Analysis* (FA - Fabrigar et al. (1999)). These methods are used to test for *unidimensionality* within indicators of the same sub-dimensions, within sub-dimensions of the same dimension, and ultimately, between dimensions. Finally, a robustness analysis of the methodological choices can be used to assess the impact that changes have on the final scores. By serving as a *quality assurance* tool, this will demonstrate the index sensitivity to any changes made during its construction process, which will significantly decrease the chances of conveying any misleading information (Saisana et al. 2005). The first instrument used in the context of robustness analysis is *impact analysis* (IA), which is used to test for the robustness of the final scores, by changing some of the inputs. Alternatively, *sensitivity analysis* (SA) quantifies the extent to which uncertainties contribute to the variation in the overall output, as stated by Saisana et al. (2005).

1.2 On digitalization and composite indicators

The concept of digitalization is becoming increasingly important within socio-economical changes discourse and takes a prominent position in promoting a human-centric, sustainable vision for a digital society. At the same time, digitalization is multifaceted and hard to capture, also due to its contin-

uous evolution in the digital era. Measuring the digital divide of the Member States of the European Union and defining digitalization components are primary steps within Europe's Digital Decade (European Commission 2021), a comprehensive framework that will guide all actions related to digital in Europe towards 2030. The aim of the Digital Decade is to ensure all aspects of technology and innovation work for people. Within the Digital Decade framework, targets have been set, which are measurable goals for each of the four defined areas: *Connectivity*, *Digital Skills*, *Digital Business*, and *Digital Public Services*. To achieve its digital goals and aims, the European Commission intends to expedite and simplify the initiation of multi-country initiatives, involving large-scale projects that individual Member States could not independently undertake.

These initiatives have the potential to:

1. Pool investments from sources including the EU budget, the Recovery and Resilience Facility, Member States, and private sector entities.

2. Address deficiencies in critical capacities within the European Union that have been identified.

3. Promote a Digital Single Market that is interconnected, compatible, and secure.

The DESI composite indicator arises from the need to track digital progress, and thus the effectiveness of the Digital Decade policy investments and synergies, by ranking the 27 Member States according to their level of digitalization. Since its introduction in 2014, its framework has been updated yearly. In the following dissertation, we refer to the most recent update, the 2022 edition. Nevertheless, DESI combines a relevant number of indicators, 33, classified into 4 dimensions, corresponding to the four cardinal points of the Digital Decade. Moreover, each dimension contains a different number of sub-dimensions, a total of 10, representing different sub-aspects of the same dimensional domain (see Chapter 2 for more on DESI structure).

In addition to DESI, presented by the European Commission, several institutions, and organizations have been developing proposals for digitalization proxy concepts measurement. The first composite indicator measuring the Digital Divide was introduced in 2001 by the United Nations, the *Technology Achievement Index* (TAI) and described in detail in the Human Development Report (United Nations 2001), followed the same year by the *Networked Readiness Index* (NRI - World Economic Forum (2019)). TAI 2001 contains 8 individual indicators, divided into 4 dimensions, i.e.*Creation of technology*, *Diffusion of recent innovations*, *Diffusion of old technologies*, and *Human skills development*. No lower hierarchical structure exists for this CI (absence of division into sub-dimensions). NRI 2001 instead, contains 53 individual indicators which are respectively divided into 4 dimensions (and 12 dimensions total), namely *Environment (Political and regulatory, Business and Information)*, *Readiness (Infrastructure, Affordability, Skills)*, *Usage (Individual, Business, Government)*, *Impact (Economic, Social)*.

Despite some differences, the three CIs described share three key aspects related to ICT:

1. the development of ICT infrastructure, which includes connection networks infrastructure, internet connection quality, and users' access to the service;

2. ICT skills, which refer to users' abilities and knowledge in using computer devices and internet services;

3. ICT usage, which highlights the various ways in which internet services are used by individual users, private companies, and public organizations.

DESI, released in 2014, is relatively innovative compared to those primordial composite indicators. In fact, it incorporates individual indicators that capture a *wider spectrum* of the digital. In fact, differently from the other two, DESI integrates with information of Cloud and Artificial Intelligence (AI) under the digital business domain dimension and also includes a dimension

that captures the demand and supply of e-government as well as open data policies. This structure is in line with its more recent story and falls under the Digital Decade policy. However, having individual indicators that capture a larger range of the digital does not necessarily translate into an accurate CI. In fact, if individual indicators do not reach a satisfactory level of internal consistency within the same construct, the final score may give misleading information when non-compensatory aggregation functions are used, as in the case of DESI 2022.

In terms of standardization and aggregation, DESI uses min-max standardization and a hierarchical approach for aggregation: it applies a weighted arithmetic mean to aggregate individual indicators within each sub-dimension, followed by a weighted arithmetic mean of sub-dimensional scores for each dimensional score, that is aggregated with a simple average into the final overall score. While NRI has a similar hierarchical structure to DESI, TAI quite differs from it. In fact, in terms of framework, TAI holds two levels of complexity only, the dimensional and the overall one. Moreover, TAI is much less informative, including only 8 simple indicators compared to the 33 of DESI and 53 of NRI. Further, TAI aggregates indicators using a simple arithmetic mean at both sub-dimensional and dimensional levels. In fact, Saisana et al. (2005) shows that, in the case of the TAI, changing the weights of certain indicators seems to affect several of the units evaluated, especially those that are ranked in middle positions. DESI individual indicators weights are instead based on the 2030 Digital Compass targets (European Commission 2021): target indicators are double-weighted within their sub-dimensions. Also, sub-dimensions scores are aggregated according to weights assigned by experts while the four dimensions of the Digital Compass are of equal importance, thus the arithmetic mean is used for their aggregation. Finally, the three CIs use a common aggregation approach, namely the arithmetic mean with equal weights, at an overall level, which implies a strong compensability of dimensions.

We can conclude that DESI power, compared with previous proposed CIs,

resides in extending the digital domain by considering novel technologies and more refined elements, and not in its methodological approach to weighting and aggregation.
The internal consistency analyses performed on the DESI 2022 framework will thus reveal whether a wider structure leads to a satisfactory measure of the current digital domain.

1.3 Measures of homogeneity

As previously mentioned, the internal consistency of indicators within and across sub-dimensions/dimensions is a necessary property of CIs when using compensatory approaches for aggregation. In fact, if a satisfactory level of homogeneity is not reached, composite indicators scores may return misleading results. Internal consistency can be seen as a proxy to test for homogeneity of CIs, where a homogeneity analysis in its broader sense is defined in Gifi (1990) as: 'A class of criteria for analyzing multivariate data in general, sharing the characteristic aim of optimizing the homogeneity of variables under various forms of manipulation and simplification'. Homogeneity can thus help in defining which individual indicators are the most appropriate to capture latent variables and contribute in the same direction for the same concept. In the literature, we can find extensive use of homogeneity techniques. In the CIs context, for instance, PCA can serve the purpose of examining the underlying nature of the data and exploring whether their different dimensions are statistically well-balanced. In the absence of a conceptual model, within a defined framework PCA can be used within dimension/sub-dimension to test for internal consistency of individual indicators. An example is the *ICT Development Index* (IDI - International Telecommunication Union (2017)), a CI published from 2009 until 2017. IDI is made of three main dimensions and within each, PCA is applied to find the most consistent indicators and eliminate the ones that are not. Additionally, IDI uses factor loadings of the first component of PCA as a weight elicitation technique, to serve as weights

for the indicators (Greyling and Tregenna 2017). However, when the first principal component is not able to explain an adequate portion of the variance of the indicators, more components are needed. Nicoletti et al. (2000) develop indicators of product market regulation, illustrating how these can be accomplished using FA. In order to reduce the number of indicators with high loadings on each component, the authors used the principal component method for factor extraction and rotated the components according to the varimax method. By examining the factor loadings of all retained factors, as described in Nicoletti et al. (2000), they were able to preserve the largest proportion of variation in the original dataset.

Finally, what we would like to achieve with this project is assessing the homogeneity of dimensional/sub-dimensional structures within DESI 2022 framework. This assessment is essential to validate two key methodological choices related to the composite indicator, being the *major objectives*:

- The use of compensatory rules for aggregation;
- The inclusion of detailed individual indicators to capture a wider spectrum of the digital domain while aligning with the Digital Decade policy objectives.

Additionally, following the experts' decision of not correcting for univariate outlying behaviors of certain countries for DESI 2022, we will compare the Classical PCA method with its robustified version (ROBPCA - Rousseeuw 1984). This comparison will be carried out at both the overall CI level, considering all individual indicators in DESI 2022, and at the dimensional level. The comparison aims to accomplish two *minor objectives*:

- Test the robustness of the PCA procedure against outliers;
- Detect possible outlying countries at the overall or dimensional level by employing measures of multivariate outlyingness.

While the first listed insight deriving from the robust PCA analysis mostly addresses statistical considerations, the second can provide valuable insights

for policy purposes. Indeed, identifying measures of multivariate outlyingness can justify further in-depth analyses of countries that exhibit distinct behaviours.

Chapter 2

DESI 2022: data collection, descriptive analysis and aggregation

This chapter provides an overview of the conceptual framework of the DESI 2022 and the individual indicators included in its computation (Paragraph 2.1). Additionally, we delve into the data sources and the methods employed for the imputation of missing data (Paragraph 2.2). A descriptive analysis of individual indicators is presented in Paragraph 2.3, offering valuable insights into their characteristics. Furthermore, we detail the data normalization procedures (Paragraph 2.4) and the aggregation methods utilized in the DESI computation (Paragraph 2.5). These essential components collectively form the foundation of our study, enabling a comprehensive analysis of the DESI 2022.

2.1 Conceptual framework

As previously mentioned, DESI 2022 composite indicator should help a country to assess its position relative to others, possibly in order to benchmark policies. Thus, its need for redefinition in the digital age calls on policymakers to take a new look at the current digital achievements. While

acknowledging that many elements make up a country's digital achievement, the index suggests that an overall assessment is more clear based on a single composite measure. Like other existing CIs, DESI 2022 is suggested for summary purposes, to be followed by individual analysis of the underlying indicators.

The DESI 2022 focuses on 33 individual indicators measured in the 27 EU member countries. The indicators are divided into the four main **dimensions** of digital development, which are in turn split into respective **sub-dimensions**. The DESI dimensions are:

- ***Human Capital* (HC)**: the human capital dimension assesses both the internet user skills of citizens and the advanced skills of specialists. For this reason, two sub-dimensions are used to capture the two domains, respectively called Internet user skills, labeled as **1a**, and Advanced skills and development (**1b**).

- ***Connectivity* (CN)**: the connectivity dimension considers both fixed and mobile broadband through indicators measuring the supply and the demand side as well as retail prices. The connectivity spectrum is then divided into four sub-dimensions, namely Fixed broadband take-up (**2a**), Fixed broadband coverage (**2b**), Mobile broadband (**2c**), and Broadband prices (**2d**).

- ***Integration of Digital Technology* (IDT)**: this dimension captures the extent to which countries enable access to digital services for all citizens, to maintain their digital prosperity. This is performed by capturing 3 main concepts, which correspond to three defined sub-dimensions, namely the use of different digital technologies at an enterprise level, corresponding to Digital Intensity (**3a**) sub-dimension, take-up of selected technologies, named Digital technologies for businesses by enterprises (**3b**) and e-commerce, contained in e-Commerce (**3c**) sub-dimension.

- ***Digital Public Services*** **(DPS)**: describes the demand and supply of e-government as well as open data policies, without any internal division. The unique sub-dimension, corresponding entirely to the dimension, is called e-Government (**4a**).

The DESI three-level structure is exhaustively depicted in Table 2.1, where the 33 included individual indicators are divided into the respective sub-dimensions and dimensions.

Dimension	**Sub-dimension**	**Indicator**
1 Human capital	1a Internet user skills	**1a1*** At least basic digital skills
		1a2 Above basic digital skills
		1a3 At least basic digital content creation skills
	1b Advanced skills and development	**1b1*** ICT specialists
		1b2* Female ICT specialists
		1b3 Enterprises providing ICT training
		1b4 ICT graduates
2 Connectivity	2a Fixed broadband take-up	**2a1** Overall fixed broadband take-up
		2a2 At least 100 Mbps fixed broadband take-up
		2a3 At least 1 Gbps take-up
	2b Fixed broadband coverage	**2b1** Fast broadband (NGA) coverage
		2b2* Fixed Very High Capacity Network (VHCN) coverage
		2b3 Fibre to the Premises (FTTP) coverage
	2c Mobile broadband	**2c1** 5G spectrum
		2c2* 5G coverage

		2c3 Mobile broadband take-up
	2d Broadband prices	**2d1** Broadband price index
3 Integration of digital technology	3a Digital intensity	**3a1*** SMEs with at least a basic level of digital intensity
	3b Digital technologies for businesses	**3b1** Electronic information sharing
		3b2 Social media
		3b3* Big data
		3b4* Cloud
		3b5* AI
		3b6 ICT for environmental sustainability
		3b7 e-Invoices
	3c e-Commerce	**3c1** SMEs selling online
		3c2 e-Commerce turnover
		3c3 Selling online cross-border
4 Digital public services	4a e-Government	**4a1** e-Government users
		4a2 Pre-filled forms
		4a3* Digital public services for citizens
		4a4* Digital public services for businesses
		4a5 Open data

Table 2.1: DESI structure

In total, there are ten sub-dimensions, differently distributed among the four dimensions of DESI. Additionally, 11 of the DESI 2022 indicators measure the Digital Decade targets and each dimension contains a different number of EU-level targets, called ***Key Principal Indicators*** (KPIs - noted with $*$ in Table 2.1).

For each of the 33 individual indicators the respective *description*, *unit of measure*, *source of collection*, and *authority of reference*, can be found in Figures B.1, B.2, B.3, B.4 of Appendix B.

2.2 Data sources, quality and availability

The data used to construct the DESI are obtained from relevant authorities of the Member States by the European Commission (Directorate-General for Communications Networks, Content and Technology as well as Eurostat) and from ad hoc studies launched by the Commission. In Appendix B, Figures B.1, B.2, B.3 and B.4 contains more information about each indicator *data source*, while Table B.1 lists the *authorities* in charge of those.
The data used for calculating DESI 2022 is collected one year prior to its publication. Therefore, when referring to 2021 data, we are specifically referring to the data used for the computation of DESI 2022 edition, and the same principle applies to the previous publications.

DESI 2022 was computed for 27 countries, for which data were partially available and of acceptable quality. However, it is important to note that data availability is not always guaranteed, and some countries may not have provided certain data points due to reasons such as non-delivery by the authority responsible for data collection. In such instances, these data points are considered as *missing*. Additionally, there were cases where the data quality was disputed after collection. Specifically, for certain countries, the authority responsible for data collection provided data that did not align with the country-specific time series of individual indicators. This resulted in anomalies, such as a drastic drop in a particular time point, leading to inconsistencies within the expected time evolution of the indicator, and affecting the country's score for that year. To address this issue, post-cleaning procedures were carried out at the discretion of DESI experts after data collection. The primary goal of these procedures was to substitute problematic data points with imputed data. During this process, the data in question

were *censored* to ensure the accuracy and integrity of the resulting DESI scores and rankings. This allowed for more reliable and meaningful comparisons between countries and facilitated the generation of a comprehensive and consistent assessment of their digital performance.

For DESI 2022, both in the case of missing *or* censored data points, the two main *imputation methods* are based on 2020 data, which were already cleaned and imputed for the computation of DESI 2021. The two methods are following described:

- *Imputation using the growth rate*: the method involves calculating the EU average growth rate of the x_i indicator between the years 2020 and 2021 ($rate_{2021-20}$). This growth rate is then multiplied by the data point for the same indicator in the year 2020 ($x_{i,c|2020}$) corresponding to the country (indexed with c) and indicator of interest (x_i). The formula to obtain the imputed data point $\hat{x}_{i,c|2021}$ is

$$\hat{x}_{i,c|2021} = rate_{2021-20} \times x_{i,c|2020}. \tag{2.1}$$

  The resulting value is used as an imputed data point for the missing or disputed data, providing an estimate for the indicator value in the specific year where data were unavailable or of poor quality. This approach allows for a reasonable estimation of the indicator's value in a consistent and standardized manner across the countries in the study.

- *Imputation using the rank*: this method addresses missing or disputed data points by employing rank-based information. This method aims to estimate the values of specific indicators for a particular country and year where data is not available or of questionable quality. To implement this method, the researchers first determine the rank of the country of interest (indexed with c) for the individual indicator x_i in the year 2020. This rank is denoted as $rank_{i,c|2020}$ and indicates the country's relative position in the data distribution for the specific indicator in 2020. Next, the researchers identify two countries for which

data is available in the year 2021: one corresponding to the rank immediately below the country of interest (indexed as $rank_{i,c|2020} - 1$), denoted as country c', and another corresponding to the same rank as the country of interest in 2020, denoted as country c''. Subsequently, the researchers calculate the average value of the indicator x_i for country c' and country c'' in the year 2021. This average value is then used to impute the missing data point for the country of interest (c) and indicator (x_i) in the year 2021.

The formula to obtain the imputed value $\hat{x}_{i,c|2021}$ is in this case

$$\hat{x}_{i,c|2021} = \frac{1}{2}(x_{i,c'|2021} + x_{i,c''|2021}). \tag{2.2}$$

As per the DESI experts, when employing each imputation method, special consideration is given to whether the imputed data point aligns with the overall trend and evolution of the indicator's time series for the specific country in question. In other words, after performing data imputation using the designated methods, the DESI experts thoroughly assess whether the estimated values logically fit into the historical progression of the indicator for the chosen country. This assessment is crucial to ensure that the imputed data points accurately reflect the trend and behaviour of the indicator over time, thus maintaining the coherence and reliability of the final score. Additionally, the assessment helps in picking an imputation method. In fact, for each missing or censored data point, the method is chosen in an arbitrary manner by selecting the one that offers the most accurate fit of the data point within the time series of the individual indicator.

2.3 Descriptive analysis

The statistics for the data with imputed values are given in Appendix A. For each raw indicator, x_i, several statistics are measured that are:

- *% of missing values*: counts the percentage of missing data or censored data;

- *average*: compute the individual indicator average (μ_{x_i});

- *standard deviation*: compute the individual indicator standard deviation (σ_{x_i});

- *coefficient of variation*: defined as the ratio of the standard deviation to the average ($\frac{\sigma_{x_i}}{\mu_{x_i}}$). It measures for the individual indicator dispersion, the higher the coefficient of variation the higher the dispersion;

- *skewness*: computes the sample skewness of the individual indicator x_i using the formula $\frac{1}{n}\sum_i((x_i - \mu_{x_i})^3)/(\sigma^3_{x_i})$;

- *kurtosis*: computes the sample kurtosis of the individual indicator x_i as $\frac{\frac{1}{n}\sum_i((x_i-\mu_{x_i})^4)}{\sigma^4_{x_i}}$;

- *skewness correction*: it is a dummy variable that indicates whether it is necessary to correct for an asymmetry. When the absolute value of skewness is larger than 2 and kurtosis is larger than 3.5, the data include outliers, which are treated;

- *maximum value*: takes the maximum value of the individual indicator;

- *country corresponding to maximum value*: shows the country label holding the maximum value for the indicator;

- *minimum value*: takes the minimum value of the individual indicator;

- *country corresponding to minimum value*: shows the country label of the country corresponding to the minimum.

Except for the % *of missing values* statistics, all the summary statistics in Appendix A were computed on the data as used for the computation of the final DESI 2022, thus on the data set with imputed data. As is commonly understood, for the computation of the missing value statistic a different data set was used, which contained the missing and censored data. The summary statistics contained in Appendix A help in having a first idea of countries'

behaviour within each individual indicator. Following, for each dimension in DESI 2022, individual indicators missing patterns and asymmetry will be discussed.

For HC dimension, looking at line % *of missing values* in Table A.2, we can observe that the only indicator having missing data is "ICT graduates" (1b4), with 3.70% of missing values. The rate corresponds to *Czech Republic* and the data point was imputed using the growth rate method. Relatively to asymmetric behaviors of individual indicators, Tables A.1 and A.2 instead, show that for the HC dimension the data do not exhibit a particularly high dispersion, since all the individual indicators in the HC dimension hold a quite low coefficient of variation. Kurtosis is high for most of the indicators but skewness does not cross the threshold value of 2 for any, thus no skewness correction is applied to outliers of any indicator.

In the CN dimension, we find in Table A.4 that the percentage of missing values is 3.70% for "At least 1 Gbps take-up" (2a3) indicator. In this case, *Croatia* is the country corresponding to the missing data but imputation was not performed. In fact, the missing data was later suggested to be equal to 0 by Croatia itself. Referring to Tables A.4, A.5 and A.6 we can note that the data does not show significant dispersion, as all the simple indicators in CN have relatively low coefficients of variation. In this dimension, for one indicator, i.e. "At least 1 Gbps take-up" (2a3), both kurtosis and skewness cross the threshold values. However, for this indicator, upon the decision of DESI experts, no correction was applied. In fact, outlying observations, corresponding to countries *France* and *Hungary*, were considered representations of excellent performances, thus very relevant in their respective final score computations.

Concerning the IDT dimension instead, a larger number of individual indicators show missing values as we can see by looking at the % *of missing values* in Tables A.9 and A.10. "ICT for environmental sustainability" (3b6), where the missing values for *Cyprus* and *Malta* were imputed using the rank method; "e-Invoices" (3b7) has a missing value for *Greece*, that was

imputed using the growth rate method; "e-Commerce turnover" (3c2) indicator where the missing values for *Finland*, *Luxembourg*, and *Poland* were all imputed using the growth rate method. By Tables A.8, A.9 and A.10, indicators "AI" (3b5) and "e-Invoices" (3b7) exhibit the highest coefficient of variation. However, no indicator was corrected for skewness.

Finally, the DPS dimension does not contain any missing value for the 2021 data, and the summary statistics contained in Tables A.12, A.13 show that while kurtosis crosses the 3.5 threshold for indicators "e-Government users" (4a1) and "Digital public services for businesses" (4a4), skewness is low on those thus no correction is applied.

While missing values have been treated, according to imputation methods in Section 2.2, no correction for skewness was adopted even though some outlying behaviors of single indicators were observed. However, the adoption of Robust PCA (Hubert, Rousseeuw, and Vanden Branden 2005) in Chapter 4 will help with this choice by addressing the two minor goals of the work, i.e. testing the robustness of Classical PCA against outliers and detecting outlying countries in a multivariate space.

2.4 Data normalization

Normalization is a crucial step in the construction of composite indicators as it allows for the comparison and aggregation of indicators with different measurement scales. In DESI 2022, the method used to reduce data to the same scale before aggregating into the final CI is *Min-Max scaling* (Nardo et al. 2008). The method normalizes indicators to have an identical range $[0, 1]$. In fact, for all original individual indicators x_i, with $i = 1, \ldots, p$, the 0 value in the normalized scale is anchored to the minimum value, $x_{min,i}$, in the indicator original scale, and the value 1 in the normalized scale was anchored to the maximum value, $x_{max,i}$, in the indicator's scale. All the individual indicators are then normalized, in order to be positively oriented with levels of digital development. The normalised value X_{ic}, for country c on indicator

x_i, is obtained according to the following transformation:

$$X_{ic} = \begin{cases} \frac{x_{ic} - x_{min,i}}{x_{max,i} - x_{min,i}} & \text{if indicator } x_i \text{ is positively oriented} \\ 1 - \frac{x_{ic} - x_{min,i}}{x_{max,i} - x_{min,i}} & \text{if indicator } x_i \text{ is negatively oriented} \end{cases}$$

for $i = 1, \ldots, p$, and $c = 1, \ldots, n$, with n the total number of observed countries.

Extreme values or outliers could distort the transformed indicator. On the other hand, depending on the choice of minimum and maximum values, min-max normalization could widen the range of indicators lying within a small interval, increasing the effect on the composite indicator more than the z-score transformation (Nardo et al. 2008). To allow for inter–temporal comparisons of index scores in the case of time-dependent studies, the minima ($x_{i,min}$) and maxima ($x_{i,max}$) across countries for the normalization of each indicator, are calculated for a reference year (usually the initial time point).

In DESI 2022 (source Methodological Note, European Commission 2022), minimum and maximum values were fixed based on the 2019 data and were computed as follows:

- Minimum: actual minimum value in the basket multiplied by 0.75.

- Maximum: actual maximum value in the basket multiplied by 1.25.

The multiplicators ensure that actual values do not go below the minimum and above the maximum values over time. Minimum and maximum values have not been updated based on the 2020 and 2021 data to avoid updating 2019 figures. The indicator "Broadband price index" (2d1) is normalized to a score between 0 and 100, where 100 is the best performance.

2.5 DESI 2022: weights and scores computation

A composite indicator using an additive method, based on *weighted mean* as aggregation technique, as the case of DESI 2022, takes the form

$$\boldsymbol{I} = \sum_{i=1}^{p} W_i X_i, \tag{2.3}$$

where $\boldsymbol{I}$ indicates the final composite index, X_i the ith normalised variable; W_i the corresponding weight at the overall score level, with $\sum_{i=1}^{p} W_i = 1$ and $0 \leq W_i \leq 1$, $i = 1, \ldots, p$ and p the total number of involved variables.

The *final weights* W_i, for each individual ith indicator, are computed in three steps, following the hierarchical structure of DESI 2022:

- **At a sub-dimensional level**, each KPI of the Digital Decade is initially assigned a weight of 2, while all other individual indicators receive a weight of 1, denoted as $\boldsymbol{w_{i,init}}$, where $i = 1, \ldots, p$ and k represents the sub-dimension to which the indicator belongs. This choice aims at giving double importance to KPIs when compared to the other individual indicators within the same sub-dimension. Next, the initial weights $w_{i,init}$ of each individual indicator are normalized within their respective kth sub-dimension, ensuring that they collectively sum up to 1. This process yields the normalized weight for each indicator $\boldsymbol{w_{i,sub=k}}$:

  $$w_{i,sub=k} = \frac{w_{i,init}}{\sum_{i \in k} w_{i,init}}. \tag{2.4}$$

  with $0 \leq w_{i,sub=k} \leq 1$.

- **At a dimensional level**, each sub-dimensional ith indicator weight $w_{i,sub=k}$ is multiplied by the weight assigned to the kth sub-dimension of the jth dimension, $u_{sub=k,dim=j}$, where $\sum_{k \in j} u_{sub=k,dim=j} = 1$. The sub-dimensional weights are differently assigned: EW is used for sub-dimensions in HC, while in the other dimensions, sub-dimensions are differently weighted according to the experts' opinion. In this way the *dimensional ith indicator weight* is obtained as:

$$\boldsymbol{w_{i,dim=j}} = u_{sub=k,dim=j} \times w_{i,sub=k} \tag{2.5}$$

The $w_{i,dim=j}$ weights, with $0 \leq w_{i,dim=j} \leq 1$, sum up to 1 in the respective jth dimension.

- In the end, **at the overall level**, the dimensional ith indicator weight $w_{i,dim=j}$ is multiplied by the weight assigned to the dimension, $u_{dim=j}$, through EW. In fact, all the $j = 1, \ldots, 4$ dimensions have equal importance at the DESI level. This step is returning the overall weight $\boldsymbol{W_i}$, such as:

$$W_i = u_{dim=j} \times w_{i,dim=j} \tag{2.6}$$

It holds that $0 \leq W_i \leq 1$ and $\sum_{i=1}^{p} W_i = 1$.

In the same way, the *country-specific scores* can be computed at three different levels of complexity, expressed as the aggregation of the individual indicators:

- **At a sub-dimensional score level**, scores can be obtained separately for each kth sub-dimension, with $k = 1, \ldots, 10$, as

$$\boldsymbol{i_{sub=k}} = \sum_{i \in k} w_{i,sub=k} X_{i,k}, \tag{2.7}$$

with $w_{i,sub=k}$ defined in 2.4 and X_i being the ith normalized individual indicator.

- **At a dimensional score level**, a score is computed for each jth dimension, with $j = 1, \ldots, 4$, as

$$\boldsymbol{i_{dim=j}} = \sum_{i \in j} w_{i,dim=j} X_{i,k}, \tag{2.8}$$

with $w_{i,dim=j}$ defined in 2.5 and X_i being the ith normalized individual indicator.

- **At a overall level**, the score computed for DESI, $\boldsymbol{I}$, is obtained as in formula 2.3.

Finally, formula 2.3 can be rewritten in terms of dimensional scores additive aggregation as

$$\boldsymbol{I} = \sum_{j=1}^{4} u_{dim=j} i_{dim=j}, \tag{2.9}$$

with the jth dimensional score $i_{dim=j}$ defined in 2.8 and $u_{dim=j} = 25\%$. Thus, in the final score computation, the aggregation method turns into a simple average, allowing for full compensation of dimensions.

In Appendix B we can find Table B.2 containing all the necessary quantities for the score computation at the sub-dimensional and dimensional level of complexity ofDESI 2022. Specifically, we have that:

- *Sub-dim. weight*: corresponds to the previously defined weight assigned to the kth sub-dimension of the jth dimension, $u_{sub=k,dim=j}$;

- *Weight*: is the aforementioned $w_{i,init}$ and it takes value 2 if the ith individual indicator is a KPI, 1 otherwise;

- *Normalized Weight*: corresponds to $w_{i,sub=k}$ defined in formula 2.4.

Chapter 3

Methodology

This chapter contains all the methods involved in testing for internal consistency and outlier detection in the process of definition of DESI 2022. In Section 3.1 a definition of homogeneity is provided, and the internal consistency concept is explained, followed by the methods employed in the testing. The theoretical frameworks of those methods and their application to DESI 2022 are provided in Section 3.2 for Cronbach's alpha and in Section 3.3 for Principal Component Analysis (PCA). Then, a robust version of PCA (ROBPCA) is introduced in Subsection 3.3.4 for the detection of outlying observations in a multivariate setting. Finally, Factor Analysis (FA) and PCA are compared in Section 3.4, specifically in the context of detecting latent variables. By undertaking this comparison, we aim to shed light on the suitability of these techniques within the realm of DESI 2022.

3.1 Homogeneity and Internal consistency

The homogeneity idea is closely related to the picture that different variables may measure the same concept (Gifi 1990). It can be studied at different degrees of complexity. One of those is through *internal consistency* which estimate refers to item homogeneity, namely the extent to which items within a group measure various aspects of the same characteristic or construct (Hen-

son 2001).

In the context of composite indicators, and specifically for DESI 2022, the interest is to test for mono-dimensionality and internal consistency within existing dimensions and sub-dimensions. The statistical methods employed to test for that and assess the validity of each indicator belonging to the respective dimension/sub-dimension are:

- *Correlation analysis* between pairs of indicators/sub-dimensions;
- *Cronbach's alpha*, a reliability coefficient that provides a method of measuring internal consistency of indicators (Cronbach 1951).
- *Principal Component Analysis*, used in an exploratory way for accounting multivariate correlation between indicators, thus accounting for their internal consistency (Jolliffe 2002).

While correlation analysis is a bivariate exploratory method to test the consistency of pairs of simple indicators, PCA is a more comprehensive method that provides information regarding the contribution that each indicator gives to the same latent construct, captured by the dimension/sub-dimension. Instead, Cronbach's alpha is a simpler statistic that only measures the average correlation between indicators in the same dimension/sub-dimension.

Additionally to the aforementioned methods, in section 3.4 Factor Analysis is compared to PCA, to show how the two substantially diverge in an internal consistency analysis setting, focusing on the reasons why the two methods should not be exchanged.

3.2 Cronbach's alpha

In the realm of social and psychological measurement, particularly in the evaluation of measurement instruments such as surveys and questionnaires, the *concept of reliability* holds paramount significance. Reliability, within the

context of internal consistency analysis, pertains to the degree of coherence and consistency exhibited by a set of items designed to gauge a common latent construct or attribute. It reflects the extent to which these items yield congruent and stable outcomes upon repeated observations. This applies to the CIs case too, where reliability assumes a critical role, underpinning the credibility and robustness of the synthesized measures. Specifically, it delineates the degree to which the constituent indicators collectively and consistently measure the underlying construct.

One widely employed method for gauging internal consistency is *Cronbach's Alpha Coefficient* (Cronbach 1951). Cronbach's alpha (α) is a simple and the most common measure of internal consistency, which assesses the average intercorrelation among the constituent indicators. It generally ranges in value from 0 to 1, the higher the score, the more reliable the generated scale is. With negative correlations between some indicators, the coefficient alpha can have a negative value. The larger the overall alpha coefficient, the more likely that the indicators all contribute to a single and then reliable latent construct. Nunnally (1978) has indicated 0.7 to be an acceptable value for internal consistency detection of the Cronbach's alpha, but lower thresholds are sometimes used in the literature depending on the different disciplines. Cronbach's alpha is defined as:

$$\alpha = \frac{p}{p-1}\Big(1 - \frac{\sum_{i=1}^{p} \sigma_i^2}{\sigma_t^2}\Big) \tag{3.1}$$

with p the number of indicators, σ_i^2 the variance associated with the ith indicator, $i = 1, \ldots, p$, and σ_t^2 the variance associated with the total score obtained by summation of the single indicators. In the context of composite indicators, Cronbach's alpha statistic quantifies the degree to which the score of an indicator is influenced by general and group factors, rather than country-specific factors (Taber 2018). Thus, Cronbach's alpha measures how well a set of indicators measures a single latent construct in a unidimensional manner. When data have a multidimensional structure, Cronbach's alpha will usually be underestimated.

Concerning DESI 2022, Cronbach's alpha will be used at a dimensional/sub-dimensional level to get an idea of the unidimensionality nature of the underlying dimensional/sub-dimensional construct, using 0.7 as a threshold value.

3.3 Principal Component Analysis

Let us assume that we observe the value of p indicators: $X_1, X_2, \ldots, X_p$ on a set of n statistical units. These values are usually organized in a $\boldsymbol{X}_{(n,p)}$ matrix made by n rows and p continuous features. The total variance of $\boldsymbol{X}_{(n,p)}$ is defined as the sum of the variances of the indicators $X_1, X_2, \ldots, X_p$.

Principal Component Analysis (PCA - Jolliffe 2002) is a descriptive, multi-variate method that aims at finding a smaller dimensional space capturing most of the total variance of the original one.

Algebraically, Principal Components (PCs) are weighted linear combinations of the (standardized) original p indicators. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by orthogonal transformation of the original system with $X_1, X_2, \ldots, X_p$ as the coordinate axes.

Since p components are required to reproduce the total variance, and PCs are inherently orthogonal, each PC is accountable for a percentage of the explained variance, which expresses the percentage of total variance that the kth principal component accounts for ($k = 1 \ldots, p$). Additionally, by construction, the variance explained by the first component is higher than the one explained by the second one, and so on for the other components. So, one looks for the k-dimensional subspace such that the projection of the data on this subspace contains most of the information of the original data.

The *score matrix*, denoted as $\boldsymbol{T}_{(n,k_{max})}$, signifies the values corresponding to each original observation across the kth component. It can accommodate up to a maximum of k_{max} components, constrained by the count of original variables p. The relation between the original data and the final projection can be written like this:

$$\boldsymbol{T}_{(n,k_{max})} = (\boldsymbol{X}_{(n,p)} - \mathbf{1_n}\hat{\boldsymbol{\mu}}')\boldsymbol{P}_{(p,k_{max})}, \tag{3.2}$$

where $\boldsymbol{P}_{(p,k_{max})}$ is the *loading matrix* which contains the eigenvectors of the estimated variance-covariance matrix, $\hat{\boldsymbol{e}}_{\mathbf{1}}, \hat{\boldsymbol{e}}_{\mathbf{2}}, \ldots, \hat{\boldsymbol{e}}_{\boldsymbol{k_{max}}}$ and $\hat{\boldsymbol{\mu}}$ is the center of the data. The loadings of the PCs, contained by column in the $p \times k_{max}$ matrix $\boldsymbol{P}$, represent the coefficients of the linear combinations. Most importantly, the loadings indicate the correlation between each simple indicator and the component.

Thus, in a CI context, taking the first PC as an expression of the main underlying latent factor, the $\boldsymbol{e}_{\mathbf{1}j}$ loading value is an expression of how the jth simple indicator relates to the main latent concept, with $j = 1, \ldots, p$. In fact, the first PC represents the directions along the maximum variability of the original indicators. Additionally, in p. 87 of Gifi (1990), it is shown how the objective of choosing scores and weights so as to maximize homogeneity of simple indicators is one of the possible definitions of finding the (first) principal component. The subsequent PCs, similarly to the first, can be expressions of ulterior latent factors, still less relevant than the first. In fact, by construction, those PCs capture decreasing portions of the total variance of $\boldsymbol{X}_{(n,p)}$, under the constraint to be orthogonal to the previous components.

Following, the loading matrix $\boldsymbol{P}_{(p,k_{max})}$ is computed, where the first column is made by

$$\hat{\boldsymbol{e}}_1 = \underset{\|\boldsymbol{e}_1\|=1}{\operatorname{argmax}} \boldsymbol{S}(\boldsymbol{e}_1'(\boldsymbol{x_1} - \hat{\boldsymbol{\mu}}), \ldots, \boldsymbol{e}_1'(\boldsymbol{x_n} - \hat{\boldsymbol{\mu}})), \tag{3.3}$$

and the other jth columns with $j = 2, \ldots, k_{max}$ are defined as

$$\hat{\boldsymbol{e}}_j = \underset{\|\boldsymbol{e}_j\|=1, \boldsymbol{e}_j \perp \hat{\boldsymbol{e}}_1, \ldots, \boldsymbol{e}_j \perp \hat{\boldsymbol{e}}_{j-1}}{\operatorname{argmax}} \boldsymbol{S}(\boldsymbol{e}_j'(\boldsymbol{x_1} - \hat{\boldsymbol{\mu}}), \ldots, \boldsymbol{e}_j'(\boldsymbol{x_n} - \hat{\boldsymbol{\mu}})), \tag{3.4}$$

where $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{S}$ are the location and scale estimators and $\boldsymbol{x_i}$ are the p-dimensional vectors of observations contained by row in $\boldsymbol{X}$.

Wwhen the original variables are measured on different scales, data are *normalized* to prevent variables with higher variance from dominating the

first principal components. It should be noted that PCA is *orthogonally equivariant*, meaning that the rotation of the data results in a corresponding rotation of the PCs. However, in the case of a half-turn rotation, only the signs of the loadings change and not the values, and thus the interpretation doesn't change. It is important to keep in mind that PCA is sensitive to variable standardization and is therefore *not affine equivariant*. As a result, the information provided by the PCs obtained from centered data and standardized data will differ. In the case of DESI 2022, PCA with internal consistency purposes has been applied over z-score normalized data.

Finally, when using PCA to test for internal consistency of indicators within a dimension/sub-dimension, the number of components k_{max} to be considered as the expression of k_{max} latent underlying factors has to be determined. This concern is very similar to a central question arising in the context of PCA used for data reduction, that is: how many components to retain? Thus, methods from the literature used for selecting the number of components, k_{max}, to retain will be applied to choosing the total number of existing underlying latent dimensions. Typically, this number is much smaller than the number of initial features p. There are various techniques to determine the exact number of components needed. One widely used approach is the *scree plot*, also known as the *elbow method*, which visualizes the calculated eigenvalues after eigendecomposition on the principal components in descending order, and searches for a drastically curving point (*elbow*) which decides the maximum number of retained components (Cartell 1966). However, selecting a clear elbow sounds too abstract and does not provide comprehensible but rather ambiguous evidence (Ledesma et al. 2015). Another widely-known criterion to choose k_{max} PCs is to look at the percentage of the total variance explained by the retained k_{max} PCs. The popular rule-of-thumb here is the *Kaiser's rule*, suggesting that the PCs whose eigenvalues exceed 1 should be retained. The assumption behind the rule is that a component explaining less variance than an original variable, equal to 1 when z-score standardization is used, is not worth retaining (Kaiser 1960). However, this

intuitive rule has also been criticized since the strict cut-off rule implies, for instance, α component with 1.01 variance would be retained while β with 0.99 variances would not (Kaufman and Dunlap 2000).

In the assessment of DESI 2022, PCA has been utilized within each dimension and sub-dimension to assess internal consistency. Following the PCA process, the Kaiser rule is employed to determine the count of latent variables present within the analyzed component. Subsequently, after identifying the number of measures that define the dimension or sub-dimensions, the loading of each indicator is leveraged as a representation of the correlations between individual indicators and the PCs of interest.

3.3.1 PCA: Eigendecomposition

PCA as its core employs *eigendecomposition*, a mathematical process that uncovers the inherent structure of data. Eigendecomposition begins by calculating the *covariance matrix* $\mathbf{\Sigma}$, which captures the relationships and variability between different variables in the dataset. Through this matrix, we extract eigenvalues and their corresponding eigenvectors. Eigenvalues signify the amount of variance present in each eigenvector's direction, while eigenvectors represent the orientation of maximum variance. The key lies in selecting the most significant eigenvectors, known as principal components. These principal components define new coordinate axes that align with the directions of maximum data variability, as noted in Formula 3.4.

A classical result for eigendecomposition used in a PCA setting is shown in Johnson and Wichern (1998). Assuming that the real variance of the data is available, namely $\mathbf{\Sigma}$, its eigendecomposition is provides us with one factoring of the covariance matrix into its canonical form. Let $\mathbf{\Sigma}$ have eigenvalue-eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)$ with $\lambda_1 \geq \lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Then

$$\mathbf{\Sigma} = \boldsymbol{Q}_{(p,p)} \mathbf{\Lambda}_{(p,p)} \boldsymbol{Q}'_{(p,p)} \tag{3.5}$$

with $\boldsymbol{Q}$ the square matrix whose jth column is the eigenvector $\boldsymbol{e_j}$ of $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding

eigenvalues, $\Lambda_{jj} = \lambda_j$.

The Formula 3.5 can be translated in estimated quantities as shown by Johnson and Wichern (1998), where instead the *sample covariance matrix* is used. Let $\boldsymbol{X}$ denote a $n \times p$ matrix of the data, $\hat{\boldsymbol{\Sigma}}$ its sample covariance matrix and $\hat{\boldsymbol{\mu}}$ the sample p-dimensional mean vector. Let $\hat{\boldsymbol{\Sigma}}$ have eigenvalues-eigenvector pairs $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), \ldots (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$. By applying an eigendecomposition to the symmetric estimated covariance matrix, $\hat{\boldsymbol{\Sigma}}$ is factorized into its canonical form, which is

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{V}_{(p,p)} \hat{\boldsymbol{\Lambda}}_{(p,p)} \boldsymbol{V}'_{(p,p)}, \tag{3.6}$$

with $\boldsymbol{V}$ the square matrix whose jth column is the eigenvector $\hat{\boldsymbol{e}}_{\boldsymbol{j}}$ of $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Lambda}}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\hat{\Lambda}_{jj} = \hat{\lambda}_j$.

The jth principal component, is given by the n-dimensional vector

$$Y_j = \hat{\boldsymbol{e}}'_j \boldsymbol{X}, \qquad j = 1, \ldots, p. \tag{3.7}$$

With this choices, it follows that

$$\begin{aligned} \mathrm{Var}(Y_j) &= \hat{\boldsymbol{e}}'_j \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{e}}_j = \hat{\lambda}_j, \qquad j = 1, \ldots, p, \\ \mathrm{Cov}(Y_j, Y_k) &= \hat{\boldsymbol{e}}'_j \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{e}}_k = 0, \qquad j \neq k. \end{aligned}$$

A fundamental result of PCA is that the sum of the principal component variances is equal to the sum of the estimated variances of the original variable, thus

$$\hat{\sigma}_{11} + \hat{\sigma}_{22} + \cdots + \hat{\sigma}_{pp} = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(Y_i) \tag{3.8}$$

Fixing $k = k_{max}$, with $k_{max} < p$, the subspace spanned by the loading vector, the score matrix $\boldsymbol{T}_{(n,k_{max})}$ is obtained accordingly to Formula 3.2, with $\boldsymbol{T}_{(n,k_{max})} = (Y_1 \ldots Y_{k_{max}})$. Furthermore, every row of $\boldsymbol{T}_{(n,k_{max})}$ contains the scores $\boldsymbol{t_i}$, which represents the ith observation in the k_{max}-dimensional subspace spanned by the loading vectors.

If data are centered, the ith score is formuled as follows

$$\boldsymbol{t}_i = \boldsymbol{P}'_{(k_{max},p)} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}). \tag{3.9}$$

where $\boldsymbol{x}_i$ is a p-variate vector denoting the ith observation (ith row of $\boldsymbol{X}$). From Formula (3.2) and obtained $\boldsymbol{P}_{(p,k_{max})}$ from $\hat{\boldsymbol{\Sigma}}$ eigendecomposition (Formula 3.6), it follows that an estimate for matrix $\boldsymbol{X}$, namely $\hat{\boldsymbol{X}}$, the $n \times p$ predicted values matrix for each observation is given by

$$\hat{\boldsymbol{X}}_{(n,p)} = \boldsymbol{T}_{(n,k_{max})}\boldsymbol{P}'_{(k_{max},p)}. \tag{3.10}$$

If data are centered, the prediction for the ith observation $\hat{\boldsymbol{x}}_i$ assumes the following formulation

$$\begin{aligned} \hat{\boldsymbol{x}}_{i,k_{max}} &= \boldsymbol{P}_{(p,k_{max})}\boldsymbol{t}_i + \hat{\boldsymbol{\mu}} \\ &= \boldsymbol{P}_{(p,k_{max})}\boldsymbol{P}'_{(k_{max},p)}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}) + \hat{\boldsymbol{\mu}}. \end{aligned} \tag{3.11}$$

3.3.2 PCA: Singular Value Decomposition

In Wall et al. (2002) it is shown that instead of passing through the $\hat{\boldsymbol{\Sigma}}$ eigendecomposition for finding eigenvalues-eigenvectors pairs, an alternative is using a Singular Value Decomposition (SVD - Golub and Van Loan 1996) on the centered data matrix.

Let $\boldsymbol{X}_c = (\boldsymbol{X} - \boldsymbol{1_n}\hat{\boldsymbol{\mu}}')$ now denote the $n \times p$ matrix of the centered data. The equation for singular value decomposition of $\boldsymbol{X}_c$ is the following:

$$\boldsymbol{X}_c = \boldsymbol{U}_{(n,r)}\boldsymbol{D}_{(r,r)}\boldsymbol{V}'_{(r,r)}, \tag{3.12}$$

where $\boldsymbol{U}$ is a unitary matrix such that $\boldsymbol{U}'\boldsymbol{U} = I_r$; $\boldsymbol{D}$ is the diagonal matrix with elements on the diagonal $d_j^{1/2}$ $(j = 1, \ldots, r)$ and r is the rank of $\boldsymbol{X}$. The columns of $\boldsymbol{U}$ are called the left singular vectors, the rows of $\boldsymbol{V}'$ contain the elements of the right singular vector while the element of $\boldsymbol{D}$ are only nonzero on the diagonal and are called the singular values. We know that $\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\boldsymbol{X'}_c\boldsymbol{X}_c$ and by substituting $\boldsymbol{X}_c$ decomposition as in (3.12), we get that

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{V}\Big(\frac{\boldsymbol{D^2}}{\boldsymbol{n-1}}\Big)\boldsymbol{V}'.$$

When $p < n$, SVD can provide computationally efficient methods to find the PCs. However, when $p >> n$, deriving the eigenvectors from $(\boldsymbol{X} -$

$\mathbf{1}_n\hat{\boldsymbol{\mu}})(\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}})'$ will take less computational time than performing SVD on the centered data matrix.

In the case of DESI 2022, when PCA is applied at any dimensional/sub-dimensional level, it always holds that $p < n$ and thus SVD is the most efficient method. The only case when SVD is not preferred is when PCA is used over all $p = 33$ indicators of DESI, thus $p > n$ and decomposing $\hat{\Sigma}$ is slightly more efficient than using SVD.

3.3.3 Classical Principal Component Analysis

Classical PCA (CPCA - Jolliffe 2002) centers $\boldsymbol{X}$ with its column-wise mean vector $\hat{\boldsymbol{\mu}}$, and uses the classical variance of the projected data ad the scale measure in sections 3.3 and 3.4. Equivalently CPCA successively solves

$$\text{maximize} \quad \boldsymbol{e}_j'\hat{\boldsymbol{\Sigma}}\boldsymbol{e}_j \tag{3.13a}$$

subject to

$$\boldsymbol{e}_j'\boldsymbol{e}_j = 1, j = 1, \ldots, k, \tag{3.13b}$$

$$\boldsymbol{e}_j'\boldsymbol{e}_i = 0, j>>2, i<j \tag{3.13c}$$

where $\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}(\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}')'(\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}')'$ is the sample covariance matrix of the data. The eigenvalues are equal to the variance of the projections in the corresponding directions. Considering Formula 3.12, one can also obtain the PC solutions by using the SVD of the centered data matrix

$$\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'. \tag{3.14}$$

Then the squared jth diagonal element of $\boldsymbol{D}$, divided by $n-1$, i.e. $\frac{d_j}{n-1}$, is the jth eigenvalue of $\hat{\boldsymbol{\Sigma}}$, with the jth column of $\boldsymbol{V}$ as the corresponding PC direction. The score of the ith observation is given by $\boldsymbol{t}_i = \boldsymbol{u}_i'$, with $\boldsymbol{u}_i'$ the ith row of $\boldsymbol{U}$.

3.3.4 Robust Principal Component Analysis

Robust statistics arise from the need of providing methods that are robust against the possibility that some outlying behaviours may occur in the data. In the context of composite indicators where there are very few observations and each one is fundamental in computing each country score, researchers may not be satisfied with removing outlying observations. Instead, in a multivariate setting a *robust approach for PCA* can help in detecting possible outlying observations that are otherwise hidden in classical methods. Robust PCA can also show the extent to which hidden outliers affect the classical PCA outputs. In Hubert, Rousseeuw, and Vanden Branden (2005) it is argued the issue of anomalous observations in classical PCA and how it can lead to unreliable latent factor detection. Specifically, it explains how the classical variance and covariance matrix are sensitive to outliers, which can attract the first components toward them and fail to capture the variation of regular observations. To address this issue, in this project the robust PCA methods are compared to classical PCA in the internal consistency analysis of DESI 2022. The aim of the comparative analysis is to detect outlying observations which possibly affect the skewness of the PCA first components outputs. A lot of work has been done on robust PCA models after the development of robust location and scatter estimators. There are two main approaches to creating robust PCA: the covariance-based PCA and the ROBPCA. Covariance-based robust PCA and ROBPCA differ in their approach to handling outliers in data analysis. While both methods aim to extract principal components, covariance-based robust PCA employs a modified covariance matrix to reduce the impact of outliers on principal component estimation. In contrast, ROBPCA employs robust statistics and optimization techniques explicitly designed to detect and down-weight outliers, leading to more accurate and reliable principal component extraction.

In a covariance-based PCA setting, as see in Subsection 3.3.1, the loadings of classical PCA are obtained by spectral decomposition of the covariance matrix. On the other hand, a covariance-based *robust* PCA approach

can be referred to as a *plug-in method*, where the empirical covariance matrix is replaced with a robust multivariate scatter estimator, and then performs spectral decomposition to get the robust loading vectors. Different robust scatter estimators can be used, like multivariate S, MM-estimator, or Minimum Covariance Determinant (MCD) (see Hubert, Rousseeuw, and Aelst 2008 for additional information). The most popular and used estimator is the MCD estimator (Rousseeuw 1984) in particular, involved when $p << n$. This is the case of internal consistency within dimension/sub-dimensions of DESI 2022, where the number of observed countries n remains fixed at 27, and it is never exceeded by the number of individual indicators p in each dimension/sub-dimensions.

The MCD method looks for the h observations (out of n) whose classical covariance matrix has the lowest possible determinant. The MCD location and scatter estimator is then the mean and the covariance matrix of those h observations ($[n + p + 1]/2 \leq h \leq n$). The first k eigenvectors of the MCD covariance matrix, sorted in descending order with respect to eigenvalues, provide robust loadings for the PCA procedure. MCD is affine equivariant, so the estimator transforms well under any non-singular reparametrization of the space of initial data. Consequently, data can be rotated, translated, or rescaled without affecting the outlier detection diagnostics. Moreover, MCD can reach a breakdown value of 50% by taking $h = [n + p + 1]$, where the If α is the percentage of total observations considered: $\alpha = h/n$, we reach a breakdown value equal to 50% when $\alpha = 0.5$, where the *breakdown value* signifies the minimum proportion of impurities within a sample that leads the estimator to malfunction, resulting in it yielding values that are arbitrarily unfavorable or lack meaningful interpretation (Hubert and Debruyne 2009). Thus, the breakdown value is a popular measure of the robustness of an estimator against outlying observations and a higher breakdown value implies greater robustness, as the procedure can withstand a larger percentage of contaminated data while still producing reliable results. In this work, we take $\alpha = 0.75$ which gives accurate results if the dataset contains at most

25% of deviant values, which is a reasonable assumption for most datasets (Verboven and Hubert 2005). One of the main goals of robust statistics is to find a trade-off between robustness and efficiency. The efficiency of the MCD estimator increases with an increment of α, but at the expense of robustness. In fact, the breakdown value decreases with an increasing α, the percentage of total observations considered within the estimate. Yet, existing methods i.e. reweighted MCD estimator, assure high efficiency of the MCD method while guaranteeing a high breakdown value (Gervini 2003).

However, in a high-dimensional setting ($p > n$), it is not possible to use the MCD estimator because the determinant of a covariance matrix of $h < p$ observations will always be zero and thus cannot be minimized. This applies to DESI 2022 when robust PCA is used over all $p = 33$ simple indicators, while n is fixed to 27 countries. The robust PCA, ROBPCA method (Hubert, Rousseeuw, and Vanden Branden 2005), circumvents this problem by combining projection pursuit ideas in the high-dimensional space with MCD estimation in a lower-dimensional subspace.

3.3.5 Outlier detection

To detect outliers in a robust PCA analysis, plots based on two different kinds of distances are built, namely the *score distance* and the *orthogonal distance* (Hubert, Rousseeuw, and Vanden Branden 2005).

Any PCA model will result in estimates of the location $\hat{\boldsymbol{\mu}}$, a loading matrix $\boldsymbol{P}_{(p,k_{max})}$ a score matrix $\boldsymbol{T}_{(n,k_{max})}$ and a diagonal matrix $\boldsymbol{\Lambda}_{(k_{max},k_{max})}$ with the eigenvaules $\lambda_{k_{max}}$ in decreasing order as the kth diagonal elements.

With these values, one can predict an observation $\boldsymbol{x}_i$ using its orthogonal projection on the PCA subspace: $\hat{\boldsymbol{x}}_i = \boldsymbol{P}_{(p,k_{max})}\boldsymbol{t}_i + \hat{\boldsymbol{\mu}} = \boldsymbol{P}_{(p,k_{max})}\boldsymbol{P}'_{(k_{max},p)}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}) + \hat{\boldsymbol{\mu}}$ (see Equation 3.11). The *score distance* (SD) measures the robust statistical distance from the scores for a specific observation to the center within the k-dimensional subspace. When CPCA is performed, the distance corresponds to the Mahalanobis distance. For an observation $\boldsymbol{x}_i$, the score distance

is defined as

$$\mathbf{SD}_i = \sqrt{\sum_{j=1}^{k_{max}} \frac{(t_{ij})^2}{\lambda_j}} = \sqrt{\boldsymbol{t}_i' \boldsymbol{\Lambda}^{-1} \boldsymbol{t}_i}, \tag{3.15}$$

where t_{ij} is the jth component of the score vector $\boldsymbol{t}_i$ of observation $\boldsymbol{x}_i$.

The *orthogonal distance* (OD) measures the Euclidean distance of an observation to the estimated PCA subspace. For an observation $\boldsymbol{x}_i$, the orthogonal distance is defined as

$$\mathbf{OD}_{i,k_{max}} = \| \boldsymbol{x}_i - \hat{\boldsymbol{x}}_{i,k_{max}} \| = \| \boldsymbol{x}_i - (\boldsymbol{P}_{(p,k_{max})} \boldsymbol{t}_i + \hat{\boldsymbol{\mu}}) \| \tag{3.16}$$

The *diagnostic plot*, or *outlier map*, developed by Hubert, Rousseeuw, and Vanden Branden (2005) is a useful tool to identify multivariate outliers. It plots the score distances on the horizontal axis and the orthogonal distances on the vertical axis with two corresponding cut-off lines. The cut-off value on the horizontal axis is $c_{SD} = \sqrt{\chi^2_{k;0.975}}$, with $\chi^2_{k;0.975}$ the 97.5% quantile of a chi-squared distribution with k degrees of freedom. It is justified by the fact that the score distances are approximately normally distributed and so the squared distances are approximately χ^2_k-distributed. As for the orthogonal distance, Box (1954) proposed that the unknown distribution of the squared orthogonal distances can be well approximated by a scaled chi-squared distribution $g_1 \chi^2_{g_2}$ with g_2 degrees of freedom. The unknown parameters can be estimated by the method of moments (Nomikos and MacGregor 1995). Based on the Wilson-Hilferty approximation for a chi-squared distribution, Hubert, Rousseeuw, and Vanden Branden (2005) proposed to approximate the distribution of orthogonal distances to the power 2/3 with a normal distribution with mean $\mu = (g_1 g_2)^{(\frac{1}{3})}(1 - \frac{2}{9g_2})$ and variance $\sigma^2 = \frac{2g_1^{\frac{2}{3}}}{9g_2^{\frac{1}{3}}}$ and use $c_{OD} = (\hat{\mu}_{MCD} + \hat{\sigma}_{MCD} z_{0.975})^{(\frac{3}{2})}$, with $z_{0.975}$ the 97.5% quantile of the standard normal distribution, as the cut-off value on the vertical axis, where $\hat{\mu}_{MCD}$ and $\hat{\sigma}_{MCD}$ are the estimates of μ and σ by the univariate MCD estimator of location and scale where there are possible outliers.

On the left side of Figure 3.1 is shown a spatial representation of a set of observations (not related to DESI 2022), the plane that is orthogonal to

those, their orthogonal distances to the plane and their projections on it. On the respective diagnostic plot, see Figure 3.1 (on the right), one can identify three different types of outliers: the good leverage points, the bad leverage points, and orthogonal outliers.

The *good leverage points* are the observations close to the PCA subspace while their scores are outlying within the PCA subspace, i.e. with $OD \leq c_{OD}$ but $SD > c_{SD}$ (observations 1 and 2 in Figure 3.1). The *bad leverage points* are defined as the observations with both outlying orthogonal distances and score distances, i.e. with $SD > c_{SD}$ and $OD > c_{OD}$ (observations 4 and 5 in Figure 3.1). Finally, the *orthogonal outlier* are observations which are far from the subspace but can not be identified within the subspace, i.e. with $SD \leq c_{SD}$ and $OD > c_{OD}$ (observations 3 in Figure 3.1). We call the good leverage points *good* because they only influence the estimates of location and scatter measure within the k-dimensional subspace. However, the orthogonal outliers and bad leverage points will shift or tilt the estimated subspace away from the k-dimensional subspace which fits the true data structure.

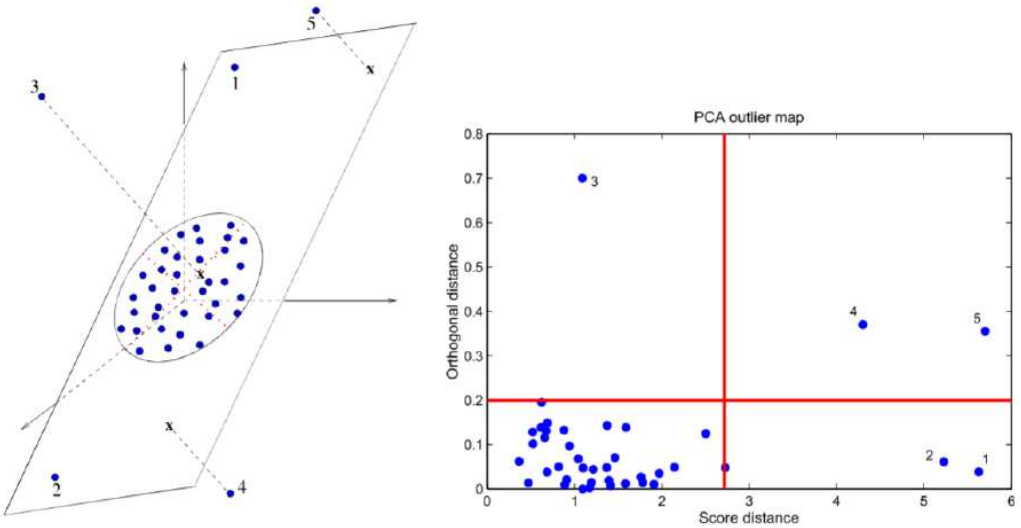


Figure 3.1: Types of outliers with respect to a two-dimensional estimated plane from a three-dimensional dataset. (Hubert, Rousseeuw, and Vanden Branden 2005)

In our analyses, score distances and orthogonal distances for both CPCA and ROBPCA are juxtaposed when using diagnostic plots. In fact, CPCA could either treat outliers as part of the natural variability or be less effective in detecting them compared to ROBPCA. Comparing the two plots helps in recalling which and in which ways countries affect the estimated subspaces. In fact, with the ROBPCA procedure, the estimated principal components result in being less influenced by outliers, helping in shedding light on countries, i.e. *orthogonal outliers*, representing cases that might exhibit behaviour not captured by the main principal components. They could indicate distinct subgroups or unexplained variations. Countries detected as *good leverage points*, instead only exert a modest influence on the principal components, thus slightly affecting the direction and magnitude of the principal components (this is reflected by comparing the ROBPCA and CPCA results). Finally, when *bad leverage points* are spotted in the ROBPCA diagnostic plot but not in the CPCA, that might be an indication of poor robustness and sensitivity of the CPCA estimates, thus of inferior efficacy of the classical procedure for latent factors detection.

We can conclude that the ROBPCA procedure leads to identifying outlying measurements that are otherwise hidden in classical methods. In reality, researchers may not be satisfied with removing outlying observations, especially in the context of CIs, where the dataset consists of a limited number of observations (in the case of DESI, $n = 27$) and the prohibition of missing values for any country requires imputation. In contrast, the DESI experts could express interest in flagging these atypical observations. This approach would serve the purpose of pinpointing countries with outlier data for reporting to the relevant authorities responsible for data collection in subsequent years. Moreover, the identification of these outlier countries opens avenues for a more thorough analysis of the individual indicator performances. Through nuanced examinations across various levels of complexity, it becomes possible to discern the underlying reasons behind these deviations.

3.4 Latent variable analysis: PCA vs FA

In a *latent variable analysis*, one tries to estimate the number of potential latent variables in an observable response pattern, i.e. *exploratory analysis*, or to test hypotheses about expected latent variables in a response pattern, i.e. a *confirmatory analysis* (Gruijters 2019). Given the composite indicators domain, FA and PCA methods are here discussed and analyzed to investigate their appropriateness in a homogeneity analysis setting, with a particular focus on DESI 2022.

The essential purpose of FA is to describe the covariance relationship among many items in terms of a few underlying, but unobservable, quantities called factors. The factor model is motivated by the following argument: suppose indicators can be grouped by their correlations. That is, suppose all indicators within a particular group are highly correlated among themselves but have relatively small correlations with indicators in different groups. Then it is conceivable that each group of indicators represents a single underlying construct or factor, that is responsible for the observed correlations. It is this type of structure that FA seeks to confirm.

PCA is instead concerned with explaining the variance-covariance structure of a set of indicators through a few linear combinations of those indicators (see 3.3). Thus PCA often reveals relationships that were not previously suspected and might be highly explanatory, and hard to detect by looking at correlation only. In the context of composite indicators, PCA can be considered a useful method to test for the internal consistency of simple indicators when a conceptual framework to rely on is missing. In fact, PCA uses its purely exploratory nature through components, that allow for the detection of potential latent variables, without assuming any model. Considering that latent variables are commonly attributed as the underlying causes behind observable data patterns, such as indicator correlations when using PCA such interpretation on components is lost. This is because PCA methodology frames constructs as being causally influenced by indicators, rather than

the conventional viewpoint where indicators are influenced by latent variables (Edwards and Bagozzi 2000).

FA instead is more suitable in case of validation purposes, as endowed with a more interesting casual interpretation, that is, the latent variable is conceptualized as a cause of response variation on observable indicators. Moreover, factor rotation allows finding the most suitable configuration for data, so that they are consistent with a prescribed structure. For these reasons, FA is properly used when validating psychological scales but caution must be paid in the context of internal consistency of composite indicators. In fact, in this context, if no conceptual model is given, PCA serves the purpose of testing for dimensional/sub-dimensional homogeneity without the need for a factor latent model and factor rotation, which are instead the main features of FA.

3.5 Factor Analysis

Factor analysis is here described just to understand the previously presented intrinsic differences with PCA, thus not applied to DESI 2022. Different methods of estimation exist for parameters of the factor model, the most popular are the principal component method and the maximum likelihood method. The solution from either method can be rotated in order to simplify the interpretation of factors. Following, we will only report the PC method since our aim is to compare FA to PCA method.

Let's assume the same model for data as in section 3.3. For each of the observed n statistical units, a $p \times 1$ random vector $\boldsymbol{X}$ exists, with p components, mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The *factor model* (Johnson and Wichern 1998) postulates that $\boldsymbol{X}$ is linearly dependent upon a few unobservable random variables $F_1, F_2, \ldots, F_m$, called *common factors*, and p additional sources of variation $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ called *specific factors*. In particular, the factor analysis model is reported in matrix notation as

$$\boldsymbol{X}_{(p,1)} - \boldsymbol{\mu}_{(p,1)} = \boldsymbol{L}_{(p,m)}\boldsymbol{F}_{(m,1)} + \boldsymbol{\varepsilon}_{(p,1)} \tag{3.17}$$

The coefficient l_{ij} is called the loading of the ith variable on the jth factor, so the matrix $\boldsymbol{L}$ is the *matrix of factor loadings*; matrix $\boldsymbol{F}$ contains the independent variables representing the common factors. The random variables $F_1, F_2, \ldots, F_m, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are unobservable and thus require some assumptions for direct verification of the factor model. We assume that

$$E(\boldsymbol{F}) = \mathbf{0}_{(m,1)},\ \mathrm{Cov}(\boldsymbol{F}) = E[\boldsymbol{F}\boldsymbol{F}'] = \boldsymbol{I}_{(m,m)}$$

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}_{(p,1)},\ \mathrm{Cov}(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \boldsymbol{\Psi}_{(p,p)} = \begin{bmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_3 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & & \ldots & \psi_p \end{bmatrix}$$

and that $\boldsymbol{F}$ and $\boldsymbol{\varepsilon}$ are independent, so

$$\mathrm{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{F}) = E[\boldsymbol{\varepsilon}\boldsymbol{F}'] = \mathbf{0}_{(p,m)}.$$

These assumptions and the relation 3.17 constitute the *orthogonal factor model* (OFM). The OFM implies that

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi} \tag{3.18}$$

and also

$$\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{F}) = \boldsymbol{L}. \tag{3.19}$$

From 3.18 follows that the total variance for the ith variable $\mathrm{Var}(X_i) = \sigma_{ii}$, the ith element on the diagonal of $\boldsymbol{\Sigma}$, can be split into two components:

$$\sigma_{ii} = \underbrace{l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2}_{\text{communality}} + \underbrace{\psi_i}_{\text{specific variance}} \tag{3.20}$$

- *communality*: the portion of variance contributed by the m common factors;

- *specific variance*: the portion of variance due to the specific factor.

Given m the number of common factors, with $m > 1$, there is always some inherent ambiguity associated with the factor model. Let $\boldsymbol{T}$ be any $m \times m$ orthogonal matrix, so that $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}$. Then the expression 3.17 can be written

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{L}^*\boldsymbol{F}^* + \boldsymbol{\varepsilon} \tag{3.21}$$

Since

$$E(\boldsymbol{F}^*) = \boldsymbol{T}'E(\boldsymbol{F}) = \boldsymbol{0}$$

and

$$\mathrm{Cov}(\boldsymbol{F}^*) = \boldsymbol{T}'\mathrm{Cov}(\boldsymbol{F})\boldsymbol{T} = \boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}$$

it is impossible, on the basis of observations on $\boldsymbol{X}$, to distinguish the loadings $\boldsymbol{L}$ from the loadings $\boldsymbol{L}^*$. That is, the factors $\boldsymbol{F}$ and $\boldsymbol{F}^*$ have the same statistical properties and even though the loadings $\boldsymbol{L}$ and $\boldsymbol{L}^*$ are in general different, they both generate the same covariance matrix $\boldsymbol{\Sigma}$. That is,

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi} = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{L}' + \boldsymbol{\Psi} = (\boldsymbol{L}^*)(\boldsymbol{L}^*)' + \boldsymbol{\Psi}$$

This ambiguity provides the rationale for *factor rotation* since orthogonal matrices correspond to rotations of the coordinate system of $\boldsymbol{X}$. Usually, rotation is determined by some *ease-of-interpretation* criterion.

Chapter 4

Internal consistency assessment

4.1 An homogeneity assessment of DESI 2022

The DESI 2022 CI design reflects two specific concerns. The first concern is related to comparing the effectiveness of the Digital Decade policy across all countries, regardless of their level of digital development, by referring to the EU-level targets, which are the Key Principal Indicators (KPIs). This goal is outlined by choices of political nature and reflected by the weighting system used for individual indicators, recalled in the initially assigned weight of 2 for KPIs and 1 for the rest of individual indicators (Section 2.5). In this work, no further investigation of the Digital Decade KPI is done. While the initially assigned weights are kept fixed as in DESI 2022, it is further investigated through sensitivity analysis (SA) the assigned sub-dimensional weights when a framework update of sub-dimensions is introduced. The application of different sub-dimensional weighting sets during the aggregation process allows for evaluation of the extent the final scores and rankings are affected by those, thus the efficacy of the Digital Decade policy. Indeed, as a measure of digital advancement, DESI will yield distinct country scores and rankings in response to each variation (see Chapter 5).

The second concern is to comply with the four cardinal points under

which the KPIs are arranged. These cardinal points are represented by the four dimensions of digital capacity, i.e. Human Capital (HC), Connectivity (CN), Integration of Digital Technologies (IDT), and Digital Public Services (DPS). To achieve this second objective, in Chapter 4 we take the compensatory aggregation rules described in Section 2.5 for DESI 2022 as fixed, i.e. Equal Weighting (EW) of dimensions and different weighting schemes based on experts' opinions at the sub-dimensional level. Then, we validate the aggregation methodological choice through an internal consistency analysis over sub-dimensions of DESI 2022. In fact, when a solid conceptual framework is lacking, as in the case of DESI 2022, it is expected for the CI reliability to increase with the internal consistency among components, i.e. individual indicators of the same construct and sub-dimensions of the framework. Instead, when internal consistency is tested at an overall level among all individual indicators, we dispute the existence of the four cardinal points. Therefore, when reframing individual indicators into sub-dimensions/dimensions we are helping to optimize the internal consistency of DESI and lessening compensation issues due to aggregation (first *major objective*). To this aim, updates of the DESI 2022 framework are introduced and help in addressing the second *major objective* of the project that relates to whether to include more detailed individual indicators to capture a wider spectrum of the digital domain. Additionally, this chapter contains a comparison of the classical PCA (CPCA) method with its robust version (ROBPCA), to address the two *minor objectives* of the work: the robustness of CPCA as a latent factor detection method against outlying observations, and countries outlying diagnosis in a multivariate setting. For this purpose, the number of PCs to retain must be selected for both the classical and robust methods, to compute orthogonal distances (OD) and score distances (SD) of each observation and consequently flag outliers.

The following sections contain, for each of the four dimensions of DESI 2022, an internal consistency analysis of the single dimensions (see Section 4.2 for HC, Section 4.3 for CN, Section 4.4 for IDT, and Section 4.5 for

DPS) and sub-dimensions as defined in DESI 2022 (the *"status-quo"* scenario, depicted in Table 2.1). Specifically, CPCA is firstly applied over all the individual indicators contained in each dimension, to detect the number of possible latent dimensions that exist, that should correspond to hierarchical lower components, i.e. sub-dimensions. CPCA used for internal consistency purposes comes along with the ROBPCA method, to test for the robustness of the CPCA procedure and to detect possible outlying countries at the dimensional level. At the dimensional level ($p << n$), the ROBPCA will be directly using the MCD estimator. Following, CPCA is applied separately over each sub-dimension, to check whether a single construct underlies each of them, which is one of the requirements when constructing a CI. We expect high internal consistency within each sub-dimension. The same holds for internal consistency across sub-dimensions, measured using correlation when the number of sub-dimensions is just two, otherwise using Cronbach's alpha. In fact, the correlation coefficient is the most straightforward method for testing internal consistency in the case of the bidimensionality of components. Additionally, according to the internal consistency results, different scenarios suggesting adjustments to the "status quo" are proposed. Those adjustments have the aim of improving both the within and the across internal consistency of sub-dimensions, by re-organizing existing sub-dimensions (*framework adjustment*) and/or excluding/including indicators (*framework updating*).

After verifying the internal consistency of each dimension and the respective sub-dimensions, the last step (Section 4.6) considers applying CPCA over all indicators ($p = 33$) and all observations ($n = 27$), to check whether the four-dimensional structure of DESI holds. The robustness of the CPCA method towards outlying observations in a high-dimensional setting (p > n) is tested here by using ROBPCA as a comparative method.

4.2 Human Capital dimension: Internal consistency analysis

The HC dimension of DESI 2022 consists of two sub-dimensions: basic (1a) and advanced (1b) digital skills (see Table 2.1 for more information on how individual indicators are divided). We expect high internal consistency within each of the two sub-dimensions and across sub-dimensions, even if consistency across sub-dimensions is generally lower than within sub-dimensions.

For the "status quo" analysis, Cronbach's alpha coefficient for the HC dimension is equal to 0.84, which is over the set threshold value of 0.7, meaning that internal consistency among all indicators is acceptable. Instead, Cronbach's alpha within sub-dimensions is 0.97 and 0.59, respectively for sub-dimensions 1a and 1b, suggesting a low internal consistency of the latter.

The scree plots for the CPCA and the ROBPCA analysis of the "status quo" scenario over all indicators in HC dimension are shown in Figure 4.1(a) and Figure 4.1(b) respectively.

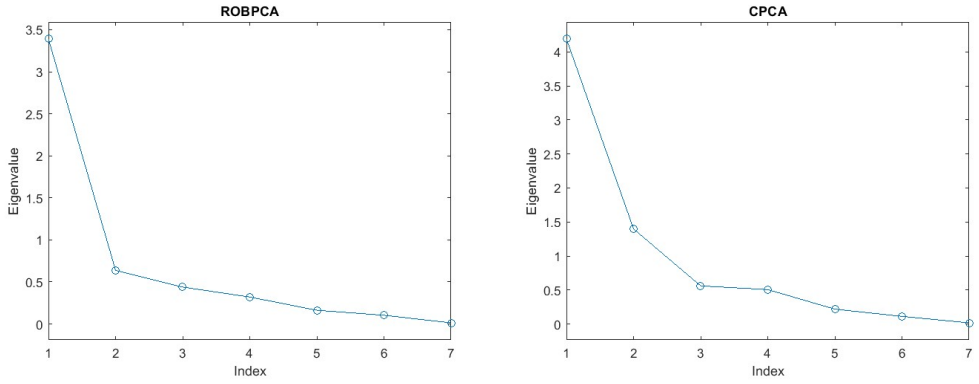


Figure 4.1: Scree plots showing the first 7 PCs with (a) ROBPCA; and (b) CPCA over HC indicators

It is clear that the two scree plots quite differ: while having only the first PC explaining more variance than the unit value threshold according to the

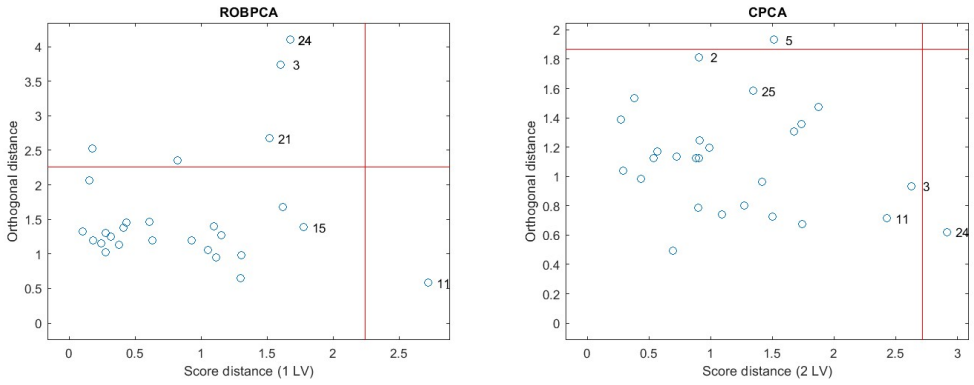


Figure 4.2: Outlier maps of the 27 countries obtained with (a) ROBPCA (k=1); and (b) CPCA (k=2)

Kaiser rule, for the ROBPCA method (Figure 4.1(a)), looking at the CPCA scree plot (Figure 4.1(b)) the Kaiser rule would consider both the first and the second PC to be retained. What appears clear from the scree plots is that apparently there are some observations shifting the estimated subspace away from the one-dimensional subspace which fits the true data structure. It might be ideal then to look at the outlier maps of both the classical and robust methods to see if this is the case. Fixing the number of retained PCs k to 1 for the ROBPCA and to 2 for the classical PCA according to the Kaiser rule, the outliers maps are shown in Figure 4.2(a) and Figure 4.2(b).

The CPCA method (Figure 4.2(b)) flags outlying observations corresponding to the Czech Republic (5) and Romania (24), respectively as orthogonal outliers and good leverage points. Instead, according to the ROBPCA diagnostic plot (Figure 4.2(a)), 4 outliers with a substantially increased score or orthogonal distances are detected. Observations 3, 21, and 24, corresponding to Bulgaria, The Netherlands, and Romania, are flagged as orthogonal outliers, while observation 11 (Finland) represents a good leverage point. Additionally, Romania (observation 24) is flagged as an outlier by both robust and classical PCA, indicating that the data point exhibits extreme behavior consistently across the two methodologies. This suggests that Romania

point represents an anomaly within the HC dimension. No bad leverage point is detected either by using CPCA or ROBPCA, and the two outlier maps slightly differ. As a result, the classical approach can be retained for conducting internal consistency analysis while being not overly influenced by outliers. Nevertheless, ROBPCA suggests that uni-dimensionality is a more representative choice for the HC structure.

Indicators	PC1	PC2	PC3
1a1	0.47	-0.09	-0.10
1a2	0.46	-0.03	-0.27
1a3	0.45	-0.12	-0.06
1b1	0.42	0.12	0.16
1b2	-0.03	0.73	-0.66
1b3	0.40	-0.11	-0.03
1b4	0.18	0.65	0.68

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	4.19	59.88
2	1.4	19.98
3	0.56	7.98
4	0.51	7.22
5	0.22	3.12
6	0.11	1.60
7	0.02	0.24

(b) Variances for each component

Table 4.1: Results of CPCA on HC dimension

The CPCA scree plot in Figure 4.1(b) clearly shows the existence of one main latent sub-dimension of Human Capital, holding almost 60% of the total explained variance, followed by a minor sub-dimension that explains approximately 20% (colored in yellow). This is in line with the current framework of the HC dimension of DESI 2022, where two sub-dimensions exist.

Further analyses examine the loading values in Table 4.1(a) that show how the individual indicators in the HC dimension contribute to the main underlying latent variables. The values that cross the 0.4 threshold are colored in green and signify a high contribution of the individual indicator to the first PC. Looking at PC1 column in Table 4.1(a), we notice that all indicators have a positive loading except for 1b2 "Female ICT specialists", which does not contribute to the first and most important PC. Additionally, indicator 1b4 "ICT graduates" slightly contributes to the first PC. It seems that one main

latent component exists and is supported by individual indicators within the HC dimension. We then decided to test a possible removal of the indicator "Female ICT specialists" in a successive adjustment of the Human Capital dimension (Subsection 4.2.2). Moreover, still looking at PC1 in Table 4.1(a), we notice that indicator 1b3 "Enterprises providing ICT training" has a loading of 0.4, which is closer to the values of indicators in sub-dimension 1a. Thus, it appears to be more correlated with those indicators than with the ones in sub-dimension 1b. We then tested indicator 1b3 "Enterprises providing ICT training" in the sub-dimension 1a.

4.2.1 Internal consistency analysis over sub-dimensions of HC

In the second part of the "status quo" scenario analysis, we apply CPCA to the two groups of indicators separately.

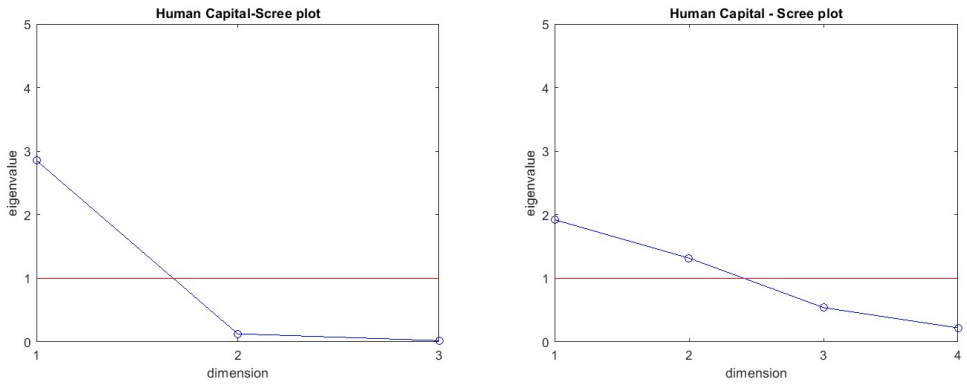


Figure 4.3: Scree plots for (a) sub-dimension 1a; and (b) sub-dimension 1b

The results for sub-dimension 1a are optimal both for the loadings and the variances, as shown in Table 4.2(a) and 4.2(b). In fact, the loadings for the first component are all positive and high while the first component explains more than 95% of the total variance. Cronbach's alpha for sub-dimension 1a is 0.97, well above the 0.7 threshold.

On the other hand, results for sub-dimension 1b are not so satisfactory.

Indicators	PC1	PC2	PC3
1a1	0.59	-0.04	0.81
1a2	0.57	-0.68	-0.45
1a3	0.57	-0.73	-0.38

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	2.85	95.08
2	0.13	4.24
3	0.02	0.68

(b) Variances for each component

Table 4.2: Results of CPCA on sub-dimension 1a of HC

Indicators	PC1	PC2	PC3
1b1	0.67	-0.15	-0.04
1b2	0.13	0.75	-0.65
1b3	0.59	-0.38	-0.33
1b4	0.44	0.52	0.69

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	1.92	48.06
2	1.32	32.94
3	0.54	13.49
4	0.22	5.51

(b) Variances for each component

Table 4.3: Results of CPCA on sub-dimension 1b of HC

The first PC accounts for only 48% of the total variance with the second one contributing to almost 33% of it. The first PC loading of indicator 1b2 "Female ICT specialist" is again the lowest one (0.13) (Table 4.3(a)). The analysis, therefore, suggests a low internal consistency of sub-dimension 1b that is attributable to indicator 1b2. In line with the CPCA results, Cronbach‘s alpha is 0.59, below the 0.7 threshold.32

Figure 4.4 relates the values of "ICT specialists" (1b1) with "Female ICT specialists" (1b2). By isolating this particular relationship, we can effectively highlight and investigate the gender dynamics within the ICT workforce and study the correlation between the two individual indicators. The cloud of points, each representing one of the 27 member states, has no shape that defines any correlation between the two, according to their correlation value of 0.03 in Table A.3, Appendix A. Due to all considerations made, the indi-

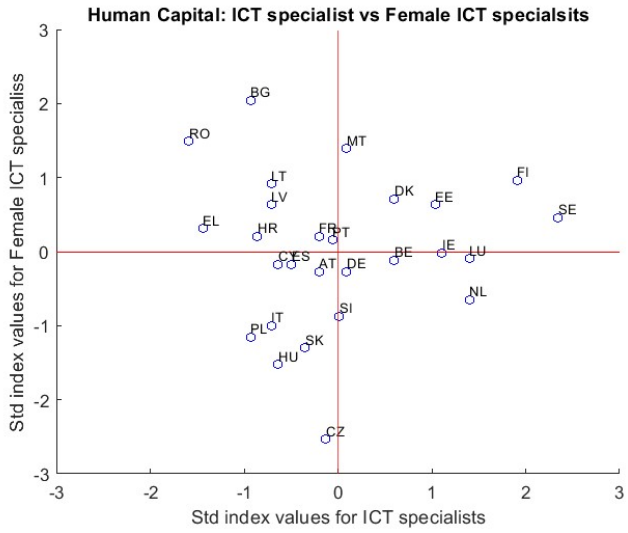


Figure 4.4: Plot of z score of "ICT specialists" against z score of "Female ICT specialist" indicators.

cator of "Female ICT specialists" (1b3) shows inconsistency within both the 1b sub-dimension and HC dimension, thus removed.

As a final step of the "status quo" analysis, correlation is measured across the two sub-dimensions, holding a value of 0.63, which is a significant correlation value and denotes the existence of across-dimension internal consistency.

Thus, the proposed adjustment attempts to raise the internal consistency of sub-dimension 1b, while preserving a high correlation degree among sub-dimensions.

4.2.2 Proposed adjustment for HC

The second step of the analysis investigates the possibility of adding two new indicators (source Eurostat), after the removal of the indicator "Female ICT specialists" (1b3):

- *Internet use* (Eurostat code: ISOC_CI_IFP_IU; year of reference 2021) which we will refer to as indicator "Never used internet", which has a negative direction, i.e. where lower is better;

- *Frequency of internet access*: at least once a week, including every day (Eurostat code: ISOC_CI_IFP_FU/I_USE; year of reference 2021), called "Frequency of use" indicator, with a positive direction, accordingly with the indicators already included in HC dimension.

CPCA is applied to all the indicators of the "status quo" scenario but "Female ICT specialists" (1b2) plus the two new indicators to be tested. We expect the indicator "Never used internet" to have loadings with an opposite sign, given its opposite direction. In this case, the opposite sign is an indication of a positive multivariate correlation.

Indicators	PC1	PC2	PC3
1a1	0.40	-0.26	0.30
1a2	0.38	-0.26	0.31
1a3	0.38	-0.26	0.28
1b1	0.38	0.12	-0.12
1b3	0.34	-0.20	-0.22
1b4	0.17	0.78	0.55
Never used internet	-0.36	0.28	-0.48
Frequency of use	0.38	0.23	-0.38

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	5.49	68.62
2	1.02	12.81
3	0.66	8.25
4	0.44	5.46
5	0.21	2.59
6	0.12	1.55
7	0.04	0.49
8	0.02	0.23

(b) Variances for each component

Table 4.4: Results of CPCA on revised HC dimension

Table 4.4(b) shows the existence of one PC explaining almost 69% of the variance. Looking at Table 4.4(a) instead, the new two indicators - "Never used internet" and "Frequency of internet use" - both contribute to this component to an extent that is similar to the indicators in sub-dimension 1a. In this case, the value of Cronbach's alpha coefficient is equal to 0.93, which is over the threshold value of 0.7, indicating a good internal overall consistency.

While the uni-dimensionality of HC dimension is suggested both here and from the ROBPCA analysis (Section 4.2), the proposal is to revise the two existing sub-dimensions. This choice allows for alignment with organizational objectives, i.e. having the two sub-dimensions of basic and advanced ICT skills divided. Moreover, the proposal of dividing into two sub-dimensions is less conservative while allowing for separated countries' performance analysis within the two sub-dimensions. The proposal is therefore to include in the revised sub-dimensions:

- **1a new**: 1a1 "At least basic digital skills", 1a2 "Above basic digital skills", 1a3 "At least basic digital content creation skills", 1b3 "Enterprises providing ICT trainings", "Never used internet" and "Frequency of internet use".

- **1b new**: 1b1 "ICT specialists", 1b4 "ICT graduates". In fact, based on Table A.3 in Appendix A, 1b1 is the individual indicator exhibiting the highest correlation value (0.41) with 1b4, the less fitting individual indicator in PC1 of Table 4.4(a), suggesting the belonging to another latent construct.

Finally, CPCA is applied to the revised sub-dimension 1a new. Table 4.5(b) shows that the percentage of variance explained by the first component is 77%, which is a clear indication of the existence of a unique latent factor for the sub-dimension. Looking at the loadings of component 1 (Table 4.5(a)), all the indicators contribute in a balanced way to this first, major component. Accordingly, Cronbach's alpha value is 0.94.

Correlation across the two new sub-dimensions is measured too and is equal to 0.73, thus confirming internal consistency across sub-dimensions.

In Appendix B, Table B.2 summarizes the proposed revision concerning the postulated sub-dimensions for Human Capital while following, Table 4.6 contains the internal consistency within and across sub-dimensions measures.

Indicators	PC1	PC2	PC3
1a1	0.44	-0.35	-0.15
1a2	0.42	-0.35	-0.19
1a3	0.42	-0.31	-0.23
1b3	0.37	-0.04	0.92
Never used internet	-0.38	0.62	-0.18
Frequency of use	0.41	0.52	-0.07

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	4.62	77.04
2	0.76	12.73
3	0.42	7.06
4	0.13	2.12
5	0.05	0.75
6	0.02	0.31

(b) Variances for each component

Table 4.5: Results of CPCA on revised sub-dimension *1a new* of HC

		CPCA: First component explained variance	Cronbach's alpha within sub-dim.	Correlation among dimensions
Status quo	1a	95%	0.97	0.63
	1b	48%	0.59	
Proposed adjustment	1a new	77%	0.94	0.73
	1b new	70%	0.58	

Table 4.6: Summarized measures of internal consistency for HC

The summarized outcomes displayed in Table 4.6 demonstrate notable patterns. In the context of *within internal consistency*, the first sub-dimension experiences a decrease in the percentage of variance explained by the initial PC from 96% to 77%. This decrement, however, maintains a satisfactory level. Conversely, the second sub-dimension exhibits a significant increase from 48% to 70%. Meanwhile, the reliability of Cronbach's alpha remains relatively consistent between the two. Notably, the internal consistency between

sub-dimensions, as indicated by the Pearson correlation coefficient, grows from 0.63 to 0.73. Consequently, the *proposed adjustment* results to be a less conservative choice compared to the result-supported uni-dimensionality of HC dimension. However, while leading to a reduction in internal consistency within the first sub-dimension of basic digital skills (but still satisfactory) it concurrently yields a distinct enhancement in internal consistency within the second sub-dimension and across sub-dimensions. Thus, overall the adjustment avoids compensability issues.

4.3 Connectivity dimension: Internal consistency analysis

The CN dimension of DESI 2022 consists of four sub-dimensions: Fixed broadband take-up (2a), Fixed broadband coverage (2b), Mobile broadband take-up and coverage (2c), and Broadband prices (2d) (see Table 2.1 for more information on the individual indicators). Analyzing the "status quo" of CN dimension, Cronbach's alpha coefficient among all indicators is equal to 0.4, well below the threshold value of 0.7, showing the poor overall internal consistency within the dimension. Also, Cronbach's alpha within sub-dimensions reflects poor internal consistency results, with values of 0.36 within 2a sub-dimension, 0.5 within 2b, and 0.46 within 2c. Results of internal consistency of sub-dimension Broadband prices (2d) are not reported since the sub-dimension is made of one indicator only, i.e. "Broadband price index" (2d1).

The scree plots for the CPCA and the ROBPCA analysis of the "status quo" scenario over all indicators in the CN dimension are shown in Figure 4.5(a) and Figure 4.5(b) respectively.

It is clear that the two scree plots slightly differ: in fact, looking at both scree plots (Figure 4.5(a) and Figure 4.5(b)), the first two PCs explain substantially more variance than the unit value threshold according to the Kaiser rule. Instead, the third PC is more explicative in the case of the classical pro-

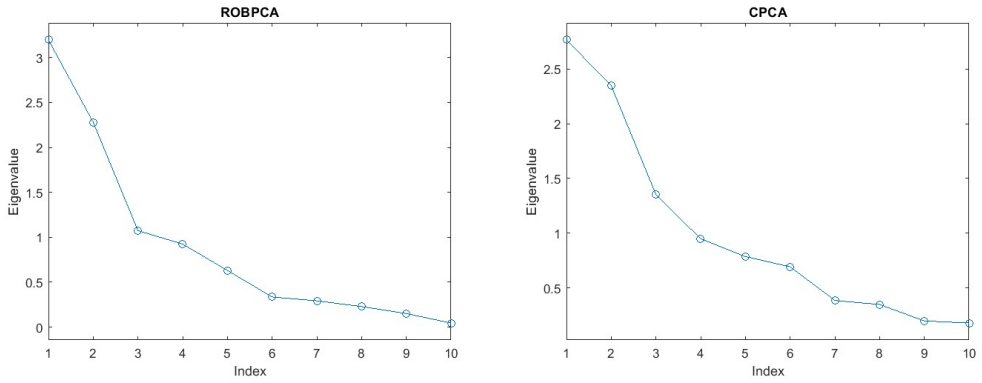


Figure 4.5: Scree plots showing the first 10 PCs with (a) ROBPCA; and (b) CPCA over CN indicators

cedure than for the ROBPCA one, where it narrowly crosses the 1 threshold. This is an indication of the existence of outliers in CN dimension that inflates the third eigenvalue estimate of CPCA. The next step consists of looking at the outlier maps of both the classical and robust methods if outlying countries are detected. Fixing the number of retained PCs k to 3 for both CPCA and ROBPCA, according to the Kaiser rule, the outliers maps are following shown.

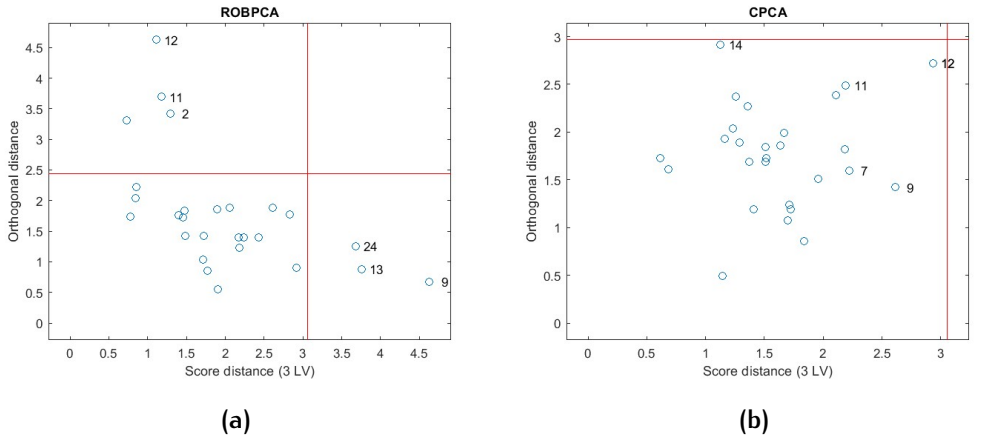


Figure 4.6: Outlier maps of the 27 countries obtained with (a) ROBPCA (k=3); and (b) CPCA (k=3)

The CPCA method (Figure 4.6(b)) does not flag any observations as outlying while according to the ROBPCA diagnostic plot (Figure 4.6(a)), 6 outliers with an increased score or orthogonal distances are detected. Observations 2, 11, and 12, corresponding respectively to Belgium, Finland, and France are good leverage points while observations 9, 13, and 24, corresponding to Greece, Croatia, and Romania, are detected as orthogonal outliers. No bad leverage points are detected, however, CPCA results are biased since the outliers map differs significantly between ROBPCA and classical PCA, which might indicate that the presence of outliers has a strong influence on the results of the analysis. Additionally, the outlying orthogonal observations detected with ROBPCA are a sign of underlying latent factors not captured by the retained PCs. Anyhow following considerations on the internal consistency of the dimension/sub-dimension will be based on the classical PCA procedure, for consistency with the methods used within the work.

The results of the CPCA analysis on all indicators are shown in the following tables. In Table 4.7(a), indicators that are positively related and mostly contribute to the first PCs are highlighted in green. The others are either counter-related or do not contribute to the first component but to the second one (orange-coloured).

The CPCA scree plot in Figure 4.5(a) shows the existence of multiple sub-dimensions in Connectivity. Table 4.7(b) and Figure 4.5(a), show that the first three principal components account for a significant portion of the total variance, crossing the unit value set by the Kaiser rule. This deviates from the current framework of the CN dimension of DESI 2022, where four sub-dimensions exist. Additionally, apart from some exceptions, the distribution of the indicators in the different sub-dimensions, as defined in DESI 2022 framework, does not correspond to what is suggested by the CPCA (Table 4.7(a)). Besides, the two 5G indicators ("5G spectrum" (2c1) and "5G coverage" (2c2)) describe a sub-dimension that is anti-related to the rest (as can be deducted by their loadings sign that is opposite to the others). In fact, a negative loading value indicates a negative correlation between the

Indicators	PC1	PC2	PC3
2a1	-0.01	0.56	-0.02
2a2	0.49	0.21	0.12
2a3	0.11	-0.05	0.53
2b1	0.04	0.44	-0.37
2b2	0.53	0.12	0.04
2b3	0.48	-0.16	0.10
2c1	-0.26	0.11	0.43
2c2	-0.33	0.12	0.44
2c3	0.15	0.39	0.39
2d1	0.19	-0.48	0.14

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	2.77	27.66
2	2.35	23.48
3	1.35	13.52
4	0.95	9.48
5	0.79	7.86
6	0.69	6.92
7	0.38	3.84
8	0.35	3.48
9	0.2	1.97
10	0.18	1.78

(b) Variances for each component

Table 4.7: Results of CPCA on CN dimension

variable and the principal component, which in this case captures the main latent variable.

4.3.1 Internal consistency over sub-dimensions of CN

In the second part of the "status quo" scenario analysis, we apply the CPCA to the sub-dimensions separately. In all the cases the internal consistency is not optimal. By looking at the variance tables for each sub-dimension, Table 4.8(b), Table 4.9(b) and Table 4.10(b), we can see that none of the sub-dimensions is described by a unique latent construct, which is also confirmed by looking at the scree plots in 4.7. In fact, all the eigenvalues for sub-dimensions 2a, 2b, and 2c are very close to the unit Kaiser threshold, meaning that the 3 PCs contribute in a balanced way to each sub-dimensional construct when unidimensionality is instead expected. As previously mentioned, the CPCA results are confirmed by Cronbach's alpha value, which is respectively 0.36, 0.5, and 0.46, all below the 0.7 threshold.

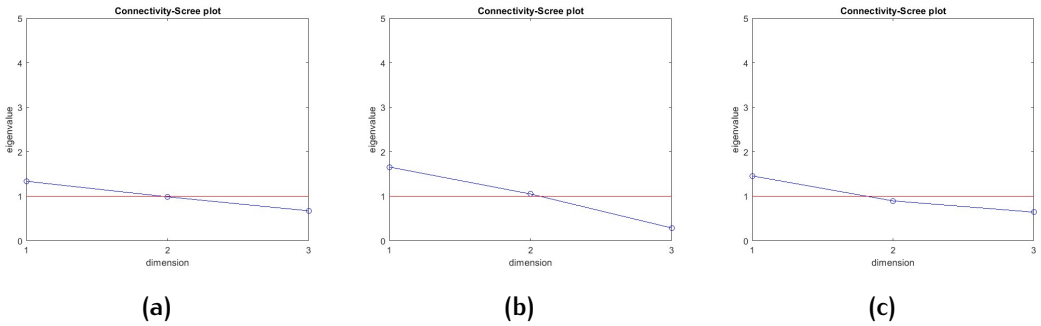


Figure 4.7: Scree plots for (a) sub-dimension 2a; (b) sub-dimension 2b; and (c) sub-dimension 2c (the red lines correspond to the threshold value 1 set by the Kaiser rule)

The Cronbach's alpha across the four sub-dimensions is -0.33. This is an indication of anti-relation between the four sub-dimensions of CN, accoridngly to the loading values of Table 4.7(a). By definition, Cronbach's alpha assumes negative values whenever the average correlation among the indicators is negative (Ursachi et al. 2015). This result leads to compensability issues among sub-dimensions of CN.

Given the low internal consistency showed *within* and *across* sub-dimensions, the proposed adjustments of CN dimension aim at:

- Reframing sub-dimensions into more consistent constructs that better reflect the latent dimension within CN;
- Lessen the compensability effect among sub-dimensions.

Additionally, based on the CPCA results and on what was suggested by experts in DG CONNECT, who have been constantly consulted on the analysis results and possible ways forward, the following indicators are excluded from the analysis:

- 2a1 "Overall fixed broadband take-up": not fitting the CPCA and conceptually overlapping with the other indicators in sub-dimension 2a (2a2 "At least 100 Mbps fixed broadband take-up");

Indicators	PC1	PC2	PC3
2a1	0.47	0.75	-0.46
2a2	0.70	-0.004	0.71
2a3	0.53	-0.66	-0.53

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	1.34	44.58
2	0.99	32.96
3	0.67	22.46

(b) Variances for each component

Table 4.8: Results of CPCA on sub-dimension 2a of CN

Indicators	PC1	PC2	PC3
2b1	0.10	0.96	-0.28
2b2	0.72	0.13	0.69
2b3	0.69	-0.27	-0.67

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	1.66	55.27
2	1.03	35.09
3	0.29	9.63

(b) Variances for each component

Table 4.9: Results of CPCA on sub-dimension 2b of CN

Indicators	PC1	PC2	PC3
2c1	0.59	-0.52	-0.61
2c2	0.66	-0.13	0.74
2c3	0.46	-0.84	-0.27

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	1.46	48.61
2	0.90	29.92
3	0.64	21.47

(b) Variances for each component

Table 4.10: Results of CPCA on sub-dimension 2c of CN

- 2a3 "At least 1 Gbps take-up": the distribution of this indicator across the 27 countries is highly negatively skewed (due to the presence of two outliers Hungary and France with values equal to 0.22 and 0.27 respectively). This is typical behaviour for an indicator of this type, describing the taking-up of an advanced type of fixed broadband. This indicator is put on-hold and will be reconsidered once new data becomes

available.

- 2b1 "Fast broadband (NGA) coverage": not fitting the CPCA and it is a too basic indicator according to CONNECT experts.

4.3.2 Proposed adjustment for CN

The second step of the analysis investigates the possibility of adding a new indicator:

- *5G stations in the 3.6 band*: number of 5G base stations in the 3.4/3.8 GHz band, per 1000 inhabitants (COCOM, year of reference 2022). This indicator is added to describe the quality of the 5G services, as it focuses on the 3.6 GHs band. We will refer to this new indicator as "5G stations".

The new indicator presents 4 missing values, corresponding to Estonia, Finland, Luxemburg, and Sweden. This data characteristic required an adjustment of the indicator weights for these countries (for the weights to always sum to 1, see Section 5.3).

Also, the latest version of data has been provided and used for indicators:

- 2a2 "At least 100 Mbps fixed broadband take-up" (updated data to be reassessed in 2023);

- 2d1 "Broadband price index" (updated data sent on 09/11/2022).

CPCA results on this revised set of indicators are displayed in Table 4.11(a) and Table 4.11(b). These results will help in deriving an understanding for re-framing the CN dimension into new and optimal sub-dimensions.

Table 4.11(b) suggests the presence of three latent sub-dimensions according to the Kaiser rule. Anyway, the third dimension explains a very small percentage of total variance, approximately 13%, thus we decide to retain only the first and second PCs that respectively explain 38% and 21% of the total variance. This hypothesis is supported by the loading values too.

Indicators	PC1	PC2	PC3
2a2 (*new data*)	0.49	0.00	-0.16
2b2	0.53	0.20	-0.18
2b3	0.45	0.07	0.15
2c1	-0.29	0.38	0.06
2c2	-0.22	0.37	0.59
2c3	0.16	0.55	-0.15
2d1 (*new data*)	0.37	-0.13	0.74
5G stations	0.03	0.60	-0.06

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	3.11	37.90
2	1.71	20.83
3	1.12	13.62
4	0.97	11.83
5	0.50	6.13
6	0.31	3.78
7	0.26	3.21
8	0.22	2.7

(b) Variances for each component

Table 4.11: Results of CPCA on new CN dimension

In fact, by looking at Table 4.11(a), the PC1 column suggests the existence of two opposite groups of indicators:

- **Group 1** (*Fixed and broadband price sub-dimension*): indicators with a positive and high value of loadings of the first PC (green coloured), including "At least 100 Mbps fixed broadband take-up (new data)" (2a2), "Fixed Very High Capacity Network (VHCN) coverage" (2b2), "Fibre to the Premises (FTTP) coverage" (2b3) and "Broadband price index (new data)" (2d1);
- **Group 2** (*Mobile sub-dimension*): with a negative or low value for the first component of the CPCA (in Table 4.11(a), orange coloured) and a high value on the second component of the CPCA (in Table 4.11(a), purple coloured), including indicators "5G spectrum" (2c1), "5G coverage" (2c2), "Mobile broadband take-up" (2c3) and "5G stations".

The next step is to apply CPCA separately to the two new revised sub-dimensions. For the "Fixed" sub-dimension, we both consider the case with

the Broadband price index excluded and then re-included, to assess its impact on the internal consistency of the "Fixed" sub-dimension.

Fixed sub-dimension without Broadband price index (*First proposal*)

Here CPCA is applied to the "Fixed" sub-dimension with the "Broadband price index" *excluded*. In this case, 75% of the percentage of variance is accounted for by the first component (Table 4.12(b)). All the indicators contribute in a balanced way to this first, major component (Table 4.12(a)). Accordingly, Cronbach's alpha value is 0.8, above the 0.7 threshold. These results clearly indicate the existence of a unique latent factor.

Indicators	PC1	PC2	PC3
2a2 (*new data*)	0.57	-0.65	-0.50
2b2	0.60	-0.08	0.79
2b3	0.56	0.76	-0.34

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	2.25	75.11
2	0.47	15.51
3	0.28	9.38

(b) Variances for each component

Table 4.12: Results of CPCA on new "Fixed" sub-dimension 2c without broadband price index

Fixed sub-dimension with Broadband price index included (*Second proposal*)

Results slightly deteriorate for the analysis of the "Fixed" sub-dimension with the "Broadband price index" *included*, even if they remain satisfactory. The percentage of variance accounted by the first component is now 64% (Table 4.13(b)) and all the indicators contribute at the same level and in the same direction to this first, major component (Table 4.13(a)). It is of interest to observe that the indicator "Broadband price index" (2d1), even if

fitting well with the others in the first component, contributes to the largest part of the second principal component (purple colored in Table 4.13(a)). This suggests that, as expected, the price indicator is less consistent with the other indicators in this sub-dimension. The Cronbach's alpha value is in this case 0.81, still well above the 0.7 threshold.

Indicators	PC1	PC2	PC3
2a2 (*new data*)	0.52	-0.30	-0.64
2b2	0.54	-0.38	0.04
2b3	0.50	0.01	0.74
2d1 (*new data*)	0.40	0.88	-0.21

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	2.55	63.88
2	0.73	18.17
3	0.45	1.26
4	0.27	6.69

(b) Variances for each component

Table 4.13: Results of CPCA on new "Fixed" sub-dimension with broadband price index

Mobile sub-dimension

Table 4.14(b) shows that the percentage of variance explained by the first component is 43%, while the second and the third component respectively account for 25% and 21%. All the loadings of the first principal component have comparable values and the same sign (Table 4.14(a)). Cronbach's alpha value is 0.55, below the 0.7 threshold, as expected given the relatively low share of variance accounted by the first component. The "Mobile" sub-dimension is less performant than the "Fixed" ones but, considering the nature of the indicators here, internal consistency can be still considered satisfactory. In fact, when individual indicators come from different surveys, as in the case of indicators 2c1, 2c2, 2c3, and "5G stations" (see Figure B.2, Appendix B), a lower internal consistency is expected but can be attributed to diverse popu-

lations, scale heterogeneity, data quality, etc., not further investigated in this work.

Indicators	PC1	PC2	PC3
2c1	0.48	-0.63	-0.14
2c2	0.51	-0.19	0.72
2c3	0.46	0.75	0.15
5G stations	0.55	0.09	-0.66

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	1.69	42.70
2	0.99	25.00
3	0.83	21.00
4	0.45	11.30

(b) Variances for each component

Table 4.14: Results of CPCA on new "Mobile" sub-dimension

To determine which of the revised "Fixed" sub-dimension to retain, we regard the *across* sub-dimension internal consistency, since the internal consistency *within* is satisfied for each proposed "Fixed" sub-dimension and can be considered equal. In fact, it is preferred a "Fixed" sub-dimension having a higher correlation with the "Mobile" sub-dimension. Thus, correlations between the scores of the "Fixed" sub-dimension without price index, the "Mobile" sub-dimension, and the "Broadband price index" are compared in Table 4.15. The correlation between "Fixed" scores, "Mobile" scores without price, and "Broadband price index", shows that there is a positive correlation between the "Broadband price index" and the "Fixed" sub-dimension (0.44), while the correlation is negative but close to 0 with the "Mobile" sub-dimension (-0.05) (Table 4.15). Thus, it is proved that the "Broadband price index" is consistent within the "Fixed" sub-dimension, while not affecting much the across sub-dimensional internal consistency. Also, the correlation between "Fixed" with price index included and "Mobile" is lower (-0.29) than of "Fixed" without price index and "Mobile" (-0.31). None of the two proposals solves for the compensability issue between sub-dimensions, but the second proposal mitigates it more than the first. Those are additional justifications, together with the CPCA results, for the second proposal where the

"Broadband price index" is included in the "Fixed" sub-dimension.

	Indicators		
Indicators	Fixed	Mobile	Broadband price index
Fixed	1.00	-0.31	0.44
Mobile		1.00	-0.05
Broadband price index			1.00

Table 4.15: Correlation between sub-dimension "Fixed" and "Mobile" with "Broadband price index"

		CPCA: First component explained variance	Cronbach's alpha within sub-dimension	Cronbach's alpha across sub-dim.	Correlation among sub-dim.
Status quo	2a	45%	0.36	-0.33	–
	2b	55 %	0.5		
	2c	49 %	0.46		
	2d	–	–		
First proposal	Fixed	75%	0.83	-0.91	-0.31
	Mobile	43%	0.55		
Second proposal	Fixed	64%	0.81	-0.81	-0.29

Table 4.16: Summarized measures of internal consistency for CN

In Appendix B, Table B.2 summarizes the proposed revision concerning the two proposals for Connectivity.

Results of internal consistency, both for the "status quo" and the two proposals, are summarized in Table 4.16. In this case, while internal consistency is improved *within* sub-dimensions, internal consistency *across* does not hold either for sub-dimensions in the first or second proposal, thus the compensability problem remains unsolved. In fact, in both cases, the two sub-dimensions are negatively correlated.

4.4 Integration of Digital Technology dimension: Internal consistency analysis

The IDT dimension of DESI 2022 consists of three sub-dimensions: Digital intensity (3a), Digital technologies for businesses (3b), and e-Commerce (3c). For the “status-quo” assessment, the CPCA is applied firstly over all the eleven indicators of the IDT dimension. Then separate analyses are applied to sub-dimensions 3b and 3c, to check whether a single construct underlies each of them. Sub-dimension 3a is excluded from any internal consistency analysis since it includes just one indicator.

Starting from the reliability measures, Cronbach’s alpha coefficient is equal to 0.9, well above the 0.7 threshold, indicating an optimal level of internal consistency across all indicators of IDT. Cronbach’s alpha values are also high, 0.79 and 0.87 within sub-dimensions 3a and 3b respectively. Finally, Cronbach’s alpha across the three original sub-dimensions has a value of 0.86, which is a clear indication of across-dimension internal consistency. Thus, based on Cronbach’s alpha results only we acquire some indication of overall good internal consistency within/among components of IDT dimensions. PCA results will instead help detect the number of latent dimensions existing and spotting outlying individual indicators.

The scree plots for the CPCA and the ROBPCA analysis of the "status quo" scenario over all indicators in IDT dimension are shown.

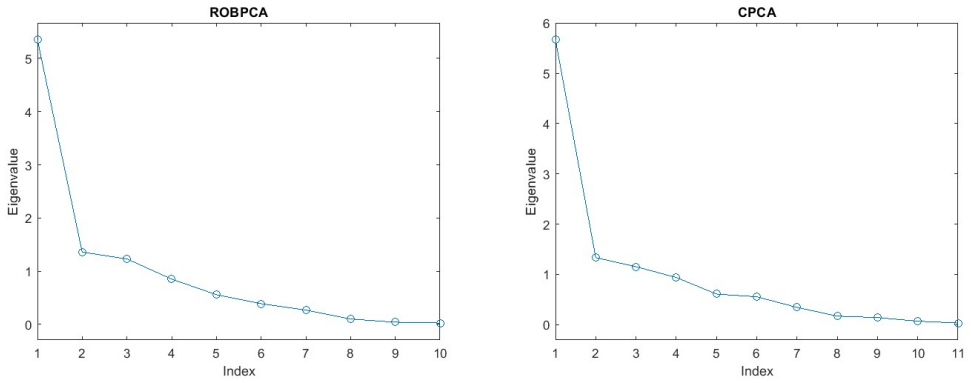


Figure 4.8: Scree plots showing the first 10 PCs with (a) ROBPCA; and (b) CPCA over IDT indicators

Looking at the two scree plots in Figure 4.8(a) and Figure 4.8(b), we notice that they do not differ much: in fact, for both, the first PC explains substantially all the variance while the second and the third slightly cross the unit value threshold according to the Kaiser rule. We double-check the robustness of the CPCA method by looking at the outliers map of both classical and robust, fixing $k = 3$.

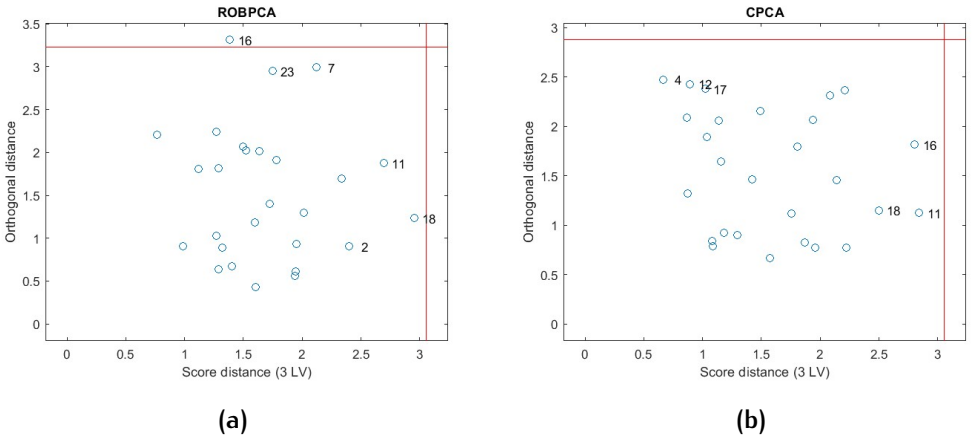


Figure 4.9: Outlier maps of the 27 countries obtained with (a) ROBPCA (k=3); and (b) CPCA (k=3)

The CPCA method (Figure 4.9(b)) does not flag any observation as out-

lying while the ROBPCA diagnostic plot (Figure 4.9(a)), detects 1 outlier, corresponding to observation 16 (Italy) with increased orthogonal distance. Thus no bad leverage point is detected in the case of the robust procedure and CPCA is considered robust enough toward outliers for the IDT dimension, given the similarity of the two outliers map.

The results of the CPCA analysis over all indicators are shown in the following Tables 4.17(a) and 4.17(b).

Indicators	PC1	PC2	PC3
3a1	0.40	0.08	0.12
3b1	0.26	0.38	-0.33
3b2	0.35	0.28	-0.02
3b3	0.32	0.04	-0.22
3b4	0.36	-0.09	0.35
3b5	0.32	0.32	-0.01
3b6	-0.01	0.55	0.21
3b7	0.20	-0.07	0.74
3c1	0.32	-0.34	-0.18
3c2	0.28	-0.47	-0.01
3c3	0.32	-0.14	-0.28

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	5.66	51.50
2	1.33	12.12
3	1.15	10.47
4	0.94	8.52
5	0.61	5.53
6	0.55	5.03
7	0.34	3.11
8	0.17	1.53
9	0.14	1.28
10	0.07	0.62
11	0.03	0.29

(b) Variances for each component

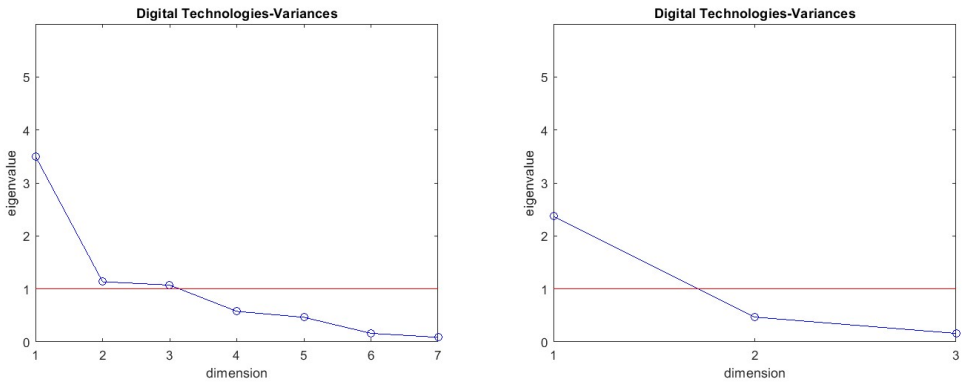
Table 4.17: Results of CPCA on IDT dimension

The analysis clearly indicates the existence of a single, main latent dimension in IDT, followed by two minor latent constructs. In fact, by looking at Table 4.17(b) we can see how the first principal component explains most of the total variance (52%), followed by the second (12%) and the third (10%). The positive and balanced loadings shown in the first column of Table 4.17(a) (green coloured) are consistent with a conceptual framework of a single dimension, well described by all the indicators but one, the "ICT for environmental sustainability" (3b6) indicator, which has a negative loading

value (orange coloured). This result is in contrast with the conceptual framework of DESI 2022, where three sub-dimensions within IDT are outlined.

4.4.1 Internal consistency over sub-dimensions of IDT

In the second part of the "status quo" scenario analysis, we apply CPCA on sub-dimensions 3b and 3c separately.



(a) **(b)**

Figure 4.10: Scree plots for (a) sub-dimension 3b; and (b) sub-dimension 3c

Indicators	PC1	PC2	PC3
3b1	0.37	0.42	-0.22
3b2	0.46	0.10	-0.04
3b3	0.42	0.02	-0.35
3b4	0.45	-0.37	0.14
3b5	0.44	0.23	0.08
3b6	0.01	0.57	0.73
3b7	0.27	-0.55	0.52

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	3.5	50.03
2	1.14	16.22
3	1.07	15.32
4	0.58	8.24
5	0.46	6.63
6	0.16	2.31
7	0.09	1.24

(b) Variances for each component

Table 4.18: Results of CPCA on sub-dimension 3c of IDT

Indicators	PC1	PC2	PC3
3c1	0.10	0.96	-0.28
3c2	0.72	0.13	0.69
3c3	0.69	-0.27	-0.67

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	1.66	55.27
2	1.03	35.09
3	0.29	9.63

(b) Variances for each component

Table 4.19: Results of CPCA on sub-dimension 2b of CN

The results for internal consistency inside the sub-dimensions are satisfactory in both cases. Each sub-dimension describes a unique latent construct (Table 4.18(b) and Table 4.19(b)) with all the indicators well behaving in their own sub-dimension apart from the indicator 3b6 "ICT for environmental sustainability" (Table 4.19(a)). Cronbach's alpha values are also high, respectively 0.79 and 0.87. Cronbach's alpha across the three original sub-dimensions has a value of 0.86, which is a clear indication of across-dimension internal consistency.

The proposed adjustment then considers taking out indicator 3b6 "ICT for environmental sustainability". Instead, we can combine all the other indicators into a new, cohesive dimension called the "revised IDT" dimension.

4.4.2 Proposed adjustment for IDT

Table 4.20(b) shows that the variance accounted by the first principal component increases to 57%, when indicator 3b6 is removed. The loadings from Table 4.20(a) are now all balanced and with concordant signs.

In this case, the value of Cronbach's alpha is of 0.89, still well above the 0.7 threshold.

Indicators	PC1	PC2	PC3
3a1	0.40	0.03	0.15
3b1	0.26	0.54	-0.08
3b2	0.35	0.29	0.14
3b3	0.32	0.20	-0.09
3b4	0.36	-0.2	0.32
3b5	0.32	0.27	0.10
3b6	0.20	0.37	0.66
3c1	0.32	-0.32	-0.41
3c2	0.28	-0.48	-0.29
3c3	0.32	-0.07	-0.39

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	5.66	56.64
2	1.22	12.20
3	1.12	11.23
4	0.62	6.23
5	0.56	5.62
6	0.35	3.49
7	0.20	1.96
8	0.15	1.53
9	0.08	0.77
10	0.03	0.33

(b) Variances for each component

Table 4.20: Results of CPCA on revised IDT dimension

The main results of the internal consistency assessment across and within sub-dimensions are shown in Table 4.21.

		CPCA: First component explained variance	Cronbach's alpha within sub-dimension	Cronbach's alpha across sub-dim.
Status quo	3a	–	–	0.86
	3b	50%	0.79	
	3c	79%	0.87	
Proposed adjustment	No sub-dim.	57%	0.89	–

Table 4.21: Summarized measures of internal consistency for IDT

From Table 4.21 we can see that the internal consistency for the proposed

"revised IDT" dimension, corresponding to the Proposed adjustment line, is high enough both in terms of variance explained by the first component (57%) and Cronbach's alpha value (0.89). Additionally, the simplified framework for IDT, where no sub-dimension is outlined, prevents from any possible trade-off effect between sub-dimensions.

4.5 Digital Public Services: Internal consistency analysis

The DPS dimension of DESI 2022 consists of one sub-dimension only, which is called e-Government. Denoted as sub-dimension 4a, it describes the demand and supply of e-government as well as open data policies.

For the "status-quo" assessment, the CPCA is applied over all the five indicators of the DPS dimension, to check whether a single construct underlies the e-Government sub-dimension. No internal consistency analysis across sub-dimension is performed since only one sub-dimension is defined at the "status quo" level. A first indication of good internal consistency within the DPS dimension is provided by Cronbach's alpha coefficient value, which is equal to 0.8, well above the 0.7 threshold.

Then, the scree plots for the CPCA and the ROBPCA analysis of the "status quo" scenario over all indicators in dimension are shown in Figure 4.11(b) and 4.11(a) respectively. This analysis will help in testing for latent variable detection within the DPS dimension and eventually notice outlying individual indicators.

Looking at the two scree plot in Figure 4.11(b) and Figure 4.11(a) we notice that they quite differ: while for the classical procedure (Table 4.11(b)) the first PC explains substantially all the variance and the second PC holds a lower value, looking at the robust scree plot (Table 4.11(a)), the first and second PCs have closer values while both crossing the unit value threshold set with the Kaiser rule. It seems like the CPCA scree plot is skewed towards a unique latent vector dimension. Thus, it is legit to double-check the ro-

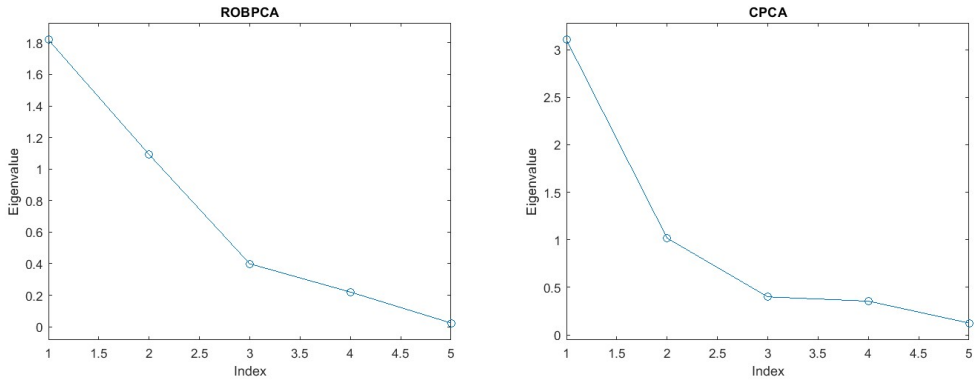


Figure 4.11: Scree plots showing the first 10 PCs with (a) ROBPCA; and (b) CPCA over DPS indicators

bustness of the CPCA method by looking at the outliers map and checking whether bad leverage points affect the estimates. K is fixed at 2 for both classical and robust procedures, according to the Kaiser rule the first two eigenvalues cross the unit value in both cases.

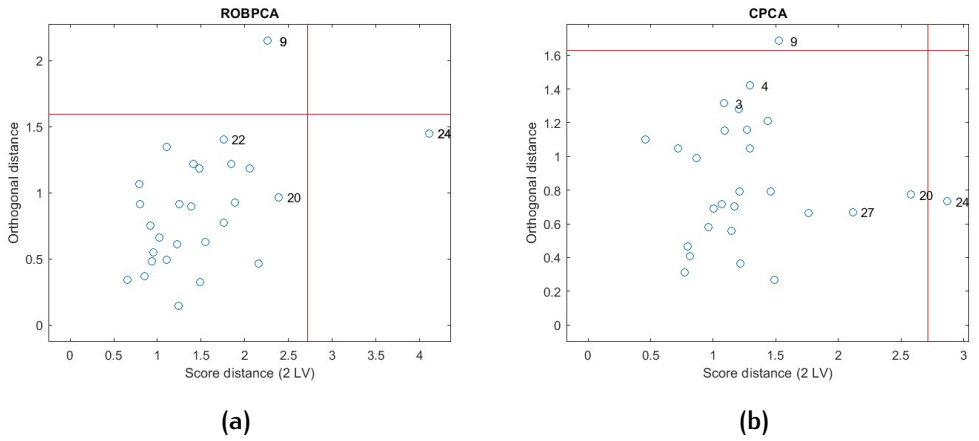


Figure 4.12: Outlier maps of the 27 countries obtained with (a) ROBPCA (k=2); and (b) CPCA (k=2)

The CPCA method (Figure 4.12(b)) flags one observation 9 (Greece) as an orthogonal outlier while observation 24 (Romania) is flagged as a good leverage point. The ROBPCA diagnostic plot (Figure 4.12(a)) detects the same

outlying points, with respectively increased orthogonal and score distances. This is an indication that Greece and Romania stand out as true anomalous points within the context of the DPS dimension. In any case, no bad leverage point is found and CPCA can be considered robust toward outliers since the outliers maps for the two methods are similar (Figure 4.12(b) and 4.12(a)). As a consequence, CPCA can be used to test for internal consistency of the dimension/sub-dimensions.

The results of the CPCA analysis on all indicators are shown in the following tables.

Indicators	PC1	PC2	PC3
4a1	0.47	0.10	0.87
4a2	0.49	-0.07	-0.151
4a3	0.526	-0.17	-0.26
4a4	0.51	0.02	-0.37
4a5	0.07	0.98	-0.13

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	3.10	62.02
2	1.02	20.34
3	0.40	8.00
4	0.36	7.09
5	0.13	2.49

(b) Variances for each component

Table 4.22: Results of CPCA on DPS dimension

The analysis clearly shows the existence of a single, main latent dimension in DPS, followed by a minor latent dimension, according to the Kaiser rule. In fact, by looking at Table 4.22(b) we can see how the first principal component explains most of the total variance, 62%, while the second accounts for the 20%. The positive and balanced loadings shown in the first column of Table 4.22(a) (green coloured) are consistent with a conceptual framework of two sub-dimensions, well described by all the indicators but one, the "Open data" (4a5) indicator. The latter holds an uninfluential loading value (orange coloured) for the first component, while it accounts for most of the second one (purple coloured), clearly indicating its contribution to a different latent factor than the one captured by the main PC. This is in contrast

with the DESI 2022 uni-dimensional structure of DPS, where all individual indicators contribute to one latent construct within the dimension. Instead of eliminating the "Open data" indicator *or* splitting the existing individual indicators into two sub-dimensions — one for "Open data" exclusively and the other encompassing all remaining individual indicators — the suggested modification involves expanding the structure of Digital Public Services. This expansion is achieved by breaking down the current individual indicators into their "National" and "Cross Border" counterparts.

4.5.1 Proposed adjustment for DPS

The second step of the analysis investigates the possibility of defining two sub-dimensions within the DPS dimension, one which captures "National" services and one "Cross-Border" (CB) services. In fact, under consultation with DESI experts, the DESI 2022 structure was considered to be a too simplistic representation of the DPS domain. This approach instead acknowledges that the DPS dimension can be multifaceted, encompassing various aspects of both national and cross-border service provision. By separating these sub-dimensions, we anticipate the potential to reveal insights into how countries perform and engage differently in each of these service types.

Indicators 4a3 "Digital public services for citizens" and 4a4 "Digital public services for businesses" are removed. In fact, in DESI 2022 both indicators are computed as the equal-weighted arithmetic mean of online availability and cross-border online availability, respectively for citizen life event (4a3) and for businesses life event (4a4). To better separate the DPS dimension into national and cross-border sub-dimensions, the score for online availability and the score for CB Online availability are separately introduced in the DPS dimension, for both citizen and businesses life event. The data source is the e-Government Benchmark, 2022 Reports

These four indicators replace original indicators 4a3 and 4a4. Following the same logic, new indicators coming from the same data source are added to the national and CB sub-dimensions:

- *Mobile friendliness*: the extent to which services are provided through a mobile-friendly interface (Score - 0 to 100);
- *User Support*: the extent to which online support, help features, and feedback mechanisms are available (Score - 0 to 100);
- *Transparency*: the extent to which services are designed with user involvement and users can manage their personal data (Score - 0 to 100);
- *CB User Support*: the extent to which online support, help features, and feedback mechanisms are available for users from other European countries (Score - 0 to 100).

The summary statistics for the newly introduced indicators of the proposed adjustment for DPS are contained in Appendix A, Tables A.16 and A.17.

The revised framework that we propose consists of two sub-dimensions, one containing only national-level indicators and the other with only cross-border indicators. Indicator 4a5 "Open data" is instead excluded as non-fitting. CPCA results on this revised set of indicators are displayed in Table 4.23(a) and Table 4.23(b).

The analysis shows the presence of one main latent dimension, with all the indicators (old and new) contributing approximately to the same extent and with the same orientation to the first principal component which explains almost 63% of the total variance (Table 4.23(b)). The positive and balanced loadings shown in the first column of Table 4.23(a) are consistent with the conceptual framework of a single dimension. This does not contrast with the possibility of two sub-dimensions, national and cross-border services, and it is in line with a conceptual framework that keeps the two concepts separated (see Section 5.5).

Finally, CPCA is applied to the newly defined National and CB services sub-dimensions individually.

Table 4.24(b) shows that the percentage of variance explained by the first component goes up to 65% when restricted to the National services

Indicators	PC1	PC2	PC3
4a1	0.30	-0.13	-0.40
4a2	0.34	0.21	-0.08
Public services for citizens (*national*)	0.34	0.28	-0.32
Public services for businesses (*national*)	0.31	-0.30	-0.38
Mobile friendliness	0.32	0.01	-0.14
User Support	0.20	0.72	0.26
Transparency	0.36	0.17	0.19
CB online availability citizen	0.36	-0.16	0.24
CB online availability businesses	0.34	-0.32	-0.15
CB User Support	0.26	-0.30	0.62

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	6.22	62.16
2	1.22	12.21
3	0.90	9.00
4	0.45	4.51
5	0.38	3.84
6	0.28	2.80
7	0.24	2.40
8	0.17	1.67
9	0.09	0.86
10	0.06	0.64

(b) Variances for each component

Table 4.23: Results of CPCA on revised DPS dimension

sub-dimension, which is also a clear indication of the existence of a unique latent factor for the sub-dimension. Looking at the loadings of PC1 (Table 4.24(a)), all the indicators contribute in a balanced way to this first, major component. Accordingly, Cronbach's alpha value is 0.91.

Indicators	PC1	PC2	PC3
4a1	0.36	-0.35	-0.32
4a2	0.42	0.11	0.15
Public services for citizens (*national*)	0.43	0.05	-0.33
Public services for businesses (*national*)	0.35	-0.50	-0.11
Mobile friendliness	0.38	-0.10	0.85
User Support	0.27	0.76	-0.06
Transparency	0.42	0.18	-0.17

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	4.54	64.83
2	1.08	15.41
3	0.44	6.25
4	0.37	5.29
5	0.26	3.69
6	0.20	2.90
7	0.12	1.64

(b) Variances for each component

Table 4.24: Results of CPCA on National services sub-dimension

Indicators	PC1	PC2	PC3
CB online availability citizen	0.60	-0.35	0.72
CB online availability businesses	0.59	-0.42	-0.70
CB User Support	0.54	0.84	-0.04

(a) Loading for the first three principal components

Components	Total variance	% of Variance
1	2.48	82.63
2	0.39	12.83
3	0.14	4.54

(b) Variances for each component

Table 4.25: Results of CPCA on CB services sub-dimension

Concerning the CB sub-dimension, the results are optimal. In fact, the variance explained by the first component is 83% (Table 4.25(b)), and loadings for the first component are all high and positive (Table 4.25(a)), giving a clear indication of the existence of a unique dimension.

Cronbach's alpha across the two new sub-dimensions is measured too and holds a 0.76 value, thus confirming internal consistency across sub-dimensions.

In Appendix B Table B.3 contains the summary table of the proposed adjustment while Table 4.26 contains the summary measures of internal consistency of both the "status quo" and the proposed revision.

		CPCA: First component explained variance	Cronbach's alpha within sub-dim.	Correlation among sub-dim.
Status quo	4a	62%	0.8	–
Proposed adjustment	National services	65%	0.91	0.79
	CB services	82 %	0.89	

Table 4.26: Summarized measures of internal consistency for DPS

Both the "status quo" and the proposed adjustment reach optimal internal consistency within and across sub-dimensions. Therefore, they do not lead to compensability issues within the DPS dimension.

4.6 CPCA and ROBPCA over all indicators

To verify the presence of the four latent dimensions within DESI 2022, an exploratory analysis is conducted considering all the 33 available individual indicators ($p = 33$) and the 27 observations ($n = 27$), corresponding to the EU member states. Data are scaled to have unit variance. We start by

Components	ROBPCA		CPCA	
	Eigenvalue	% Variance	Eigenvalue	% Variance
1	11.98	36.32	12.46	37.75
2	4.11	12.44	3.93	11.90
3	2.50	7.60	2.70	8.17
4	1.91	5.80	1.79	5.42
5	1.61	4.88	1.69	5.13
6	1.37	4.14	1.56	4.72
7	1.23	3.72	1.37	4.14
8	0.85	2.59	1.29	3.92
9	0.36	1.10	1.12	3.38
10	0.18	0.55	1.00	3.03

Table 4.27: Eigenvalues and Cumulative Percentage of Variance Explained by the first 10 PCs using ROBPCA and CPCA.

performing a ROBPCA (Hubert, Rousseeuw, and Aelst 2008) analysis on the scaled high-dimensional data ($p > n$) with a default value of $\alpha = 0.75$.

Figure 4.13(a) visualizes the eigenvalues of the components obtained from the ROBPCA procedure versus the number of the components. As mentioned earlier in section 3.3, the Kaiser rule is primarily used to determine the number of principal components to retain. In fact, the focus is on checking the robustness of the CPCA method against outliers.

We examine the eigenvalues of each component to apply the Kaiser rule, thereby gaining insight into the number of subspaces to retain. Referring to Table 4.27, for both the ROBPCA and CPCA, the first three principal components consistently explain more variance than the threshold of unity dictated by the Kaiser rule. However, in the case of ROBPCA, from the fourth to the tenth PC possesses eigenvalues ranging between 2 and 1. Similarly, employing the CPCA method, the fifth, sixth, and seventh components exhibit eigenvalues of 1.61, 1.37, and 1.23, respectively. In scenarios where several components possess eigenvalues just above 1, a more conservative ap-

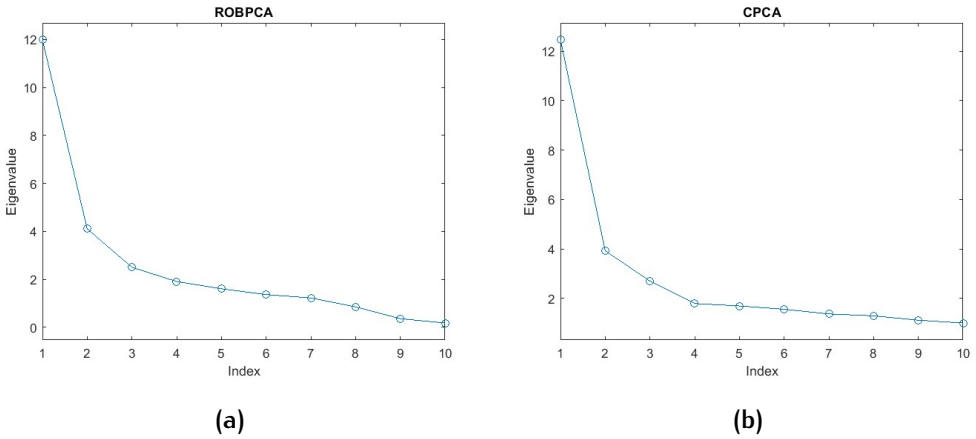
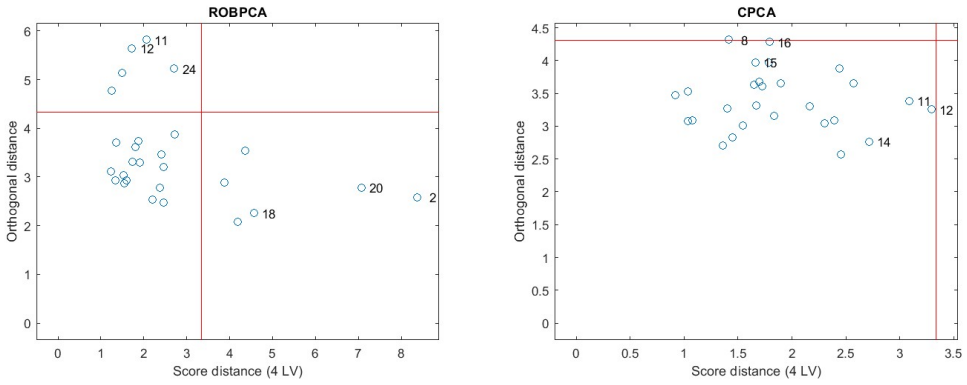


Figure 4.13: Scree plots showing 10 PCs with (a) ROBPCA; and (b) CPCA

proach is recommended, such as employing the elbow method. According to the elbow method, as depicted in Figure 4.13(a) for ROBPCA, a suggestion arises to retain four principal components, aligning with the interpretation found in Figure 4.13(b) for CPCA, where the elbow corresponds to the fourth eigenvalue. Both estimates of the underlying dimensions align well with the structure of DESI 2022, which encompasses the four core dimensions, i.e. Human Capital, Connectivity, Integration of Digital Technologies, and Digital Public Services. Based on this last criterion, we decide for both ROBPCA and CPCA to retain $k = 4$ components for outlier analysis.

The CPCA method (4.14(b)) does not flag any outliers. Just observations 8 and 16, corresponding to Estonia and Italy, lie close to the orthogonal threshold while observations 11 and 12, corresponding respectively to Finland and France, have a high score distance. On the other hand, the ROBPCA procedure detects 6 outliers with a substantially increased score *or* orthogonal distances. Observations 2, 18, and 20, corresponding respectively to countries Belgium, Luxembourg, and Malta, are regarded as good leverage points since they lie close to the PCA subspace but far from the regular observations, whereas observations 11, 12, and 24, corresponding respectively to Finland, France, and Romania, are classified as orthogonal outliers because of their large orthogonal distance to the PCA subspace (see Figure



(a) (b)

Figure 4.14: Outlier maps of the gait dataset obtained with (a) ROBPCA (k=3); and (b) CPCA (k=4)

4.14(a)). In both cases, no observations are found to be bad leverage points. The most noticeable outcome of the comparative analysis is that the outliers map quite differ. In fact, most of the mentioned observations are found to be good leverage points or orthogonal outliers in ROBPCA but CPCA, the classical method is not robust against outliers. Thus ROBPCA is more reliable in this setting, and according to Figure 4.13(a) it confirms the existence of 4 latent dimensions.

4.7 Conclusions

The internal consistency assessment over the dimensions and sub-dimensions of DESI 2022 is crucial to justify the compensatory rules employed in the final computation of DESI 2022 scores. Results concerning the "status quo" analyses were found to be not homogeneus among dimensions. In fact, in some cases, the initial framework already had a good level of internal consistency while in others, it started from a lacking consistency ground. As a consequence, the proposed adjustments would represent minor or major improvements in the respective dimension. Following, chapter findings are summarized by dimension:

- Human Capital dimension: analyzing the HC dimension revealed high internal consistency for the "status quo" analysis over all indicators, and within subdimension 1a measuring basic ICT skills, while poor internal consistency was found for the sub-dimension measuring advanced ICT skills (1b). The comparative analysis of CPCA and ROBPCA over all indicators showed the existence of potential outliers that shift the estimated subspace from the one-dimensional structure identified by the ROBPCA. Particularly, *Romania* is identified as an outlier by both robust and classical PCA implying that the data point signifies an anomaly. The *proposed adjustment* introduces a less conservative approach than the uni-dimensionality suggested by ROBPCA. In fact, it preserves the separation between the two sub-dimensions of basic and advanced ICT skills, while removing indicator 1b4 "female ICT specialist". Moreover, two new indicators "Never used interned" and "Frequency of use" are added and the sub-dimensions are reorganized according to the PCA loading results. The proposed adjustment leads to a reduction in internal consistency within the first sub-dimension of basic digital skills, while being still satisfactory, enhancing both the 1b sub-dimensional and HC dimensional robustness against compensability issues. Thus the use of compensatory aggregation rules and the inclusion of a higher number of indicators is here justified.

- Connectivity dimension: exploring the CN dimension globally, Cronbach's alpha showed a poor overall consistency within the dimension , which is also reflected within sub-dimensions. Retained principal components offered consistent insights into the true dimensional structure. The comparative analysis shows that CPCA results are biased since the outlier map of the classical procedure does not find outlying observations, while ROBPCA does. However, the results of the CPCA analysis on all indicators validate the existence of two latent sub-dimensions within the "status quo" of the CN dimension, which is in contrast with the DESI 2022 structure that formulates four sub-dimensions. The two

detected latent dimensions within CN exhibit a negative correlation with each other. Thus, the *proposed adjustment* aims at solving for both the poor internal consistency within and across sub-dimensions of CN dimension. Following that, indicators 2a1 "Overall fixed broadband take-up", 2a3 "At least 1 Gbps take-up, 2b1 "Fast broadband (NGA) coverage" are removed, and the "5G stations" indicator is added to the analysis. The proposed two adjustments, both framing a "Mobile" and "Fixed" sub-dimension, keep the "Mobile" sub-dimension fixed while modifying the "Fixed" one, by excluding (*first proposal*) or including (*second proposal*) the "Broadband price index" (2d1) indicator. At the same level of internal consistency, the second proposal is preferred since less negatively correlated with the "Mobile" sub-dimension. However, even if an optimal level of internal consistency within the "Fixed" and "Mobile" sub-dimensions is reached, the use of compensatory aggregation rules among sub-dimensions is not validated within the CN dimension. In fact, the two sub-dimensions are negatively correlated in both proposals. Additionally, it is shown that a simplified framework for the CN dimension is preferred, which encloses a lower number of individual indicators compared to the "status quo". The simplified sub-dimensional structure contains only 2 sub-dimensions compared to the initial 4.

- Integration of Digital Technologies dimension: Cronbach's alpha reliability measure over all indicators in the IDT dimension shows an optimal level of internal consistency, as well as within sub-dimensions, measuring respectively for Digital technologies for businesses (3b) and e-Commerce (3c). Sub-dimension 3a is excluded from any internal consistency analysis within the sub-dimension since it contains only one indicator. The scree plots of both CPCA and ROBPCA highlight the dominance of the first principal component in explaining the total variance. The CPCA procedure seems to be robust towards outliers since outliers maps of the robust and classical procedure are very similar,

except for one outlying observation corresponding to *Italy*, detected by ROBPCA. However, the results of the CPCA analysis on all indicators validate the uni-dimensionality hypothesis, well described by all the indicators but one, the "ICT for environmental sustainability" (3b6) indicator. This result is in contrast with the conceptual framework of DESI 2022, which instead formulates three sub-dimensions. The *proposed adjustment* involves the exclusion of the 3b6 indicator and the creation of a unique "revised IDT" dimension, under which all the remaining indicators are contained. The proposal contributes to maintaining a high level of internal consistency within IDT, reflecting the true dimensional structure detected with CPCA while thwarting potential trade-offs between sub-dimensions. Compared with the "status quo" of IDT, the dimensional framework has been simplified, and the number of individual indicators reduced.

- Digital Public Services dimension: The "status quo" of the DPS dimension does not contain sub-dimensions, thus internal consistency is tested only at an overall level. Starting from reliability coefficients, Cronbach's alpha over the DPS dimension shows a satisfactory level of internal consistency among individual indicators. Additionally, by comparing CPCA and ROBPCA it seems like the CPCA scree plot is skewed toward a unique latent vector while ROBPCA supports the existence of a double dimensionality within DPS. Both methods identify the same outliers, corresponding to *Greece* and *Romania*. This result holds a double meaning: that CPCA is robust towards outliers since can identify outlying observations similarly to ROBPCA, and that *Greece* and *Romania* stand out as true anomalous points within the context of the DPS dimension. Finally, as previously supported by the ROBPCA scree plot, CPCA loading estimates show the existence of two latent sub-dimensions within IDT. The result shows that the "Open data" (4a5) indicator contributes mainly to the second PC, while the other indicators capture the first PC. The *proposed adjustment*, which cap-

tures separately the "National" and "Cross-Border" services, reaches optimal results both in terms of within and across sub-dimension internal consistency. Moreover, it captures the bi-dimensional structure of the DPS dimension, which was not accounted for by the "status quo". Regardless, both the "status quo" and the proposed adjustment reach optimal levels of internal consistency and do not incur in compensability issues.

As a last step of the internal consistency assessment, the quadri-dimensional structure of DESI is tested and both the results from CPCA and ROBPCA support the existence of 4 dimensions within DESI 2022.

Overall, the compensatory rules are justified, both at the sub-dimensional and dimensional levels. The only case in which a satisfactory level of internal consistency is not reached is for Connectivity, where sub-dimensions of the proposed adjustments are negatively correlated.

Chapter 5

Impact of the proposed adjustments on DESI 2022

5.1 Robustness analysis

This chapter undertakes a comprehensive analysis to assess the implications of various dimensional adjustments proposed in Chapter 4, on the scores and rankings of DESI 2022. Specifically, it delves into an in-depth examination using Impact Analysis (IA) and, in some specific cases Sensitivity Analysis (SA), for each dimension of DESI.

The *impact analysis* serves as a critical tool for evaluating the robustness of the final scores and rankings. By subjecting specific inputs to variations, i.e. the dimensional structure and the individual indicator weights that will change accordingly, we gauge the resilience of the composite indicator outcomes. Notably, the IA examines the overall impact, encompassing both dimensional scores and rankings, when the proposed adjustments are applied to compute DESI. A comparative analysis is then performed between these results and the corresponding outcomes in DESI 2022. The performance measures involved in examining the repercussions of the framework updates are *QQ-plots*, which may help to see how the obtained score distribution differs

from the original distribution according to DESI 2022, and a *ranking difference* measure, which for each country computes the difference between the original ranking position and the new one, to highlight how differently are countries impacted by the proposed adjustment.

Additionally, in specific scenarios, a *sensitivity analysis* is employed to quantify the additional uncertainty stemming from testing diverse weight sets on newly defined sub-dimensional components. For example, within the context of the DPS dimension, where the adjustment introduces new sub-dimensions and necessitates the allocation of weights, SA is employed. This aims to discern the influence of distinct sub-dimensional weight configurations on the final scores. By maintaining the framework consistent with the proposed DPS adjustment, varying weight sets are examined, revealing the relative variability in scores. In contrast, the Connectivity dimension does not undergo the SA process on sub-dimensional weights, as the newly proposed sub-dimensions (Cross-Border and National services) are conceptually regarded as equivalent by DESI experts and, hence equally weighted.

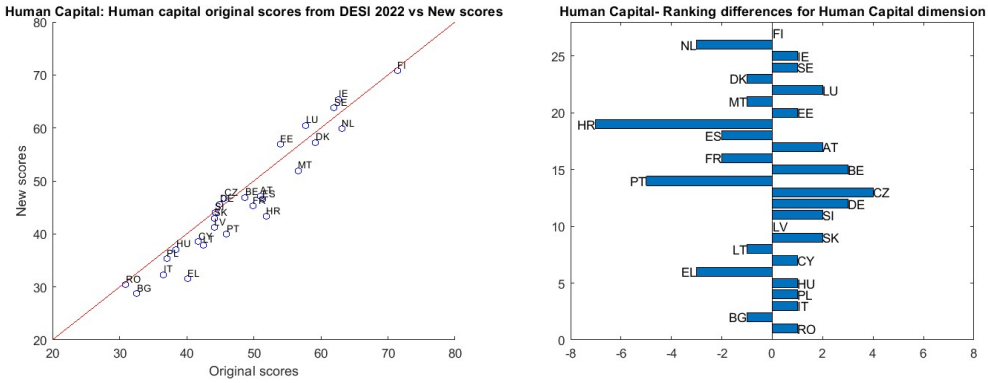
In essence, this chapter employs IA to offer a comprehensive panorama of the collective uncertainty associated with the CI. It takes into account a multitude of variance sources and their cumulative effects. On the other hand, SA is designed to isolate and quantify the influence of specific input variations, such as sub-dimensional weights, on the dimensional composite score. Importantly, for each dimension, the DESI 2022 dimensional scores and rankings are upheld as benchmarks. When updates introduce new indicators within the framework, the computation of dimensional scores entails assigning a *weight* of either 1 or 2 to each new indicator. This allocation is conducted while maintaining the remaining weights unchanged and subsequently normalizing within each sub-dimension to ensure a sum of 1 (see Table B.3 in Appendix B).

5.2 Impact analysis on Human Capital

The Human Capital proposed adjustment assumed two newly defined sub-dimensions:

- 1a new containing indicators 1a1, 1a2, 1a3, 1b3, "Never Used Internet" and "Frequency of use";
- 1b new including 1b1 and 1b4.

In the HC dimension, the newly introduced indicator "Never Used Internet" was reversed to have an orientation consistent with the others. For this indicator, the assigned dimensional weight was 1, since not related to the KPIs. In Appendix B, Table B.2 shows the original weights and the minimum and maximum values for indicators as in DESI 2022 while Table B.2 contains the new weights and the respective minimum maximum values. For the reintroduced indicator, the minimum and the maximum value for normalization are taken equally as in DESI 2020, that is the last version of the DESI where those indicators were included.



(a) QQ-plot: Original scores vs revised scores **(b)** Ranking differences at HC dimension level at dimension level

Figure 5.1: Impact of the proposed adjustment on the scores (a) and rankings (b) of HC

A direct observable consequence of the adjustment for most of the countries is that the revised dimension scores are lower than the original ones (Table 5.1(a)). This is attributable to the fact that "Never used internet" indicator has opposite orientation with respect to the rest of the indicators, thus lower scores are the result of adding it to the new HC dimension. A common phenomenon is that countries taking middle positions in the score's distribution are more sensitive to rank changes than the countries positioned on the tails of the distribution. For example, *Greece*'s revised score is remarkably lower than the original one (Figure 5.1(a)) but its ranking shifts by only 3 positions (Figure 5.1(b)). On the contrary, *Croatia* with a score close to the average, loses 7 positions while having a loss in score comparable to Greece. Additionally, Croatia is the country mostly negatively impacted by the proposed adjustment while Checz Republic is the one gaining more positions, by up scaling 4 positions in the ranking.

5.3 Impact analysis on Connectivity

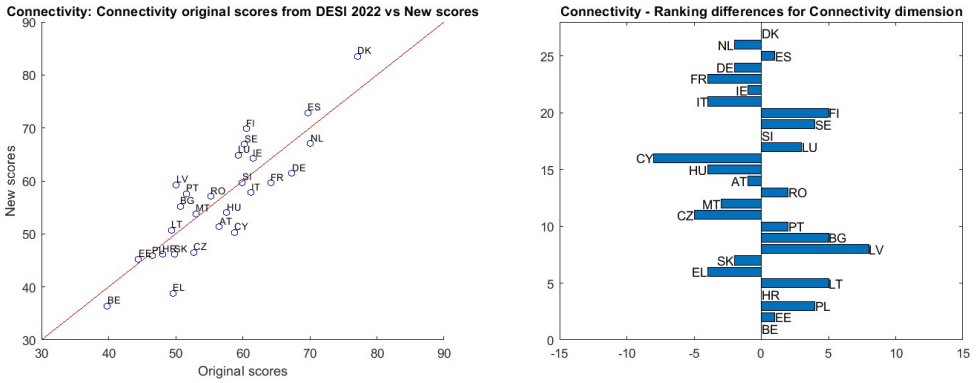
For the Connectivity dimension, we previously proposed two adjustments, each having a "Fixed" and a "Mobile" sub-dimension. While the "Mobile" sub-dimension, which includes indicators 2c1, 2c2, 2c3, and "5G stations", remains unchanged between the two proposals, the "Fixed" sub-dimension changes. In the case of the *first proposal*, "Fixed" is defined by indicators 2a2, 2b2, and 2b3, while for the *second proposal* indicator 2d1 "Broadband price index" is added to the "Fixed" sub-dimension previously described.

In the CN dimension, indicator weights at a sub-dimensional level are computed for both proposals by assigning to the "5G stations" indicator an initial weight of 1. The "Mobile" sub-dimension, where the indicator "5G stations" is added, presents an additional complexity because of missing values for four countries. For each of these countries, the score for the Mobile sub-dimension is calculated only over indicators 2c1, 2c2, and 2c3, and weights are renormalized accordingly, leading to respectively 25%, 50%, and 25% as

weights.

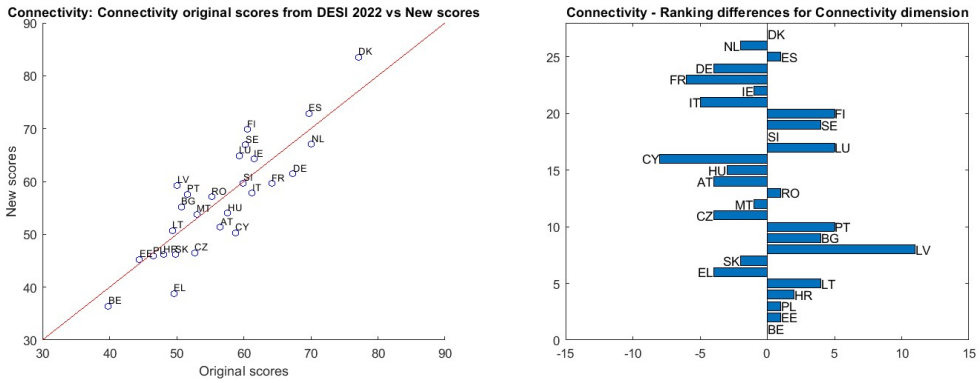
In Appendix B, Table B.2 shows the original weights and the minimum and maximum values for indicators as in DESI 2022. Table B.3 shows the new weights for respectively the first and second proposals, and the respective minimum maximum values.

Equal weighting for the "Fixed" and "Mobile" sub-dimensions is used for the computation of the final Connectivity scores for both proposals. This choice allows for the full compensability of the two sub-dimensions. For indicator normalization, the minimum and maximum values as in DESI 2022 are always used. For the newly introduced "5G stations" indicator, the minimum is taken equal to 0 while the maximum is equal to 1.14, which is double the maximum value recorded in 2022 (relative to Latvia).



(a) QQ-plot: Original scores vs revised scores **(b)** Ranking differences at CN dimension level at dimension level

Figure 5.2: Impact of the *first* proposed adjustment on the scores (a) and rankings (b) of CN



(a) QQ-plot: Original scores vs revised scores**(b)** Ranking differences at CN dimension level at dimension level

Figure 5.3: Impact of the *second* proposed adjustment on the scores (a) and rankings (b) of CN

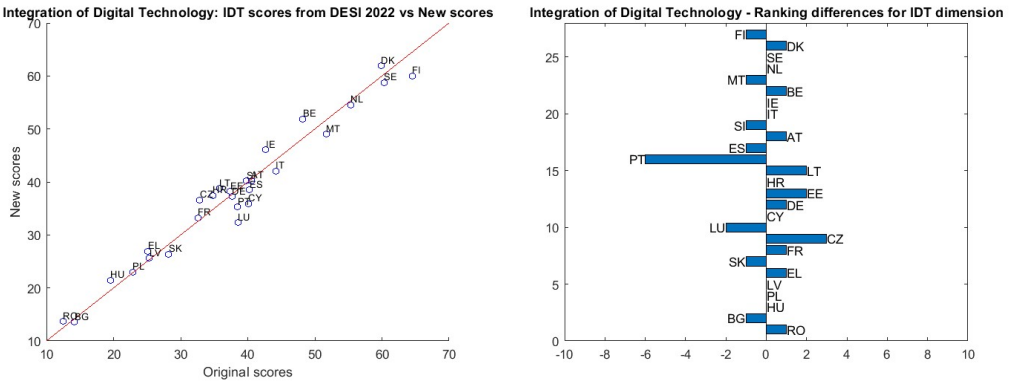
The effect on scores and ranks is very similar in the two cases (Figure 5.2 and Figure 5.3). Overall, the impact on scores is rather limited: the correlation coefficient between the original and revised scores is measured and is almost 0.9 in both cases. As expected, the effect on ranks is higher, especially for the countries with middle scores. In particular, the most affected by the two proposals countries are the same. Specifically, the one gaining more positions in the ranking is *Latvia*, which for the first proposal increases by 9 positions while for the second of 11, and the country that is most negatively impacted is *Cyprus* that for both proposals loses 9 positions.

5.4 Impact analysis on Integration of Digital Technologies

The proposed adjustment for IDT considered taking out indicator 3b6 "ICT for environmental sustainability" from the original IDT dimension and create a unique cohesive "revised IDT dimension", eliminating the need for sub-dimensions.

The IA is performed by comparing the original scores and rankings from

IDT dimension of DESI 2022 to the new ones at the dimension level. Figure 5.4(a) compares the original IDT scores versus the revised ones, while Figure 5.4(b) shows the ranking differences.



(a) QQ-plot: Original scores vs revised scores at dimension level

(b) Ranking differences at IDT dimension level

Figure 5.4: Impact of the proposed adjustment on the scores (a) and rankings (b) of IDT

Both figures show that the impact of the proposed adjustment is limited and again, more effective on the observations that are central in the original IDT score distribution. The highest negative impact on the rankings is experienced by *Portugal* (Figure 5.4(b)), which loses 6 positions, while all other observations are comparable in terms of ranking differences, which has a maximum of 3. This clearly shows how adjustments, even when applying limited framework updates, can have a different impact on different countries. This is due both to the country's position in the score distribution and to how badly/well the country performs in the removed *or* more/less weighted individual indicator. For instance, in Figure 5.5 we can see the different values of 3b6 "ICT for environmental sustainability" by country, which was removed in the proposed adjustment of IDT. It shows that the country corresponding to the highest value is Portugal, i.e. the country losing more positions compared to the DESI 2022 IDT ranking (see Figure 5.4(b)).

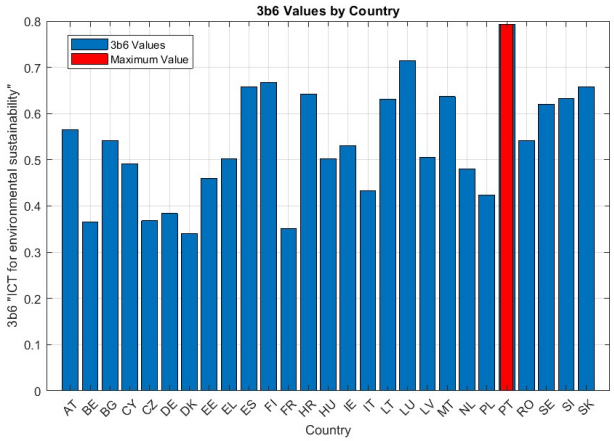


Figure 5.5: 3b6 indicator original values

5.5 Impact analysis on Digital Public Services

Within the DPS dimension, the proposed adjustment was introduced aiming at capturing the National and Cross Border (CB) services separately. In fact, two new sub-dimensions were assembled, respectively including:

- "National services" sub-dimension: 4a1 "e-Government users", 4a2 "Pre-filled forms", "Online availability citizen", "Online availability businesses", "Mobile friendliness", "User support", "Transparency";
- "CB services" sub-dimension: "CB Online availability citizen", "CB Online availability businesses", "CB User support".

The IA of the new proposed adjustment on the original scores from DESI 2022 is performed by comparing the original scores to the ones obtained using equal weights on National and CB sub-dimensions.

The new indicator weights are computed by assigning to each of the new indicators in the respective sub-dimension a weight equal to 1, keeping the other weights fixed to the original value, and then by normalizing inside each sub-dimension, so that they sum up to 1. The only exceptions are the four indicators: "CB Online availability for citizens" and "CB Online availability

for businesses", which are all double weighted since obtained by decomposing the KPIs 4a3 and 4a4 (see Section 4.5.1). Minimum and maximum values for the newly introduced indicators were obtained by following the same approach as in DESI 2022. In fact, based on the 2019 data, the minimum was computed as the actual minimum value multiplied by 0.75 and the maximum as the actual maximum value multiplied by 1.25.

The evaluation of the robustness of rankings and scores to different weighting sets, i.e. SA, for sub-dimensions, is performed here. Specifically, SA is employed for weighting allocation purposes, when uncertain whether to attribute the two new sub-dimensions equally importance (see Table B.3 in Appendix A to check the weighting sets choices).

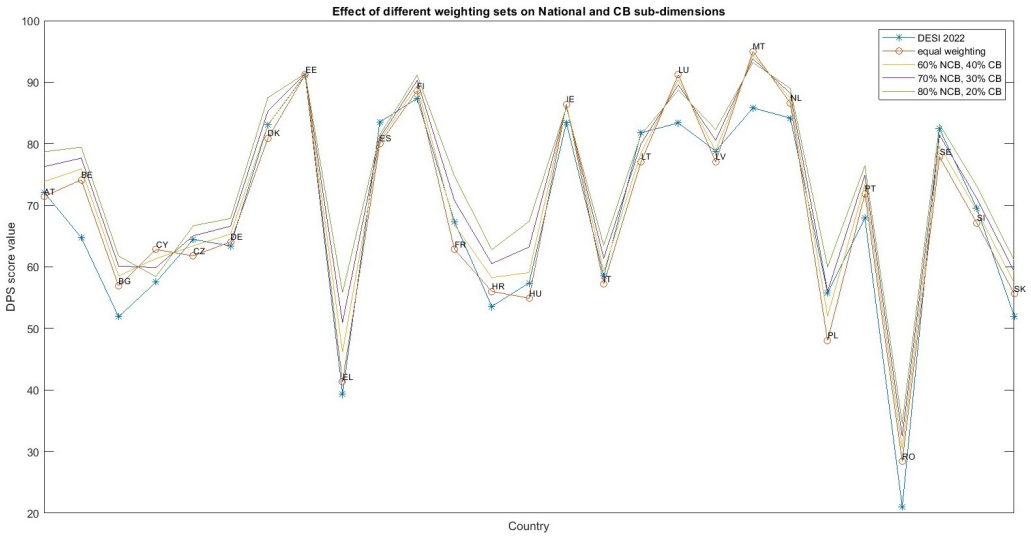


Figure 5.6: Scores of DPS dimension only using different sets of weights for the newly defined sub-dimensions

First, the variability of the DPS scores due to the 4 different sets of weights is measured and shown in Figure 5.6. We can notice that for some countries such as *Greece* and *Romania*, changing the set of weights used for weighting the sub-dimensions within DPS has a bigger impact, while for countries like *Estonia* and *The Netherlands*, different weighting systems have little or zero impact. Countries with little to no impact from varying weights may suggest

that the specific sub-dimensions within DPS contribute relatively equally to their overall performance. Conversely, countries with larger score variations could indicate that certain sub-dimensions have a more significant influence on their DPS scores. These results underscore the importance of involving relevant stakeholders, including experts and policymakers, in the weight assignment process. Engaging stakeholders can lead to a more balanced and representative choice of weights, minimizing potential bias and ensuring that the composite indicator accurately reflects the dimension importance.

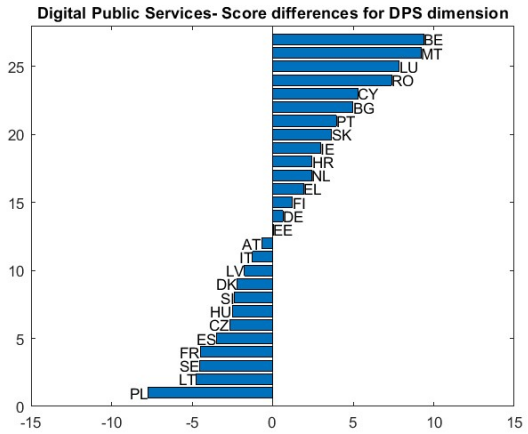


Figure 5.7: Score growth when using equal weighting

In fact, the separation between the national and the cross-border construct, when using equal weighting, leads to giving much more importance to indicators of the CB sub-dimension than it was done in the "status quo". Further exploration in Figure 5.7 identifies the countries most prominently affected by this adjustment. These countries experience the greatest shifts in scores due to the reconfiguration of weight distribution, highlighting their heightened sensitivity to the new approach.

The outcome stresses the potential implications of redefining the relationship between national and cross-border dimensions within the composite indicator framework, emphasizing the importance of careful consideration and stakeholder engagement when implementing such adjustments.

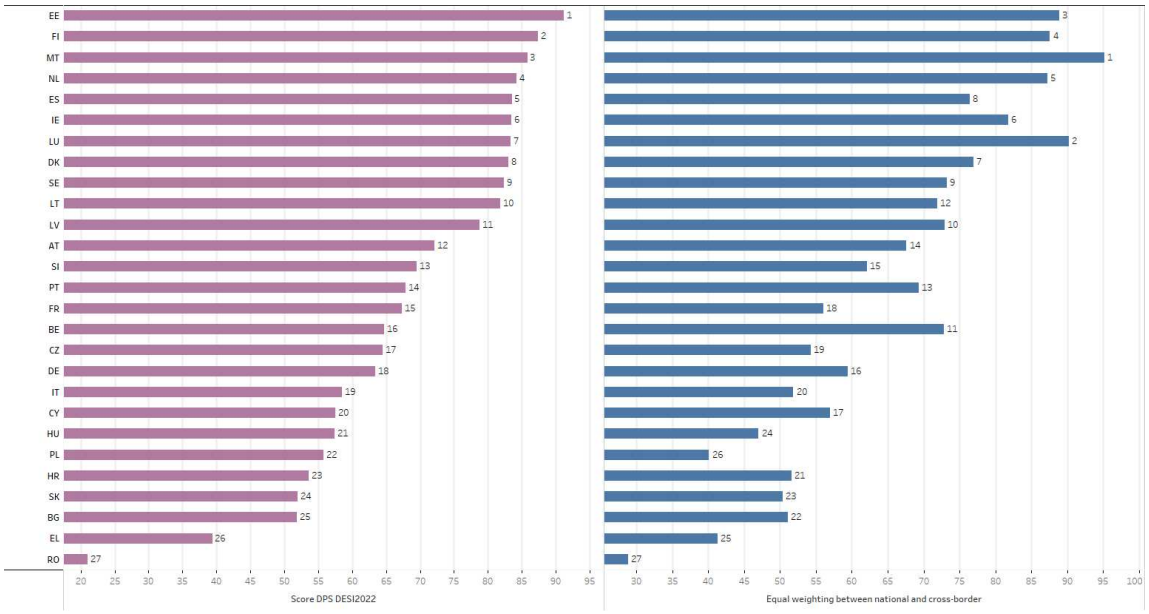


Figure 5.8: Impact of equal weighting on scores and ranks of DPS (left) compared to DESI 2022 (right)

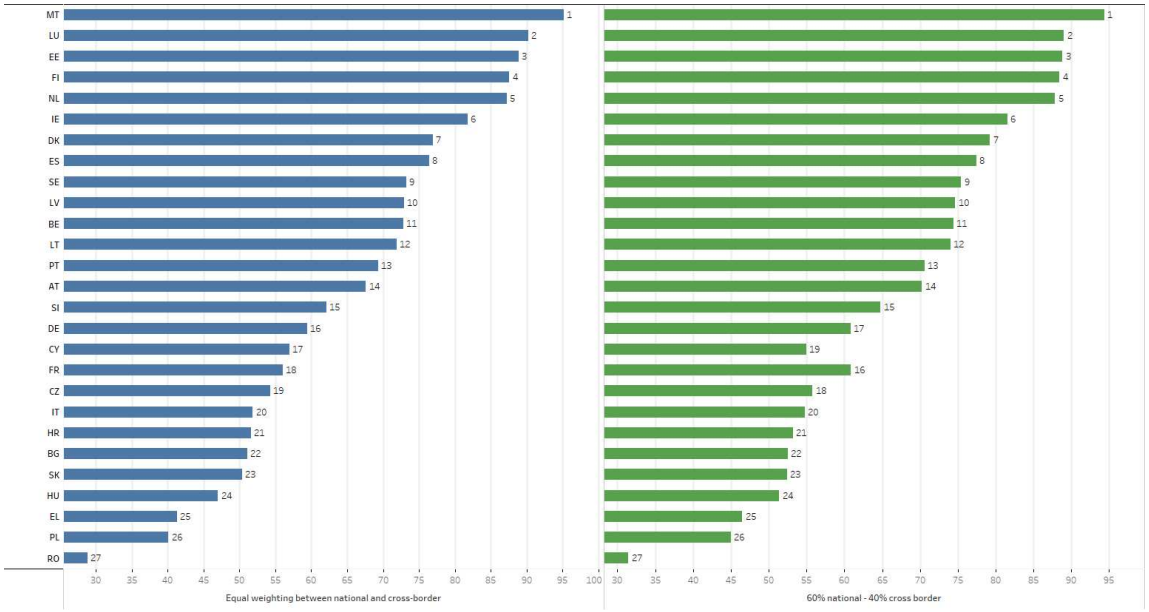


Figure 5.9: Impact of '60%-40%' weighting on scores and ranks of DPS (right) compared to equal weighting (left)

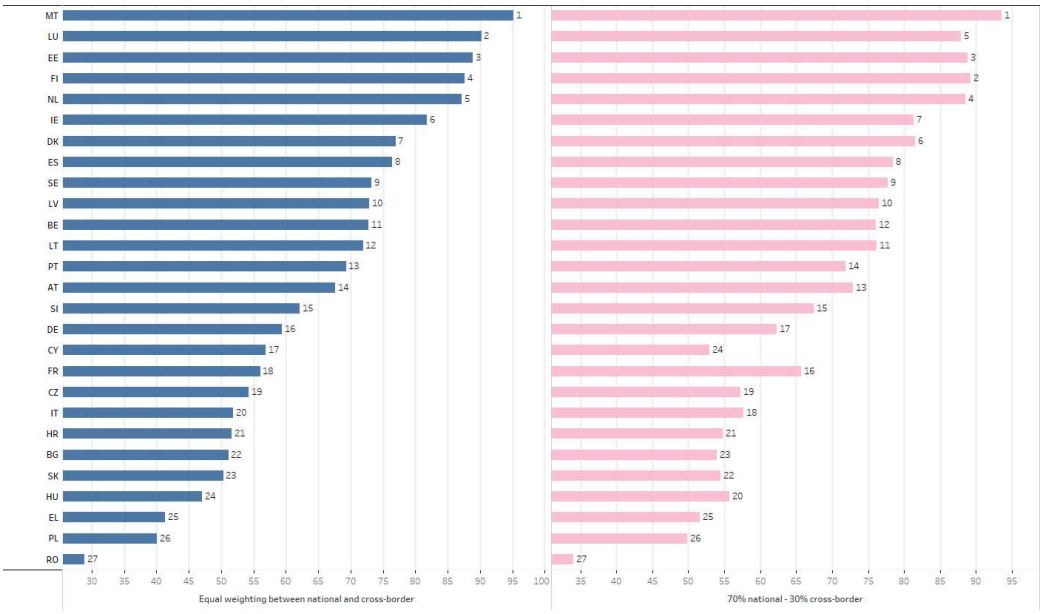


Figure 5.10: Impact of '70%-30%' weighting on scores and ranks of DPS (right) compared to equal weighting (left)

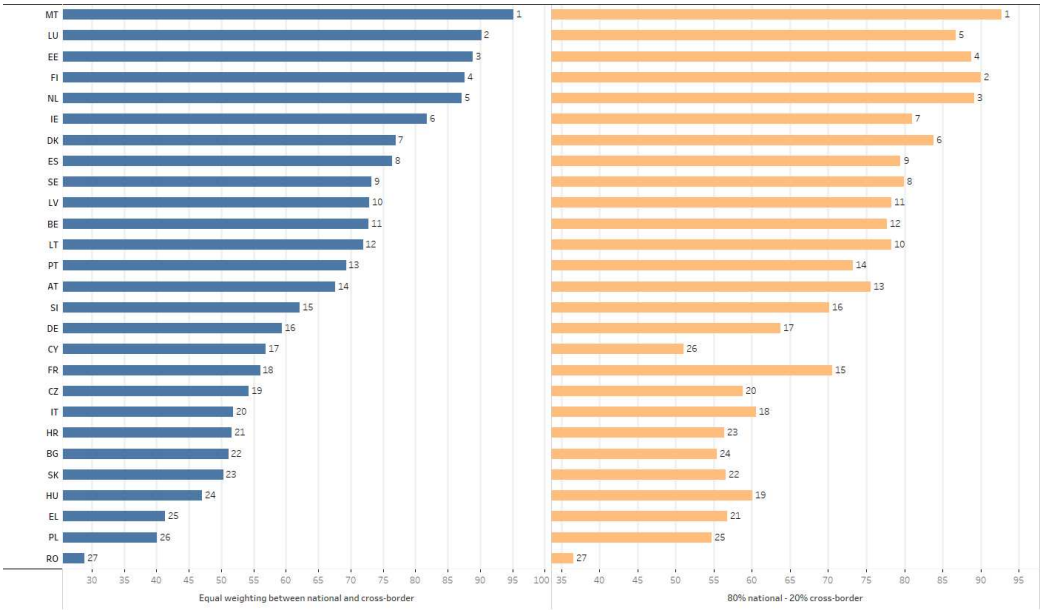


Figure 5.11: Impact of '80%-20%' weighting on scores and ranks of DPS (right) compared to equal weighting (left)

In the second part of the SA, we compare the single pairs of weighting sets scores and rankings. The equal weighting scenario for National and CB services sub-dimensions is opposed to the sets of weights 60%-40% (Figure 5.9), 70%-30% (Figure 5.10) and 80%-20% (Figure 5.11), with the higher weights always assigned to the National services sub-dimension. In terms of score impact, from Figure 5.8 we can observe that the proposed adjustment when using equal weighting, has a very high impact on most of the countries.

The countries that have achieved high scores on CB indicators, such as *Belgium* (Figure 5.12(a)), *Malta* (Figure 5.12(b)), and *Luxembourg* (Figure 5.12(c)), show the most significant increase in scores.

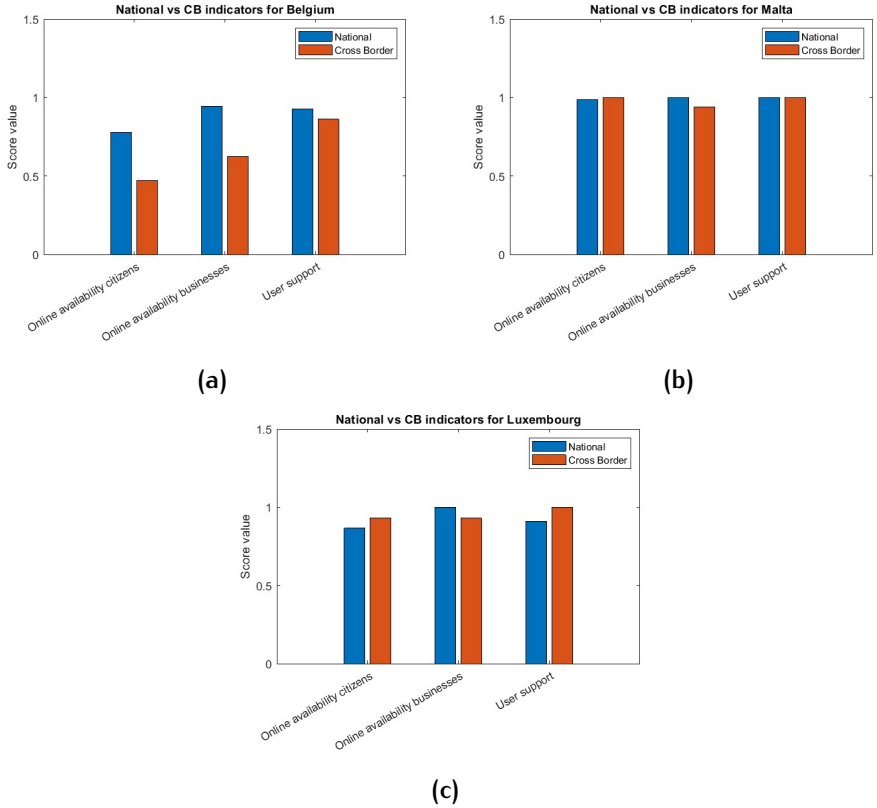


Figure 5.12: CB and National scores for (a) Belgium; (b) Malta; (c) Luxembourg

Conversely, the countries with the lowest scores on the CB dimension, such as *Poland* (Figure 5.13(a)), *Lithuania* (Figure 5.13(b)), and *Sweden*

(Figure 5.13(c)) experience the most considerable decrease in scores. The pronounced shifts in scores among these countries underscore the sensitivity of the composite indicator to changes in weight distribution.

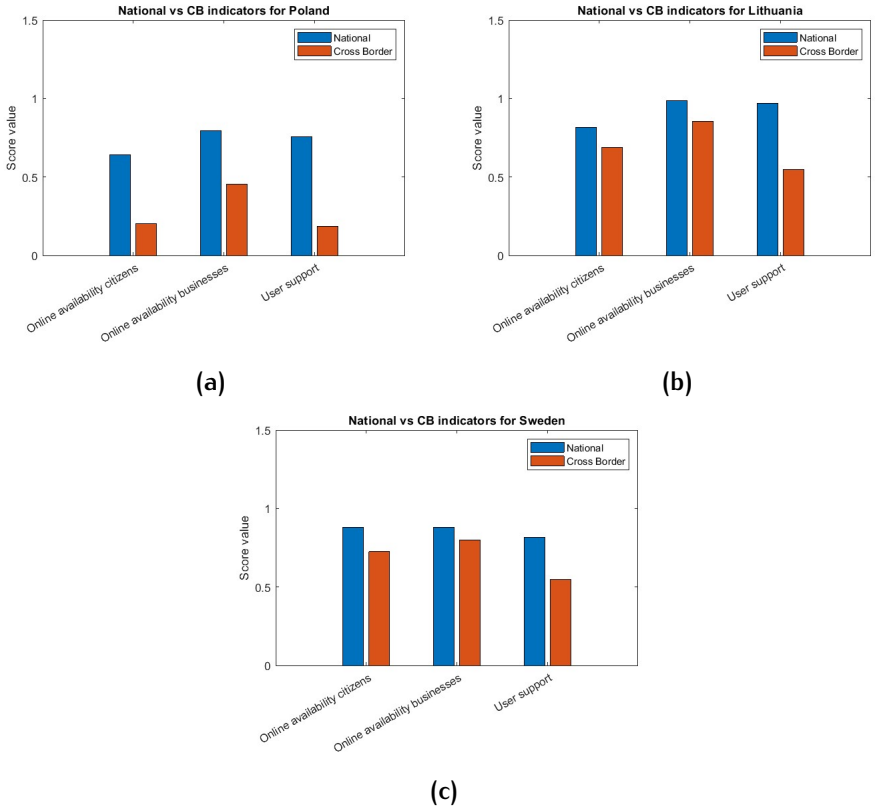


Figure 5.13: CB and National scores for (a) Poland; (b) Lithuania; (c) Sweden

On the contrary, the impact of diverse sub-dimensional weighting sets appears to be less substantial, as evidenced in Figures 5.9, 5.10, and 5.11. These findings collectively suggest that equal weighing could emerge as the most suitable option for interpretative endeavors. Moreover, equal weighting among sub-dimensions may yield a clearer and more intelligible interpretation of the composite indicator. The relatively minor fluctuations in scores under varying weight distributions emphasize the stability and consistency achieved through equal weighting, which can facilitate a more straightforward and intuitive understanding of the composite indicator implications.

5.6 Conclusions

In this chapter, impact and sensitivity analyses are employed to assess the effects of proposed adjustments to the DESI framework, highlighting the complexity of such adjustments and their varying impacts on different countries and dimensions. What is observed is that adjustment of dimensions and weights can lead to changes in country rankings. As a general behaviour, middle-ranked countries tend to be more sensitive to these changes compared to those at the higher and lower extremes of the ranking distribution.

Dimension-wise, for the HC dimension the inclusion of an indicator with an opposite orientation to the rest of the indicators results in lower scores for most countries within the HC dimension. Further, *Croatia* shows to be the country mostly negatively impacted by the proposed adjustment, and *Czech Republic* is the one gaining more positions.

In the CN dimension instead, for both proposals, it is observed that *Latvia* reaches the highest increment in ranking positions while *Cyprus* experiences the most elevated negative impact on the rankings.

Within the context of the IDT dimension, *Portugal* encounters a notable influence on its rankings, resulting in a decline of 6 positions due to the introduced adjustment. This influence can be attributed to its high performance in the removed indicator "ICT for environmental sustainability." In contrast, other countries exhibit deviations in rankings limited to a maximum of 3 positions in comparison to the DESI 2022 ranking.

Lastly, the sensitivity analysis on the DPS dimension shows that *Greece* and *Romania* are the countries with the highest variability, thus a bigger impact, on their DPS score when changing sub-dimensional weight sets. While on *Estonia* and *The Netherlands*, varying weight systems have little to no impact on their scores. This is attributable to the fact that for some countries the two new sub-dimensions count equally. In terms of score impact, we observed that the proposed adjustment when using equal weighting, has a very high impact on most of the countries. In fact, the separation between the

national and the cross-border constructs leads to giving much more importance to indicators of the CB sub-dimension than it was done in the “status quo".

The cumulative implications across dimensions affirm the thesis that changes made to the framework lead to varying effects based on specific country characteristics. This confirms the concept that enhancing internal consistency within the framework results in corresponding shifts in rankings, effectively capturing new trends stemming from the adjusted framework.

Concluding remarks

The Digital Economy and Society Index has been a reference instrument, since its creation in 2014, to benchmark EU Member States within the Digital Decade policy programme. As such, its need for redefinition is argued every year by different stakeholders (EU Commission, EU Member States, National Governments, Industry and Businesses, etc.) whose focus is set on trying to keep pace with the Digital Era. This results in the correspondence of the structure of DESI with the underlying phenomenon it seeks to capture being relegated to the background.

Lacking a strong conceptual framework, assessing the internal consistency of dimensions within DESI becomes a crucial step for two reasons: aligning the conceptual framework with the intended phenomenon of digitalization, and justifying the use of compensability rules involved in the computation of the scores. In fact, the CI uses an equal-weighted arithmetic mean among dimensions, and weighted arithmetic means for sub-dimensions within the same dimension. These aggregation rules allow respectively for full compensability of dimensions, and strong compensability (depending on the sub-dimensional weights) of sub-dimensions. Compensatory aggregation rules rely on the assumption of trade-offs between indicators. Thus, verifying internal consistency in this setting helps to confirm that DESI scores accurately represents the way in which different individual indicators contribute to the overall estimate.

The assessment conducted in Chapter 4 on DESI 2022 holds paramount importance in validating two crucial methodological choices integral to the

composite indicator framework, while addressing this work *major objectives*:

1. The utilization of compensatory rules for aggregation.
2. The inclusion of detailed individual indicators to comprehensively capture the diverse facets of the digital domain.

Additionally, Chapter 4 confronts the DESI experts' decision to not correct for univariate outlying behaviours of individual indicators of DESI 2022. Thus, to ensure the robustness of the methods employed in redefining the DESI framework, a comparison was made between Classical PCA and its robustified counterpart (ROBPCA) at both the overall composite indicator level and dimensional level. This comparison addressed two *minor objectives*:

1. Testing the robustness of PCA against outliers.
2. Detecting potential outlying countries through measures of multivariate outlyingness.

Yet, at the end of Chapter 4, another focal point is confronted, which centers around the adherence of DESI to the cardinal points, encompassing the four dimensions of digital capacity: Human Capital (HC), Connectivity (CN), Integration of Digital Technologies (IDT), and Digital Public Services (DPS).

The following conclusions are derived concerning the *major* and *minor* objectives set forth:

- In the Human Capital dimension, the *proposed adjustment*, compared to the "status quo" leads to a reduction in internal consistency within the new sub-dimension of basic digital skills, while being still satisfactory, enhancing both the advanced sub-dimensional and HC dimensional robustness against compensability issues. Additionally, the new dimension supports the inclusion of a higher number of indicators. The comparative analysis of CPCA and ROBPCA instead shows the existence of potential outliers that shift the estimated subspace away from

the one-dimensional structure identified by the ROBPCA. Anyway, results show that CPCA can be considered robust.

- For the Connectivity dimension, the internal consistency analysis of the "status quo" showed poor overall consistency. The *proposed adjustments* lead to an optimal level of internal consistency within the two new sub-dimensions, "Fixed" and "Mobile". However, the use of compensatory aggregation rules among those is not validated since the two new sub-dimensions are negatively correlated in both proposals. Additionally, the new proposals suggest a simplified framework for the CN dimension, where a lower number of individual indicators is present. Finally, the comparative analysis between CPCA and ROBPCA showed that the classical procedure is not robust against outlying observations.

- Within Integration of Digital Technologies dimension, the "status quo" internal consistency analysis provided with optimal results. However, the *proposed adjustment* is more in line with the uni-dimensionality hypothesis suggested by the results. The proposal contributes to maintaining a high level of internal consistency while preventing potential trade-offs between sub-dimensions. In this case, the dimensional framework has been simplified, and the number of individual indicators reduced. The CPCA is considered robust and there are no outlying observations that are consistently outlying among the two procedures.

- The Digital Public Services reaches a satisfactory level of internal consistency in the "status quo" analysis. The *proposed adjustment*, which captures separately the "National" and "Cross-Border" services, reaches optimal results of internal consistency while capturing the bi-dimensional structure of the DPS dimension which was not accounted for by the "status quo". This leads to preventing compensability issues. Additionally, the comparison of CPCA with ROBPCA shows that the classical procedure is robust towards outliers In fact, it identifies the same outlying observations as the robust procedure, thus identifying true

anomalous points.

Additionally, the internal consistency assessment confirms the existence of 4 latent dimensions within DESI 2022, that align with the four cardinal points. In summary, the internal consistency assessment results show that the proposed adjustments lead to an improvement of the conceptual framework robustness towards compensatory effects, except in the case of Connectivity. In fact, for the latter, sub-dimensions of the proposed adjustments are negatively correlated and lead to unreliable Connectivity scores. Hence, a policy advice is to look at the scores and rankings of Connectivity sub-dimensions separately, to avoid the information stemming from each being covered by trade-off effects.

In the concluding stages, Chapter 5 undertakes an evaluation of the repercussions stemming from the proposed adjustments on dimensional scores and rankings. It becomes evident that modifications in dimension weights can exert an influence on the rankings of countries. However, the alteration in scores and rankings does not exhibit uniformity across dimensions. The cumulative effects spanning various dimensions validate the hypothesis that alterations introduced into the framework yield divergent outcomes contingent upon distinct country attributes. This substantiates the notion that augmenting the intrinsic coherence within the framework gives rise to corresponding shifts in rankings, adeptly encapsulating emerging trends stemming from the adapted framework. It's worth noting that while this influence reverberates through the individual dimensions, its resonance at the overarching DESI level appears to be of a relatively lesser magnitude.

In summary, this comprehensive analysis provides a thorough understanding of the composite indicator framework across dimensions. Methodological choices have been scrutinized, and the framework robustness has been validated. Outliers effects vary across dimensions, with some dimensions showcasing more sensitivity than others. Overall, these findings contribute to a comprehensive understanding of digital progress, enabling policymakers to make informed decisions in line with the Digital Decade objectives.

Appendix A

Summary statistics

A.1 Human Capital Indicators

Indicator name	1a1 At least basic digital skills	1a2 Above basic digital skills	1a3 At least basic digital content creation skills	1b1 ICT specialists
% of missing values	0.00	0.00	0.00	0.00
average	0.56	0.28	0.67	0.05
standard deviation	0.12	0.10	0.11	0.01
coefficient of variation	0.21	0.36	0.16	0.28
skewness	-0.27	0.35	-0.60	0.71
kurtosis	3.24	3.12	3.06	2.73
skewness correction	no	no	no	no
orientation	positive	positive	positive	positive
maximum value	0.79	0.52	0.83	0.08

country correspon--ding to maximum value	FI	NL	NL	SE
minimum value	0.28	0.08	0.41	0.03
country correspon--ding to minimum value	RO	BG	RO	RO

Table A.1: Univariate summary statistics for indicators 1a1, 1a2, 1a3, 1b1 in HC

Indicator name	1b2 Female ICT specialists	1b3 Enterprises providing ICT training	1b4 ICT graduates
% of missing values	0.00	0.00	3.70
average	0.20	0.21	0.05
standard deviation	0.04	0.08	0.02
coefficient of variation	0.19	0.36	0.38
skewness	0.77	0.08	0.66
kurtosis	3.25	2.66	2.96
skewness correction	no	no	no
orientation	positive	positive	positive
maximum value	0.28	0.38	0.09
country correspon--ding to maximum value	BG	FI	IE
minimum value	0.10	0.06	0.01
country correspon--ding to minimum value	CZ	RO	IT

Table A.2: Univariate summary statistics for indicators 1b2, 1b3, 1b4 in HC

	Indicators						
Indicators	1a1	1a2	1a3	1b1	1b2	1b3	1b4
1a1	1.00	0.96	0.95	0.73	-0.13	0.71	0.26
1a2		1.00	0.87	0.71	-0.01	0.66	0.25
1a3			1.00	0.69	-0.15	0.65	0.24
1b1				1.00	0.03	0.73	0.41
1b2					1.00	-0.12	0.38
1b3						1.00	0.16
1b4							1.00

Table A.3: Correlation Table for indicators in HC

A.2 Connectivity Indicators

Indicator name	2a1 Overall fixed broadband take-up	2a2 At least 100 Mbps fixed broadband take-up	2a3 At least 1 Gbps take-up	2b1 Fast broadband (NGA) coverage
% of missing values	0.00	0.00	3.70	0.00
Average	0.79	0.40	0.04	0.92
standard deviation	0.09	0.18	0.06	0.07
coefficient of variation	0.12	0.44	1.78	0.08
skewness	-0.34	0.22	2.58	-1.04
kurtosis	2.38	2.05	8.96	3.33
skewness correction	no	no	no	no
orientation	positive	positive	positive	positive
maximum value	0.97	0.72	0.27	1.00
country correspon--ding to maximum value	NL	ES	FR	CY
minimum value	0.61	0.09	0.00	0.74
country correspon--ding to minimum value	FI	EL	AT	FR

Table A.4: Univariate summary statistics for indicators in 2a1, 2a2, 2a3, 2b1 CN

Indicator name	2b2 Fixed Very High Capacity Network (VHCN) coverage	2b3 Fibre to the Premises (FTTP) coverage	2c1 5G spectrum	2c2 5G coverage
% of missing values	0.00	0.00	0.00	0.00
Average	0.58	0.56	0.46	0.88
standard deviation	0.24	0.33	0.32	0.06
coefficient of variation	0.27	0.41	0.59	0.70
skewness	-0.83	-0.41	-0.28	0.25
kurtosis	3.11	2.09	1.99	1.75
skewness correction	no	no	no	no
orientation	positive	positive	positive	positive
maximum value	1	0.89	1.00	1.00
country correspon--ding to maximum value	MT	LV	DE	IT
minimum value	0.20	0.10	0.00	0.00
country correspon--ding to minimum value	EL	BE	EE	LV

Table A.5: Univariate summary statistics for indicators 2b2, 2b3, 2c1, 2c2 in CN

Indicator name	2c3 Mobile broadband take-up	2d1 Broadband price index
% of missing values	0.00	0.00
Average	72.73	0.73

standard deviation	0.06	11.21
coefficient of variation	0.07	0.15
skewness	-0.14	0.16
kurtosis	2.49	2.12
skewness correction	no	no
orientation	positive	positive
maximum value	0.98	96.52
country correspon--ding to maximum value	IE	RO
minimum value	0.73	56.25
country correspond--ing to minimum value	BG	BE

Table A.6: Univariate summary statistics for indicators 2c3, 2d1 in CN

	Indicators									
Indicators	2a1	2a2	2a3	2b1	2b2	2b3	2c1	2c2	2c3	2d1
2a1	1.00	0.22	0.01	0.49	0.09	-0.19	0.07	0.08	0.43	-0.58
2a2		1.00	0.25	0.22	0.70	0.49	-0.17	-0.35	0.36	0.04
2a3			1.00	-0.22	0.07	0.15	0.03	0.09	-0.01	0.06
2b1				1.00	0.19	-0.1	-0.07	0.07	0.08	-0.40
2b2					1.00	0.65	-0.3	-0.32	0.38	0.17
2b3						1.00	-0.18	-0.37	0.05	0.38
2c1							1.00	0.33	0.11	-0.21
2c2								1.00	0.22	-0.06
2c3									1.00	-0.20
2d1										1.00

Table A.7: Correlation Table for indicators in CN

A.3 Integration of Digital Technologies Indicators

Indicator name	3a1 SMEs with at least a basic level of digital intensity	3b1 Electronic information sharing	3b2 Social media	3b3 Big data
% of missing values	0.00	0.00	0.00	0.00
average	0.55	0.37	0.30	0.14
standard deviation	0.16	0.11	0.11	0.08
coefficient of variation	0.29	0.29	0.37	0.55
skewness	-0.11	-0.09	0.19	0.67
kurtosis	2.54	2.24	2.13	2.08
skewness correction	no	no	no	no

orientation	positive	positive	positive	positive
maximum value	0.86	0.57	0.51	0.30
country correspon--ding to maximum value	SE	BE	FI	MT
minimum value	0.61	0.09	0.00	0.74
country correspon--ding to minimum value	FI	EL	AT	FR

Table A.8: Univariate summary statistics for indicators 3a1, 3b1, 3b2, 3b3 in IDT

Indicator name	3b4 Cloud	3b5 AI	3b6 ICT for environmental sustainability	3b7 e-Invoices
% of missing values	0.00	0.00	7.41	3.70
average	0.37	0.08	0.67	0.30
standard deviation	0.16	0.05	0.09	0.23
coefficient of variation	0.45	0.65	0.13	0.76
skewness	0.38	1.09	0.09	1.50
kurtosis	2.25	4.07	2.16	4.30
skewness correction	no	no	no	no
orientation	positive	positive	positive	positive
maximum value	0.69	0.24	0.86	0.95
country correspon--ding to maximum value	SE	DK	PT	IT
minimum value	0.20	0.01	0.54	0.10

country correspon--ding to minimum value	EL	RO	DK	BG

Table A.9: Univariate summary statistics for indicators 3b4, 3b5, 3b6, 3b7 in IDT

Indicator name	3c1 SMEs selling online	3c2 e-Commerce turnover	3c3 Selling online cross-border
% of missing values	0.00	11.11	0.00
average	0.21	0.12	0.09
standard deviation	0.08	0.05	0.03
coefficient of variation	0.38	0.41	0.36
skewness	0.43	0.22	0.25
kurtosis	2.1921	2.4687	2.0635
skewness correction	no	no	no
orientation	positive	positive	positive
maximum value	0.38	0.22	0.16
country correspon--ding to maximum value	DK	IE	AT
minimum value	0.09	0.03	0.04
country correspon--ding to minimum value	LU	LU	BG

Table A.10: Univariate summary statistics for indicators 3c1, 3c2, 3c3 in IDT

	Indicators										
Indicators	3a1	3b1	3b2	3b3	3b4	3b5	3b6	3b7	3c1	3c2	3c3
3a1	1.00	0.55	0.89	0.66	0.90	0.65	0.03	0.48	0.66	0.50	0.65
3b1		1.00	0.62	0.44	0.31	0.63	0.02	0.10	0.30	0.22	0.46
3b2			1.00	0.65	0.74	0.56	0.08	0.29	0.49	0.28	0.55
3b3				1.00	0.61	0.64	-0.15	0.14	0.52	0.41	0.47
3b4					1.00	0.56	-0.08	0.64	0.58	0.56	0.51
3b5						1.00	0.17	0.35	0.42	0.37	0.48
3b6							1.00	0.01	-0.06	-0.15	-0.06
3b7								1.00	0.22	0.32	0.17
3c1									1.00	0.76	0.76
3c2										1.00	0.53
3c2											1.00

Table A.11: Correlation Table for indicators in IDT

A.4 Digital Public Services Indicators

Indicator name	4a1 e-Government users	4a2 Pre-filled forms	4a3 Digital public services for citizens
% of missing values	0.00	0.00	0.00
average	0.71	64.49	74.63
standard deviation	0.19	20.87	13.73
coefficient of variation	0.27	0.32	0.18
skewness	-1.01	-0.37	-0.29
kurtosis	3.6802	2.0815	2.4098
skewness correction	no	no	no
orientation	positive	positive	positive

maximum value	0.93	94.32	99.64
country correspon--ding to maximum value	SE	NL	MT
minimum value	0.17	19.05	44.24
country correspon--ding to minimum value	RO	RO	RO

Table A.12: Univariate summary statistics for indicators 4a1, 4a2, 4a3 in DPS

Indicator name	4a4 Digital public services for businesses	4a5 Open data
% of missing values	0.00	0.00
average	81.71	0.81
standard deviation	13.59	0.15
coefficient of variation	0.17	0.18
skewness	-1.31	-0.93
kurtosis	4.8820	2.5951
skewness correction	no	no
orientation	positive	positive
maximum value	100.00	0.98
country correspon--ding to maximum value	IE	FR
minimum value	42.27	0.50

country correspon--ding to minimum value	RO	SK

Table A.13: Univariate summary statistics for indicators 4a4, 4a5 in DPS

	Indicators				
Indicators	4a1	4a2	4a3	4a4	4a5
4a1	1.00	0.62	0.66	0.64	0.14
4a2		1.00	0.75	0.66	0.06
4a3			1.00	0.85	-0.03
4a4				1.00	0.13
4a5					1.00

Table A.14: Correlation Table for indicators in DPS

A.5 New Indicators: summary statistics

Indicator name	Never Used Internet (2021)	Frequency Internet Use (2021)	2a2 At least 100 Mbps fixed broadband take-up (*new data*)	5G stations in the 3.6 band
% of missing values	0.00	0.00	0.00	0.00
average	7.93	88.00	0.16	0.47
standard deviation	5.22	6.23	0.16	0.20
coefficient of variation	0.66	0.07	0.96	0.43
skewness	0.46	-0.39	1.30	0.03
kurtosis	2.61	2.56	2.07	4.02
skewness correction	no	no	no	no
orientation	positive	positive	positive	positive
maximum value	20.00	98.00	0.57	0.83
country correspon--ding to maximum value	EL	IE	LV	ES
minimum value	0.00	74.00	0.00	0.09
country correspon--ding to minimum value	IE	BG	NL	EL

Table A.15: Univariate summary statistics for newly introduced/updated indicators of HC and CN

Indicator name	Public services for citizens (national)	Public services for business (national)	Mobile friendliness	User Support
% of missing values	0.00	0.00	0.00	0.00
average	88.98	97.28	93.27	92.83
standard deviation	8.59	3.72	6.54	6.55
coefficient of variation	0.10	0.04	0.07	0.07
skewness	-1.43	-1.98	-0.98	-1.38
kurtosis	4.76	10.61	2.71	3.11
skewness correction	no	no	no	no
orientation	positive	positive	positive	positive
maximum value	100.00	100.00	100.00	100
country correspon--ding to maximum value	MT	CZ	FI	EL
minimum value	64.50	83.70	77.20	71.40
country correspon--ding to minimum value	CY	RO	RO	CY

Table A.16: Univariate summary statistics for newly introduced indicators of DPS

Indicator name	Transparency	Cross-border Online Availability citizens	Cross-border Online Availability businesses	Cross-border User Support
% of missing values	0.00	0.00	0.00	0.00
average	62.82	61.87	67.39	70.78

standard deviation	16.29	20.39	22.93	19.23
coefficient of variation	0.26	0.33	0.34	0.27
skewness	0.13	-0.08	-0.97	0.09
kurtosis	2.41	2.26	3.82	1.95
skewness correction	no	no	no	no
orientation	positive	positive	positive	positive
maximum value	97.95	100.00	100.00	100
country correspon--ding to maximum value	MT	MT	IE	LU
minimum value	30.57	24.25	8.33	33.33
country correspon--ding to minimum value	CY	RO	EL	PL

Table A.17: Univariate summary statistics for newly introduced indicators of DPS

Appendix B

Data sources and weights

Data source	**Data collection process**
Eurostat	Data collected and verified by the national statistical offices or by Eurostat.
Communications Committee (COCOM)	Data collected and verified by the national regulatory authorities (by data experts appointed by the members of the Communications Committee in every Member State).
Broadband coverage studies	Data collected by IHS Markit, Omdia and Point Topic and verified by the national regulatory authorities (by data experts appointed by the members of the Communications Committee.
Retail broadband prices studies	Data collected by Empirica and verified by the national regulatory authorities (by data experts appointed by the members of the Communications Committee in every Member States).
e-Government benchmark	Data collected by Capgemini and verified by relevant ministries in every Member State.
Survey of businesses on the use of digital technologies	Data collected by Ipsos and iCite, survey results have been reviewed by the Digital Single Market

	Strategic Group
European data portal	Data collected by Capgemini from representatives appointed by the relevant ministries in every Member State

Table B.1: Data sources and the role of national authorities

Indicator	Description	Unit	Source
1a1 At least basic digital skills	Individuals with 'basic' or 'above basic' digital skills in each of the following five dimensions: information, communication, problem solving and software for content creation and safety	% individuals	Eurostat - European Union survey on ICT usage in Households and by Individuals (I_DSK2_BAB)
1a2 Above basic digital skills	Individuals with 'above basic' digital skills in each of the following five dimensions: information, communication, problem solving and software for content creation and safety	% individuals	Eurostat - European Union survey on ICT usage in Households and by Individuals (I_DSK2_AB)
1a3 At least basic digital content creation skills	Individuals with at least a basic level in using software for digital content creation	% individuals	Eurostat - European Union survey on ICT usage in Households and by Individuals (I_DSK2_DCC_BAB)
1b1 ICT specialists	Employed ICT specialists. Broad definition based on the ISCO-08 classification and including jobs like ICT service managers, ICT professionals, ICT technicians, ICT installers and servicers.	% individuals in employment aged 15-74	Eurostat - Labour force survey (isoc_sks_itspt)
1b2 Female ICT specialists	Employed female ICT specialists. Broad definition based on the ISCO-08 classification and including jobs like ICT service managers, ICT professionals, ICT technicians, ICT installers and servicers.	% ICT specialists	Eurostat - Labour force survey (isoc_sks_itsps)
1b3 Enterprises providing ICT training	Enterprises who provided training in ICT to their personnel	% enterprises	Eurostat - European Union survey on ICT usage and eCommerce in Enterprises (E_ITT2)
1b4 ICT graduates	Individuals with a degree in ICT	% graduates	Eurostat (table educ_uoe_grad03, using selection ISCED11=ED5-8) and ISCEDF_13 [F06] Information and Communication Technologies

Figure B.1: Data sources for indicators in the HC dimension taken from DESI 2022 Methodological Note (European Commission (2022))

Indicator	Description	Unit	Source
2a1 Overall fixed broadband take-up	% of households subscribing to fixed broadband	% households	Eurostat - European Union survey on ICT usage in Households and by Individuals [H_BBFIX]
2a2 At least 100 Mbps fixed broadband take-up	% of households subscribing to fixed broadband of at least 100 Mbps, calculated as overall fixed broadband take-up (source: Eurostat) multiplied with the percentage of fixed broadband lines of at least 100 Mbps (source: COCOM)	% households	European Commission, through the Communications Committee (COCOM) and Eurostat - European Union survey on ICT usage in Households and by Individuals
2a3 At least 1 Gbps take-up	% of households subscribing to fixed broadband of at least 1 Gbps, calculated as overall fixed broadband take-up (source: Eurostat) multiplied with the percentage of fixed broadband lines of at least 1 Gbps (source: COCOM)	% households	European Commission, through the Communications Committee (COCOM) and Eurostat - European Union survey on ICT usage in Households and by Individuals
2b1 Fast broadband (NGA) coverage	% of households covered by fixed broadband of at least 30 Mbps download. The technologies considered are FTTH, FTTB, Cable Docsis 3.0 and VDSL	% households	Broadband coverage in Europe studies for the European Commission by IHS Markit, Omdia and Point Topic
2b2 Fixed Very High Capacity Network (VHCN) coverage	% of households covered by any fixed VHCN. The technologies considered are FTTH and FTTB for 2015-2018 and FTTH, FTTB and Cable Docsis 3.1 for 2019 onwards	% households	Broadband coverage in Europe studies for the European Commission by IHS Markit, Omdia and Point Topic
2b3 Fibre to the Premises (FTTP) coverage	% of households covered by FTTH and FTTB	% households	Broadband coverage in Europe studies for the European Commission by IHS Markit, Omdia and Point Topic
2c1 5G spectrum	The amount of spectrum assigned and ready for 5G use within the so-called 5G pioneer bands. These bands are 700 MHz (703-733 MHz and 758-788 MHz), 3.6 GHz (3400-3800 MHz) and 26 GHz (1000 MHz within 24250-27500 MHz). All three spectrum bands have an equal weight	Assigned spectrum as a % of total harmonised 5G spectrum	European Commission services, through the Communications Committee (COCOM)
2c2 5G coverage	% of populated areas with coverage by 5G	% populated areas	Broadband coverage in Europe studies for the European Commission by IHS Markit, Omdia and Point Topic
2c3 Mobile broadband take-up	Individuals who used the internet on a mobile device	% individuals	Eurostat - European Union survey on ICT usage in Households and by Individuals [I_IUG_MD]
2d1 Broadband price index	The broadband price index measures the prices of representative baskets of fixed, mobile and converged broadband offers	Score (0-100)	Broadband retail prices study, annual studies for the European Commission realised by Empirica

Figure B.2: Data sources for indicators in the Connectivity dimension taken from DESI 2022 Methodological Note (European Commission (2022))

Indicator	Description	Unit	Source
3a1 SMEs with at least a basic level of digital intensity	The digital intensity score is based on counting how many out of 12 selected technologies are used by enterprises. A basic level requires usage of at least 4 technologies.	% SMEs	Eurostat - European Union survey on ICT usage and eCommerce in Enterprises
3b1 Electronic information sharing	Enterprises who have in use an ERP (enterprise resource planning) software package to share information between different functional areas (e.g. accounting, planning, production, marketing)	% enterprises	Eurostat - European Union survey on ICT usage and eCommerce in Enterprises (E_ERP1)
3b2 Social media	Enterprises using two or more of the following social media: social networks, enterprise's blog or microblog, multimedia content sharing websites, wiki-based knowledge sharing tools. Using social media means that the enterprise has a user profile, an account or a user license depending on the requirements and the type of the social media.	% enterprises	Eurostat - European Union survey on ICT usage and eCommerce in Enterprises (E_SM1_GE2)
3b3 Big data	Enterprises analysing big data from any data source	% enterprises	Eurostat - European Union survey on ICT usage and eCommerce in Enterprises (E_BDA)
3b4 Cloud	Enterprises buying sophisticated or intermediate cloud computing services	% enterprises	Eurostat - European Union survey on ICT usage and e-commerce in enterprises (E_CC1_SI)
3b5 AI	Enterprises using any AI technology	% enterprises	Eurostat - European Union survey on ICT usage and e-commerce in enterprises (E_AI_TANY)
3b6 ICT for environmental sustainability	The indicator measures the level of support that adopted ICT technologies offered to enterprises to engage in more environmentally-friendly actions. The level of intensity is measured based on the number of environmental actions (maximum 10) reported by enterprises to have been facilitated by the use of ICT. The following categorisation was achieved: low intensity (0 to 4 actions), medium intensity (5 to 7 actions) and high intensity (8 to 10 actions).	% enterprises having medium/high intensity of green action through ICT	Survey of businesses on the use of digital technologies by Ipsos and iCite
3b7 e-Invoices	Enterprises sending e-invoices, suitable for automated processing	% enterprises	Eurostat - European Union survey on ICT usage and eCommerce in Enterprises (E_INV4S_AP)
3c1 SMEs selling online	SMEs selling online (at least 1% of turnover)	% SMEs	Eurostat - European Union survey on ICT usage and eCommerce in Enterprises (E_ESELL)
3c2 e-Commerce turnover	SMEs' total turnover from e-commerce	% SME turnover	Eurostat - European Union survey on ICT usage and eCommerce in Enterprises (E_ETURN)
3c3 Selling online cross-border	SMEs that carried out electronic sales to other EU countries	% SMEs	Eurostat - European Union survey on ICT usage and eCommerce in Enterprises (E_AESEU)

Figure B.3: Data sources for indicators in the IDT dimension taken from DESI 2022 Methodological Note (European Commission (2022))

Indicator	Description	Unit	Source
4a1 e-Government users	Individuals who used the Internet, in the last 12 months, for interaction with public authorities	% internet users	Eurostat - European Union survey on ICT usage in Households and by Individuals (I_IUGOV12)
4a2 Pre-filled forms	Amount of data that is pre-filled in public service online forms	Score (0 to 100)	eGovernment Benchmark
4a3 Digital public services for citizens	The share of administrative steps that can be done online for major life events (birth of a child, new residence, etc.) for citizens	Score (0 to 100)	eGovernment Benchmark
4a4 Digital public services for businesses	The indicator broadly reflects the share of public services needed for starting a business and conducting regular business operations that are available online for domestic as well as foreign users. Services provided through a portal receive a higher score, services which provide only information (but have to be completed offline) receive a more limited score.	Score (0 to 100)	eGovernment Benchmark
4a5 Open data	This composite indicator measures to what extent countries have an open data policy in place (including the transposition of the revised PSI Directive), the estimated political, social and economic impact of open data and the characteristics (functionalities, data availability and usage) of the national data portal.	% maximum score	European data portal

Figure B.4: Data sources for indicators in the DPS dimension taken from DESI 2022 Methodological Note (European Commission (2022))

Dimension	Sub-dim.	Sub-dim. weight	Indicator	Weight	Normalized weight	Min-Max
1 Human capital	1a	50%	**1a1***	2	50%	0% - 100%
			1a2	1	25%	0% - 66%
			1a3	1	25%	25% - 100%
	1b	50%	**1b1***	2	33.33%	0% - 10%
			1b2*	2	33.33%	0% - 10%
			1b3	1	16.67%	0% - 50%
			1b4	1	16.67%	0% - 10%
2 Connectivity	2a	25%	**2a1**	1	33.33%	50% - 100%
			2a2	1	33.33%	0% - 100%
			2a3	1	33.33%	0% - 50%
	2b	25%	**2b1**	1	25%	25% - 100%
			2b2*	2	50%	0% - 100%
			2b3	1	25%	0% - 100%
	2c	40%	**2c1**	1	25%	0% - 100%

			2c2*	2	50%	0% - 100%
			2c3	1	25%	25% - 100%
	2d	10%	**2d1**	1	100%	25 - 100
3 Integration of digital technology	3a	15%	**3a1***	2	100%	20% - 100%
	3b	70%	**3b1**	1	10%	0% - 60%
			3b2	1	10%	0% - 60%
			3b3*	2	20%	0% - 75%
			3b4*	2	20%	0% - 75%
			3b5*	2	20%	25% - 75%
			3b6	1	10%	30% - 100%
			3b7	1	10%	0% - 100%
	3c	15%	**3c1**	1	33.33%	0% - 50%
			3c2	1	33.33%	0% - 33%
			3c3	1	33.33%	0% - 25%

4	Digital public services	4a	100%	**4a1**	1	14.29%	0% - 100%
				4a2	1	14.29%	0- 100
				4a3*	2	28.57%	35 - 100
				4a4*	2	28.57%	45 - 100
				4a5	1	14.29%	0% - 100%

Table B.2: Weights and min-max values for indicators in DESI 2022

Dimension	Sub-dim.	Sub-dim. weight	Indicator	Weight	Normalized weight	Min-Max
1 Human capital	1a *new*	50%	**1a1***	2	28.57%	0% - 100%
			1a2	1	14.29%	0% - 66%
			1a3	1	14.29%	25% - 100%
			1b3	1	14.29%	0% - 50%
			Never Used Internet	1	14.29%	0% - 45%
			Frequency of use	1	14.29%	40% - 100%
	1b *new*	50%	**1b1***	2	66.67 %	0% - 10%
			1b4	1	33.33%	0% - 10%
2 Connectivity *first proposal*	*Fixed sub-dim.*	50%	**2a2**	1	25%	0% - 100%
			2b2*	2	50%	0% - 100%
			2b3	1	25%	0% - 100%

	Mobile sub-dim.	50%	**2c1**	1	20%	0% - 100%
			2c2	2	40%	0% - 100%
			2c3	1	20%	25% - 100%
			5G stations in the 3.6 band	1	20%	0 - 1.14
2	Connect-ivity *second proposal*	*Fixed sub-dim.*	50%	**2a2**	1	20%	0% - 100%
				2b2*	2	40%	0% - 100%
				2b3	1	20%	0% - 100%
				2d1	1	20%	25% - 100%
		Mobile sub-dim.	50%	**2c1**	1	20%	0% - 100%
				2c2	2	40%	0% - 100%
				2c3	1	20%	25% - 100%

				5G stations in the 3.6 band	1	20%	0 - 1.14
3	Integration of digital technology	*No sub-dim.*	100%	**3a1***	2	14.29%	20% - 100%
				3b1	1	7.14%	0% - 60%
				3b2	1	7.14%	0% - 60%
				3b3*	2	14.29%	0% - 75%
				3b4*	2	14.29%	0% - 75%
				3b5*	2	14.29%	25% - 75%
				3b7	1	7.14%	0% - 100%
				3c1	1	7.14%	0% - 50%
				3c2	1	7.14%	0% - 33%
				3c3	1	7.14%	0% - 25%
4	Digital public services	*National* services	50% 60% 70% 80%	**4a1**	1	7.69%	0% - 100%
				4a2	1	7.69%	0 - 100

			Online availab. cit.	2	15.38%	44 - 100
			Online availab. bus.	2	15.38%	54 - 100
			Mobile friend.	1	7.69%	56 - 100
			User supp.	1	7.69%	57 - 100
			Trans-parency	1	7.69%	22 - 100
	CB services	50% 40% 30% 20%	***CB Online avail cit.***	2	40%	18 - 100
			CB Online avail bus.	2	40%	6 - 100
			CB User support	1	20%	25 - 100

Table B.3: Weights and min-max values for the new proposed adjustments (red horizontal line in CN means that one of the two proposals must be chosen; matching colors in DPS sub-dimensional weights corresponds to the set of proposed weights for the SA)

Bibliography

[1]S. Alkire and J. Foster, "Counting and multidimensional poverty measurement", Journal of Public Economics **95**, 476–487 2011.

[2]P. Annoni, "Different ranking methods: potentialities and pitfalls for the case of european opinion poll", Environmental and Ecological Statistics **14**, 453–471 2007.

[3]P. Annoni and R. Brüggemann, "Exploring partial order of european countries", Social indicators research **92**, 471–487 2009.

[4]G. E. P. Box, "The exploration and exploitation of response surfaces: some general considerations and examples", Biometrics **10**, 16–60 1954.

[5]R. Brüggemann and L. Carlsen, "Multi-criteria decision analyses. viewing mcda in terms of both process and aggregation methods: some thoughts, motivated by the paper of huang, keisler and linkov", Science of the total environment **425**, 293–295 2012.

[6]B. R. Cartell, "The scree test for the number of factors", Multivariate Behavioral Research **1**, 245–276 1966.

[7]L. Cronbach, "Coefficient alpha and the internal structure of tests", Psychometrika **16**, 297–334 1951.

[8]J. Custance and H. Hillier, "Statistical issues in developing indicators of sustainable development", Journal of the Royal Statistical Society. Series A (Statistics in Society) **161**, 281–290 1998.

[9]K. Decancq and M. Lugo, "Weights in multidimensional indices of well-being: an overview", Econometric Reviews **32** 2010.

[10]K. Decancq and M. A. Lugo, "Weights in multidimensional indices of well-being: an overview", Econometric Reviews **32**, 7–34 2013.

[11]J. R. Edwards and R. P. Bagozzi, "On the nature and direction of relationships between constructs and measures.", Psychological methods **5 2**, 155–74 2000.

[12]S. El Gibari et al., "Evaluating university performance using reference point based composite indicators", Journal of Informetrics **12**, 1235–1250 2018.

[13]European Commission, *Digital economy and society index (desi) 2022 - methodological note*, tech. rep. (2022).

[14]European Commission, *Europe's digital decade*, Online, 2021.

[15]L. Fabrigar et al., "Evaluating the use of exploratory factor analysis in psychological research", Psychological Methods **4**, 272 1999.

[16]M. Freudenberg, "Composite indicators of country performance", 2003.

[17]M. Freudenberg, "Composite indicators of country performance: a critical assessment", STI Working Paper **16** 2003.

[18]D. Gervini, "A robust and efficient adaptive reweighted estimator of multivariate location and scatter", Journal of Multivariate Analysis **84**, 116–144 2003.

[19]A. Gifi, *Nonlinear multivariate analysis*, Wiley Series in Probability and Statistics (Wiley, 1990).

[20]G. H. Golub and C. F. Van Loan, *Matrix computations*, Third (The Johns Hopkins University Press, 1996).

[21]S. Greco et al., "On the methodological framework of composite indices: a review of the issues of weighting, aggregation, and robustness", Social Indicators Research **141**, 1–34 2019.

[22]T. Greyling and F. Tregenna, "Construction and Analysis of a Composite Quality of Life Index for a Region of South Africa", Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement **131**, 887–930 2017.

[23]S. Gruijters, "Using principal component analysis to validate psychological scales: bad statistical habits we should have broken yesterday ii", 2019.

[24]R. K. Henson, "Understanding internal consistency reliability estimates: a conceptual primer on coefficient alpha", Measurement and Evaluation in Counseling and Development **34**, 177–189 2001.

[25]M. Hubert, P. Rousseeuw, and K. Vanden Branden, "Robpca: a new approach to robust principal components analysis", Technometrics **47**, 64–79 2005.

[26]M. Hubert and M. Debruyne, "Breakdown value", Wiley Interdisciplinary Reviews: Computational Statistics **1**, 296–302 2009.

[27]M. Hubert, P. J. Rousseeuw, and S. V. Aelst, "High-breakdown robust multivariate methods", Statistical Science **23** 2008.

[28]International Telecommunication Union, *The ict development index*, tech. rep. (2017).

[29]R. Johnson and D. Wichern, *Applied multivariate statistical analysis*, 4. ed (Prentice Hall, 1998).

[30]I. T. Jolliffe, *Principal component analysis*, 2nd ed., Springer series in statistics (Springer, New York (N.Y.), 2002).

[31]H. Kaiser, "The application of electronic computers to factor analysis", Educational and Psychological Measurement **20**, 141–151 1960.

[32]J. Kaufman and W. Dunlap, "Determining the number of factors to retain: a windows-based fortran-imsl program for parallel analysis", Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc **32**, 389–95 2000.

[33]R. Ledesma et al., "The scree test and the number of factors: a dynamic graphics approach", The Spanish Journal of Psychology **18** 2015.

[34]G. Munda, "Choosing aggregation rules for composite indicators", Social Indicators Research **109**, 337–354 2012.

[35]G. Munda, *Social multi-criteria evaluation for a sustainable economy*, Operation Research and Decision Theory Series (Springer, Heidelberg, Germany, 2008).

[36]M. Nardo et al., *Handbook on constructing composite indicators: methodology and user guide*, 2nd ed. (OECD publishing, 2008).

[37]G. Nicoletti et al., "Summary indicators of product market regulation with an extension to employment protection legislation", 2000.

[38]P. Nomikos and J. F. MacGregor, "Multi-way partial least squares in monitoring batch processes", Chemometrics and Intelligent Laboratory Systems **30**, 97–108 1995.

[39]J. C. Nunnally, *Psychometric theory*, 2. ed (McGraw-Hill New York, 1978).

[40]F. Pennoni et al., "The 2005 european e-business readiness index", University Library of Munich, Germany, MPRA Paper 2006.

[41]J. Rawls, *A theory of justice* (The Belknap press of Harvard University Press, 1971).

[42]P. J. Rousseeuw, "Least median of squares regression", Journal of the American statistical association, 871–880 1984.

[43]M. Saisana et al., "Uncertainty and sensitivity techniques as tools for the analysis and validation of composite indicators", Journal of the Royal Statistical Society Series A **168**, 307–323 2005.

[44]S. Seth, "Inequality, interactions, and human development", Journal of Human Development and Capabilities **10**, 375–396 2009.

[45]D. Slottje, "Measuring the quality of life across countries", The Review of Economics and Statistics **73**, 684–93 1991.

[46]K. Taber, "The use of cronbach's alpha when developing and reporting research instruments in science education", Research in Science Education **48**, 1–24 2018.

[47]S. Tarantola et al., "The internal market index 2004", 2004.

[48]United Nations, "Human development report", 2001.

[49]G. Ursachi et al., "How reliable are measurement scales? external factors with indirect influence on reliability estimators", Procedia Economics and Finance **20**, 679–686 2015.

[50]S. Verboven and M. Hubert, "Libra: a matlab library for robust analysis", Chemometrics and Intelligent Laboratory Systems **75**, 127–136 2005.

[51]M. E. Wall et al., "Singular value decomposition and principal component analysis", 2002.

[52]World Economic Forum, *The network readiness index 2019: boosting ai Innovation and entrepreneurship in the Middle East and North Africa* (World Economic Forum, 2019).