Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Design of a Modular DSS for Public Health Decision-making in the Context of a COVID-19 Pandemic Landscape

Sergey Samoilenko, Kweku-Muata Osei-Bryson

PII: S0957-4174(21)01676-6
DOI: https://doi.org/10.1016/j.eswa.2021.116385
Reference: ESWA 116385

To appear in: Expert Systems with Applications

Received Date: 11 November 2020
Revised Date: 18 October 2021
Accepted Date: 6 December 2021

Please cite this article as: Samoilenko, S., Osei-Bryson, K-M., Design of a Modular DSS for Public Health Decision-making in the Context of a COVID-19 Pandemic Landscape, Expert Systems with Applications (2021), doi: https://doi.org/10.1016/j.eswa.2021.116385

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Elsevier Ltd. All rights reserved.
Design of a Modular DSS for Public Health Decision-making in the Context of a COVID-19 Pandemic Landscape
Sergey Samoilenko
Department of CIS/CS, Averett University, Danville, VA, 24541, U.S.A.
ssamoilenko@averett.edu

Kweku-Muata Osei-Bryson
Department of Information Systems, Virginia Commonwealth University, Richmond, VA, 23284, U.S.A.
KMOsei@vcu.edu

1 Corresponding Author. Email: ssamoilenko@averett.edu
Abstract: The awareness of the occurrence of a new disease involves much uncertainty and the search for answers and also appropriate questions. In this paper we focus on the perspective of public health decision-makers. Typically, they would have a standard set of questions and supporting metrics that have been found in previous disease outbreaks to be useful in assessing the effectiveness of various ‘solution’ methods on the trajectory of the disease. There may be other relevant questions with which such public health domain experts may not be familiar and/or for which they are familiar but are not aware of methods for addressing such questions when there is limited data. Decision Support Systems (DSS) can be used to facilitate the exploration of established questions and some other relevant questions. Given an initial set of questions, the DSS designer should consider which sets of data analytic methods have the capabilities to adequately address. Some of these data analytic methods may also have the capability of addressing questions that could be of interest to the public health decision makers including researchers. In this paper we present a conceptual design for a relevant easy-to-construct DSS and an example of a multi-method DSS that is based on this conceptual design. Using publicly available data on the CoViD-19 pandemic, we illustrate benefits of the multi-method DSS in action.

Keywords: Public Health; Decision Support System; Modular Design; Data Analytics.

1. Introduction

A process for dealing with a pandemic, or with the spread of any disease, could be seen as being comprised of two distinctive, yet interrelated parts that are combined into a system characterized, as a result, by a high cohesion and loose coupling of its components. The first component would be a general decision-making model (e.g. Samoilenko & Osei-Bryson, 2013; Babaei & Bamdad, 2020; Shi & Fung, 2017) that is customizable to the context of a given pandemic via the information provided by the second model. Unlike the first component, the second model is a specific to a given disease decision support system (DSS) that would be primarily utilized and relied upon by the relevant public health domain decision makers.

The field of information systems (IS) is uniquely and advantageously positioned to help fighting spread of a disease via designing and developing a DSS (Collins, Ketter & Gini, 2010) that could support public health decision makers in a pandemic. The purpose of this paper is to outline a design of such a DSS. We accomplish this goal via three objectives. First, we offer a conceptual design of a DSS based on the set of explicitly stated assumptions and premises. Second, we offer a blueprint of the system by designing an appropriate multi-method methodology that relies on the tools of data analysis and data mining that are commonly used in the field of IS. Finally, we test the resultant decision support system in the context of the current pandemic using publicly available data.
Our effort is suggestive, rather than definitive, in its purpose, for we are not claiming that our approach is the only way that the DSS could be designed and constructed. Instead, we invite our readers to contribute their ideas and to use our effort as a possible foundation in creating and testing better designs.

We present our work in the following sequence: Section 2 describes conceptual design of the DSS. Section 3 outlines the methodology that implements the design, and Sections 4 and 5 are focused on the testing of the system using publicly available data. Conclusion and Discussion’ sections close the paper.

2. Conceptual Design of the DSS
Any conceptual design of an artifact relies on a particular model of the environment within which the artifact is intended to operate. We rely on a set of assumptions (A#) that serve the purpose of the describing the environment within which a disease spreads. We invite our reader to examine each of the assumptions, along with the corresponding justifications, presented below in Table 1.

| Assumption | Justification |
|------------|--------------|
| A1         | For each disease, there is an identifiable set of demographic risk factors that provides a useful characterization of the disease. For each known disease, public health practitioners attempt to identify variables that have high correlations with its occurrence and outcomes. For each disease, the identified variables are its ‘risk factors’. Once these ‘risk factors’ have been identified public health agencies aim to collect data that can be used to estimate that associated population-level statistics. |
| A2         | The population of a geographic area (e.g., nation, state, city, county, etc.) could be described by a set of demographic factors that includes the risk factors for known diseases. In practice this typically holds at the national and state levels, and in some cases at the county and city level. |
| A3         | The population of a geographic area could be characterized by a disease-related subset of its demographic factors i.e. the risk factors for a given disease. In practice for known diseases this typically holds at the national and state levels, and in some cases at the county and city level. |
| A4         | Populated geographic areas could be grouped in terms of the risk factors for a given disease. If A3 holds then A4 should also be possible. |
| A5         | A geographic area could be assessed in terms of the impact of the risk factors for a given disease on the associated infection rate. The reasonableness of this assumption follows from the meaning of the concept of a ‘risk factor’ for a disease. |
| A6         | The spread of an infectious disease follows a path towards geographic areas with higher demographic risk factors. The reasonableness of this assumption follows from the meaning of the concept of a ‘risk factor’ for a disease. |
| A7         | A geographic area could be assessed in terms of its relative effectiveness and efficiency of containing the spread of a disease. The presence of any organized system of healthcare is associated with collecting of the relevant patient data. |
| A8         | A spread of an infectious disease follows a path towards geographic areas with lower levels of effectiveness and efficiency of containing the spread of a disease. The reasonableness of this assumption would follow from the meaning of the concepts of ‘effectiveness’ and ‘efficiency’ of disease containment. |
A geographic area could be assessed in terms of the changes in its level of efficiency of containing a disease over time. If A7 holds then A9 would also hold.

The efficiency and effectiveness of a geographic area in containing a disease could be improved via area-specific decisions (e.g. allocation of resources). If A10 does not hold then it would not make sense to have a DSS that supports the making of appropriate area-specific decisions.

Table 1. Conceptual Design of a DSS: Underlying Assumptions and Justifications

Conceptually, the design of the DSS could be perceived as consisting of two modules. The first module allows for projecting the direction/flow of contagion across geographic areas based on the demographic factors. The second module allows for modeling the spread of the disease based on the effectiveness and efficiency of the geographic areas in fighting the outbreak. The operation of the modules is supported by the following two propositions (P#):

P1: A preferred path of the spread of a disease is toward geographic area(s) characterized by high-risk demographic factors specific to the disease.

P2: A path of contagion is directed towards the geographic area(s) that is (are) least efficient and effective in containing the spread of the disease.

In order to defend the propositions, we use the rules of hypothetico-deductive logic so our reader can examine veracity of the statements.

In regard to P1, the supporting argument is as follows:

**Major premise:** A spread of a disease X is associated with a set of demographic risk factors

**Minor premise:** Area A has a greater proportion of the population with a set of X-specific risk factors that are at a higher level than that of Area B

**Conclusion:** Area A will have a greater level of contagion of X than Area B.

The supporting argument for P2 is as follows:

**Major premise:** A spread of a disease X is associated with a level of efficiency and effectiveness of containment of the outbreak

**Minor premise:** Area A has a lower level of efficiency and effectiveness of containing X than Area B

**Conclusion:** Area A will have a greater level of contagion of X than Area B.

The above-stated set of assumptions allows for outlining a set of steps that the proposed DSS should allow a decision maker for performing. This set of steps is grouped into two modules, where each module is would be applied at a different state/stage of a pandemic.

**Module 1: Identify a path of contagion based on demographic risk factors**

*State of the pandemic: Initial stage, pre-pandemic, patient zero to limited number of cases.*
**Step 1**: Identify a set of demographic risk factors for the given disease. The outcome of this step is a disease-specific sub-set of the variables that are used to describe the population of interest.

**Step 2**: Test the assumption of homogeneity of geographic areas of interest by using the available demographic data. The outcome of this step is a set of groups of geographic areas that may differ in terms of the general demographic factors.

**Step 3**: Identify the demographic variables that differentiate the set of groups of geographic areas the most. The outcome of this step is a sub-set of the variables that are, possibly, also risk factors for a given disease.

**Step 4**: Discover naturally occurring associations between demographic risk factors and the level of contagion. The outcome of Step 4 is a confirmation of the relationship between the demographic risk factors and the contraction of the disease.

**Step 5**: Test for the presence of the impact of the demographic risk factors on the spread of the disease. The result would allow for identifying causal impact of the risk factors on the contraction of the disease.

The output of using Module 1 is an *n*-tiered “projection of a path” system comprised of, if $n=3$, Low Risk, Mid Risk, and High Risk areas and constructed using the actionable information obtained in Step 5.

**Module 2: Identify a path of contagion based on efficiency and effectiveness of containing the spread**

*State of the pandemic: Developing stage, pandemic is growing, number of cases is rising.*

**Step 6**: Identify groups of geographic areas based on the effectiveness of containing the spread of the disease. This would allow for ranking the geographic areas in terms of the relative success in dealing with the given disease.

**Step 7**: Compare the groups identified in Step 6 with the groups identified in Step 2.

**Step 8**: Identify the factors differentiating the areas discerned as a result of Step 6.

**Step 9**: Assess a relative efficiency of containing the spread of the disease of a geographic area vis-à-vis other areas. The outcome of this step is a ranked order of areas reflecting their relative standing of each area vis-à-vis the better and worse performing counterparts.

**Step 10**: Assess the changes, over time, in the relative efficiency of each geographic area in regard to dealing with the contagion. The outcome of this step is an assessment, of each area, in regard to its improvement/deterioration of its performance in fighting the disease that took place over the given time period(s).

**Step 11**: Identify the drivers of the change in performance of a geographic area in fighting the disease. The outcome of this step is the discernment of the culprit of the improving or deteriorating performance.
The output of using Module 2 is an identification of *High*, *Mid*, and *Low* risk areas based on the efficiency and effectiveness in containing the spread of the diseases (and minimizing the mortality rate) with a correspondent suggestion for the improvements.

At this point we are ready to present a blueprint of the DSS supported by the tools of the data analysis and data mining commonly available to the researchers and practitioners in the field of IS.

3. Design of the DSS based on an Integrated Multi-Method Workbench

The process of translation of a conceptual model into a blueprint of the design of DSS could follow two distinct paths. First option is to rely on a custom solution that is subject- and purpose-specific. This is akin to creating a program/application by relying on custom, one-off code. Such an option, while not without its merits, is characterized by the difficulties in maintaining and adapting the design when changes are called for. The second option is to opt for a “building block” approach, where the translation of the concept into the blueprint is achieved via existing and tried-and-true components. This is similar to creating a program by utilizing existing libraries of algorithms, data structures, and classes. As a result, the design is modular, highly transparent, and adaptable.

We follow the second approach in translating a conceptual model of DSS into the design blueprint, where established data analysis methods serve as the elements comprising the end product. We believe that the selected “building blocks” approach would allow our reader to not only evaluate the appropriateness and fit for the purpose of each component, but also to appraise the soundness of the overall design of the DSS. While we do not claim that the selected methods (i.e. Cluster Analysis, Decision Tree induction, Association Rules Mining, Data Envelopment Analysis, Multiple Regression) are the only and the best options for designing a DSS, we do suggest that the chosen methods are appropriate and well-suited for the purpose. We invite our reader to consider possible substitutes to the selected methods that may contribute to a more flexible and robust design of a DSS.

We offer a brief overview of the insights offered by each method, along with some of the limitations, in Table 2 below. For further details on these data analytic methods the reader could consult various studies (e.g. Samoilenko & Osei-Bryson, 2017, Osei-Bryson & Ngwenyama, 2014).

| Method               | Offered Insight                                                                 | Limitation                                                                 |
|----------------------|--------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| CA: Cluster Analysis | Allows for testing an assumption of homogeneity of the sample and identifying presence of sub-groups in the sample. | In the presence of multiple sub-groups does not offer any insights into the sources of heterogeneity. |
| DTI: Decision Tree Induction | Given the target variable, allows for identifying attributes responsible for differentiating sub-groups of the sample. | Target variable must be provided “from outside.” Does not consider impact of differentiating variables on an “Input → Output” model of sub-groups. |
ARM: Association Rules Mining
Allows for identifying a set of “If ⇒ Then” rules present in the data set.
Does not provide any insights regarding an “Input→Output” process.

DEA: Data Envelopment Analysis
Allows for calculating the relative efficiency scores of decision-making units (DMUs), as well as changes in the scores over time via using the Malmquist Index (MI) scores.
A “black box” model of the “Input→Output” conversion process. Does not offer insights into the sources of inefficiencies.

MR: Multiple Regression
Allows for determining the significance of the impact of independent variables on a dependent variable and identifying the presence of complementarities.
Does not provide any insights regarding an “Input→Output” process and does not allow for considering multiple outputs of the process.

Table 2. Structural Components of the DSS: Offered Insights and Limitations

At this point we are ready to map the selected methods to the steps of each module. We offer our reader to examine each step, along with the intended results of the application of each method, by referring to Table 3 below.

| Module | Step | Method | Expected Outcome/Result |
|--------|------|--------|-------------------------|
| 1 | **Step 2**: Test the assumption of homogeneity of geographic areas of interest by using the available demographic data. | CA | A group of n-clusters of geographic areas that differ in terms of the demographic risk factors. |
| | **Step 3**: Identify the demographic variables that differentiate the geographic areas the most | DTI | A set of demographic factors that are responsible for the differences between the geographic areas. |
| | **Step 4**: Discover naturally occurring associations between demographic risk factors and the level of contagion. | ARM | A set of “if>then” naturally occurring associations that characterize the sample. |
| | **Step 5**: Test for the presence of the impact of the demographic risk factors and the spread of the disease. | MR | Determination of the significance/presence of the impact between “if” and “then” parts of associations discovered in Step 4. |
| 2 | **Step 6**: Identify groups of geographic areas based on the effectiveness of containing the spread of the disease. | CA | A group of n-clusters of geographic areas that differ in terms of the contagion-specific factors. |
| | **Step 8**: Identify the factors differentiating the areas discerned as a result of Step 6. | DTI | A set of contagion-specific factors that are responsible for the differences between the geographic areas identified in Step 8. |
| | **Step 9**: Assess a relative efficiency of containing the spread of the disease of a geographic area vis-à-vis other areas. | DEA | A set of scores of relative efficiency for each area, as well as for each cluster that was identified in Step 6. |
| | **Step 10**: Assess the changes, over time, in the relative efficiency of each geographic area in regard to dealing with the contagion. | DEA MI | Determination of the improvement, or deterioration of performance of each area in fighting the outbreak via Malmquist Index (MI) scores. |
| | **Step 11**: Identify the drivers of the change in performance of a geographic area in fighting the disease. | DEA MI | Determination of the reasons for the improvement/deterioration in performance of each geographic area |
Table 3. Modular Design of the DSS: Methodological Steps

At this point we are ready to outline the sequence of the methodological steps, for each module, as well as corresponding data flows in a pictorial format (see Figure 1). Despite that our system is intended to be comprised into a coherent whole, we present our DSS as a collection of two loosely coupled modules, so our readers could evaluate each module independently. Also, we invite our readers to consider a suitability of using other methods of data analysis than those selected by us (e.g., to replace DEA with Free Disposal Hull (FDH), to substitute MR with multivariate adaptive regression splines (MARS), etc.). The five techniques of data mining and data analysis that we use in the design of the proposed DSS have been widely utilized in IS research and practice in a stand-alone fashion. However, they are also very frequently used in combination to construct multi-method methodologies.

Module 1: Projection of the Path

Demographic Data → CA → DTI → ARM → MR → Pandemic Path Projection

Module 2: Assessment of the Containment

Pandemic Data → CA → DTI → DEA → MI DEA → Recommend Improvement

Figure 1. Design of the Proposed DSS

For example, DEA is widely employed for the purpose of evaluating productivity and performance (e.g. Khouja, 1995; Shao & Lin, 2001; Samoilenko & Green, 2008; Bollou & Ngwenyama, 2008; Yu & Lin, 2008; Avkiran & Rowlands, 2008; Kao & Hung, 2009; Du et al., 2010; Lozano-Vivas & Pastor, 2010; Tsolas, Charles, & Gherman, 2020), but it has also been used to complement other data analytic techniques: cluster analysis (e.g. Shin & Sohn, 2004; Hirshberg & Lye, 2001; Lemos et al., 2005; Morais & Camanho, 2011), neural network induction (e.g. Samoilenko & Osei-Bryson, 2008a; Celebi & Bayraktar, 2008; Emrouznejad & Shale, 2009; Mostafa, 2009; Wu, 2009), decision tree induction (e.g. Samoilenko & Osei-
Now, once a completed design of the system has been presented to our reader, we are ready to test the proposed DSS in action using relevant real world data.

4. Descriptions of the Illustrative Datasets

The context of the testing of the DSS is the United States, consequently, the required for the first module data were obtained from the United States Census Bureau (https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/). The topic of interest is demographic data, which is available by selecting “Demographic Characteristics” option (https://data.census.gov/cedsci/table?d=ACS%205-Year%20Estimates%20Data%20Profiles&table=DP05&tid=ACSDP5Y2018.DP05). The latest available year for American Community Survey’ demographic and housing 5-year estimates is 2018, consequently, this year was selected. The data set was augmented by adding a variable “state’ Population Density” (https://state.1keydata.com/state-population-density.php), because population density is an important factor impacting the spread of a disease. We reduced the data set by selecting only those variables that are considered to be associated with Covid-19 risk factors (https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/index.html). Overall, we ended up with 50 variables (Table 4), available for 33 states (Table 5).

Table 4. Census Data - Variables Used in Module 1

| Variable/Code       | Description                                           |
|---------------------|-------------------------------------------------------|
| PopDensity          | Population Density                                    |
| S2601C_C01_010E     | Total Population, 55 to 64 years                     |
| S2601C_C01_011E     | Total Population, 65 to 74 years                     |
| S2601C_C01_012E     | Total Population, 75 to 84 years                     |
| S2601C_C01_013E     | Total Population, 85 years +                         |
| S2601C_C01_017E     | Total Population, 65 years +                         |
| S2601C_C01_018E     | Total Population, 65 years +, Male                   |
| S2601C_C01_019E     | Total Population, 65 years +, Female                 |
| S2601C_C01_020E     | Total Population, Median age (years)                 |
| S2601C_C01_023E     | Total Population, Black or African American          |
| S2601C_C01_034E     | Total population, 15 years +, Widowed                |
| S2601C_C01_035E     | Total population, 15 years +, Divorced               |
| S2601C_C01_043E     | Total population, 25 years +, Bachelor's degree or higher |
| S2601C_C01_047E     | Total population With a disability                   |
| S2601C_C01_051E     | Total Population 18 to 64 years With a disability   |
| S2601C_C01_054E     | Total population 65 years + With a disability       |
| S2601C_C01_087E     | Total population, 16 years +, Unemployed             |
| S2601C_C01_088E     | Total population 16 years +, Unemployed, Percent of civilian labor force |
| S2601C_C01_093E     | Total population 16 years +, Service occupations     |
| S2601C_C01_106E     | Total population, poverty rate, All people          |
Table 5. List of States Used in the Study

Alabama, Arizona, Arkansas, California, Colorado, Connecticut, Florida, Georgia, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, New Jersey, New York, North Carolina, Ohio, Oklahoma, Oregon, Pennsylvania, South Carolina, Tennessee, Texas, Virginia, Washington, Wisconsin

The second module requires a disease-specific data, and Covid-19 data were obtained from COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19). For Steps 6-9 we selected the latest data available at the time of this stage of the study (07-15-2020), and for Steps 10 and 11 we used the data for April 12, May 13, June 14, July 15, and August 13 (data for August became available later in the study). Overall, this resulted in 4 time-periods of April-May, May-June, June-July, and July-August. The selected variables described in Table 6.

Table 6. Pandemic Data - Variables Used in Module 2

| Identifier | Description |
|------------|-------------|
| S2601C_C01_107E | Total population, poverty rate, 18 years + |
| S2601C_C01_108E | Total population, poverty rate, 18 to 64 years |
| S2601C_C01_109E | Total population, poverty rate, 65 years + |
| S2601C_C02_009E | Total group quarters population, 45 to 54 years |
| S2601C_C02_010E | Total group quarters population, 55 to 64 years |
| S2601C_C02_011E | Total group quarters population, 65 to 74 years |
| S2601C_C02_012E | Total group quarters population, 75 to 84 years |
| S2601C_C02_013E | Total group quarters population, 85 years + |
| S2601C_C02_017E | Total group quarters population, 65 years + |
| S2601C_C02_018E | Total group quarters population, 65 years +, Male |
| S2601C_C02_019E | Total group quarters population, 65 years +, Female |
| S2601C_C02_023E | Total group quarters population, Black or African American |
| S2601C_C02_034E | Total group quarters population, 15 years +, Widowed |
| S2601C_C02_035E | Total group quarters population, 15 years +, Divorced |
| S2601C_C02_043E | Total group quarters population, 25 years +, Bachelor's degree or higher |
| S2601C_C02_047E | Total group quarters population, With a disability |
| S2601C_C02_051E | Total group quarters population, 18 to 64 years, With a disability |
| S2601C_C02_052E | Total group quarters population, 18 to 64 years, No disability |
| S2601C_C02_053E | Total group quarters population, Disability Status, 65 years + |
| S2601C_C02_054E | Total group quarters population, 65 years +, With a disability |
| S2601C_C02_087E | Total group quarters population, 16 years +, Unemployed |
| S2601C_C02_088E | Total group quarters population, 16 years +, Unemployed, % of the labor force |
| S2601C_C02_090E | Total group quarters population, 16 years +, Not in labor force |
| S2601C_C02_093E | Total group quarters population, 16 years +, Service occupations |
| S2601C_C02_094E | Total group quarters population, 16 years +, Sales and office occupations |
| S2601C_C02_105E | Total group quarters population, Individuals With Food Stamp/SNAP benefits |
| S2601C_C02_106E | Total group quarters population, Poverty Status is Determined, All people |
| S2601C_C02_107E | Total group quarters population, Poverty Status is Determined, 18 years + |
| S2601C_C02_108E | Total group quarters population, Poverty Status is Determined, 18 to 64 years |
| S2601C_C02_109E | Total group quarters population, Poverty Status is Determined, 65 years + |

NB: Name of the State is also included as the ID variable
| Variable          | Description                                                                 |
|------------------|-----------------------------------------------------------------------------|
| Confirmed        | Aggregated confirmed case count for the state.                              |
| Deaths           | Aggregated Death case count for the state.                                  |
| Active           | Aggregated confirmed cases that have not been resolved.                    |
| Incidence Rate   | Confirmed cases per 100,000 persons.                                        |
| People Tested    | Total number of people who have been tested.                               |
| Mortality Rate   | Number recorded deaths / Number confirmed cases.                            |
| Testing Rate     | Total number of people tested per 100,000 persons.                         |

NB: *Name of the State* is also included as the ID variable.

Additionally, we used two variables to serve as a proxy for the level of medical resources available to each state: *Total Active Patient Care Physicians, Rate per 100,000* (data were obtained from [2019 State Physician Workforce Data Report](https://www.aamc.org/data-reports/workforce/report/state-physician-workforce-data-report) published by Association of American Medical Colleges, available online at https://www.aamc.org/data-reports/workforce/report/state-physician-workforce-data-report) and *Number of Hospitals* (data were obtained from [American Hospital Directory](https://www.ahd.com/state_statistics.html)). Once the data sets were compiled, we tested the DSS in action-the results are described in the next section.

### 5. Testing the DSS - Results of the Data Analysis

We present the results following the format that we used in describing the DSS- as a sequence of two modules, and the sequence of the steps within each module.

**Module 1: Projection of the Path**

**Step 1: Obtain demographic data**

We used the input data set described in Table 4.

**Step 2: Cluster Analysis (CA)**

We used the RGui ([R Core Team](https://www.R-project.org/)) data analysis to perform cluster analysis. Prior to performing CA, we standardized the imported dataset using *scale* function available in the package. Then we ran *k-means* CA with \( k=3 \). The choice of \( k \) allowed us to partition the sample in terms of “High”, “Mid”, and “Low” risk areas- the determination of the label was based on the averaging of the rankings (in terms of the confirmed cases) of the members of each cluster (see Table 7.1b). The results are presented in Table 7.1a below.

| Cluster | States                                                                 |
|---------|------------------------------------------------------------------------|
| 1       | Arizona, California, Florida, Georgia, Illinois, Indiana, Michigan, New York, North Carolina, Ohio, Oregon, Pennsylvania, Texas |
| 2       | Colorado, Connecticut, Iowa, Kansas, Maryland, Massachusetts, Minnesota, New Jersey, Virginia, Washington, Wisconsin       |
| 3       | Alabama, Arkansas, Kentucky, Louisiana, Mississippi, Missouri, Oklahoma, South Carolina, Tennessee                          |

Table 7.1a. Module 1 - Cluster Membership
Once we got the memberships of each cluster, we could use it to calculate the average ranking of each cluster based on the nation-wide ranking of each state based on the *Number of Confirmed Cases* in descending sequence. The calculation demonstrates that average rank of Cluster 1 is higher than that of Cluster 2, and significantly higher than that of Cluster 3 (see Table 7.1b below). Interestingly, the 10 states comprising Cluster 1 (with the exception of Arizona, North Carolina and Oregon) are in the top 16 states in terms of the *Number of Confirmed Cases*.

| Cluster | Cluster 1 | Cluster 2 | Cluster 3 |
|---------|-----------|-----------|-----------|
| Average rank | 12.54 | 16.09 | 28.22 |
| Assigned label | High Risk | Mid Risk | Low Risk |
| Cluster size | 13 | 11 | 9 |

Table 7.1b. Module 1 – Summary Descriptions of the Clusters

**Step 3: Decision Tree Induction (DTI)**

By running decision tree analysis in RGui (R Core Team, 2020), we were able to identify two top-split variables that differentiate the clusters the most. After that we removed the two variables from the data set and re-ran the analysis with the purpose of identifying other variables that may play the role in differentiating the clusters. Results are presented in Table 7.1c and Figure 2 below.

| DTI Iteration | Selected Splitting Variables |
|---------------|------------------------------|
| 1             | • Total population for whom poverty status is determined (*PopPoverty*)<br>• Total population with a disability (*PopDisability*) |
| 2             | • Population Density (*PopDensity*)<br>• Population 25 years and over with bachelor’s degree or higher (*PopBS*) |

Table 7.1c. Module 1 – Top Variables that differentiate the Clusters

![Figure 2. Visual Representation of the Results of DTI](image)

**Step 4: Association Rule Mining (ARM)**
In order to do the ARM analysis, we needed to augment the data set containing the demographic data with the data reflecting the number of confirmed cases. After that the data set was transformed so a record for every state could be represented as a transaction. We binned the numeric values for each variable into four quartiles—this allowed us to represent the range of values in terms of “Low”, “MidLow”, “MidHigh”, and “High” categories. After that we ran the analysis—the results are summarized in Table 7.1d below.

| Left Side (If)                  | Association (⇒)                  | Right Side (Then)                  |
|--------------------------------|----------------------------------|-----------------------------------|
| High Population Density        | ⇒                                | MidHigh Number of Confirmed Cases |
| High Population Density        | ⇒                                | High Number of Confirmed Cases    |
| Low Population Density         | ⇒                                | Low Mid Number of Confirmed Cases |
| Low Population Density         | ⇒                                | Low Number of Confirmed Cases     |
| MidHigh Age 85+                | ⇒                                | High Number of Confirmed Cases    |
| Low Poverty Level              | ⇒                                | Low Mid Number of Confirmed Cases |
| MidLow Total Population with Disability | ⇒                      | High Number of Confirmed Cases    |
| Low Total Population with Disability | ⇒                  | High Number of Confirmed Cases    |
| High Total Population with Disability | ⇒                | Low Number of Confirmed Cases     |

Table 7.1d. Module 1—Association Rules

**Step 5: Multiple Regression (MR)**

The results of the previous step yielded a number of associations, but ARM does not allow for testing a significance of a causal relationship between the “If” and “Then” side of associations. So, in order to test the significance of the impact we use MR. Based on the results of ARM we created the following model:

\[(Age\ 85+,\ \text{PopDensity},\ \text{PopDisability},\ \text{PopPoverty}) \rightarrow \text{Number of Confirmed Cases}.\]

We did not consider the interaction terms in the model.

The results of the analysis yielded Population Density, Population with Disability, and Population with Poverty to be statistically significant, while Population Age 85+ is not statistically significant. As a result, we removed that variable from the model and ran the following regression:

\[(\text{PopDensity},\ \text{PopDisability},\ \text{PopPoverty}) \rightarrow \text{Number of Confirmed Cases}.\]

The results are summarized in the Table 7.1e below.

| Model Statistics | $R^2$ | Adjusted $R^2$ | F     | Significance F |
|------------------|-------|---------------|-------|----------------|
|                  | 0.390 | 0.327         | 6.190 | 0.002          |

| Variable         | Coefficients | Standard Error | t Stat | P-value |
|------------------|--------------|----------------|--------|---------|
| PopDensity       | 107.935      | 40.872         | 2.641  | 0.013   |
| PopDisability    | -20478.073   | 7514.452       | -2.725 | 0.011   |
| PopPoverty       | 13467.399    | 5916.581       | 2.276  | 0.030   |

Table 7.1e. Module 1—Result of MR Analysis

Interestingly, the coefficient of Population with Disability is negative, which means that the states with a greater number of people with disability have a smaller number of confirmed cases. This may, possible, have to do with the limited mobility that some of the disabled people have, where the limited mobility
implies limited exposure to others and, thus, constraining impact on the spread of the disease. At this point we have identified two variables- *Population Density* and *Population Poverty* - that have an impact on the spread of the disease (proxied by the *Number of Confirmed Cases*), and, consequently, we are in a good position to construct a projection of the path of a pandemic.

**Pandemic Path Projection - Prioritizing Intervention Measures**

In projecting a path of a pandemic, knowing that *Population Density* and *Population Poverty* impact the *Number of Confirmed Cases*, we need to consider how actionable the information is. For example, while we can fairly quickly impact *Population Density* via social distancing, quarantines, and other restrictions, we cannot equally quickly impact *Population Poverty*. Consequently, let us consider only one identified variable, *Population Density*, as a predictor of the path of the spread of the disease. Based on the ranking of the states in regard to the population density, we could create a 3-tier projection system, where Tier 1 signified top 33% of the sample in terms of the population density, Tier 2 is comprised of the middle third, and Tier 3 is represented by the states in the bottom third of the sample. The results are summarized in Table 7.1e below.

| Tier 1’ States - High Priority | Tier 2’ States - Mid Priority | Tier 3’ States - Low Priority |
|-------------------------------|-------------------------------|-----------------------------|
| New Jersey, Massachusetts,   | North Carolina, Indiana,     | Alabama, Missouri, Minnesota, |
| Connecticut, Maryland, New   | Georgia, Michigan, South     | Arizona, Mississippi, Arkansas, |
| York, Florida, Ohio,         | Carolina, Tennessee, Kentucky | Oklahoma, Iowa, Colorado,    |
| Pennsylvania, California,    | Washington, Texas, Wisconsin,| Oregon, Kansas             |
| Illinois, Virginia           | Louisiana                     |                             |

**Table 7.1e. Module 1 – Priority-based Groupings of the States**

The accuracy of the projection is easy to test, for the ranking of the states in terms of the number of cases is available - 11 of Tier 1’ states are in the top-13 states with respect to *Number of Confirmed Cases* (top-14 if counting Louisiana, but this state is not included in our data set).

**Module 2: Assessment of the Containment**

**Step 6: CA**

In this step we perform k-means CA using the data set described in Table 6 (i.e. Pandemic). The results are presented in Table 7.2a below.

| Cluster | States                                      |
|---------|---------------------------------------------|
| 1 (n=13) | Colorado, Indiana, Kansas, Kentucky, Michigan, Minnesota, Missouri, Ohio, Oklahoma, Oregon, Pennsylvania, Washington, Wisconsin |
Step 7: Comparison of Clusters

In this step we compare the results of the CA of Module 1 with the results of the previous step (i.e. CA of Module 2). The purpose of this comparison is to see how closely the contents of the clusters match- in a perfect world with the absence of inefficiencies, we would expect that in both cases the clusters would be comprised of the same states. Meaning, a cluster with a higher level of demographic risk factors would also be a cluster with a greater level of the spread of the disease (Table 7.2b), and vice versa.

| Cluster | Cluster 1 | Cluster 2 | Cluster 3 |
|---------|-----------|-----------|-----------|
| Average Rank | 23.00 | 20.50 | 6.13 |
| Assigned Label | Low Prevalence | Mid Prevalence | High Prevalence |
| Cluster Size | 13 | 12 | 8 |

Table 7.2b. Module 2 – Summary Descriptions of the Clusters

As previously, we used the ranking of the states in terms of the Number of Confirmed Cases to determine the averaged rank for each cluster. The determination of the label (i.e. High, Mid, Low) was based on the averaged rank.

Overall, we found out that 14 states mapped to the cluster perfectly- meaning, 42% of the sample exhibited an “expected behavior” (e.g. if they were in the High Risk category in Module 1’ cluster they were also in the High Prevalence category of Module 2’ cluster), 7 states, or 21%, did worse, while 12 states, or 36%, performed better than expected based on the pandemic data. The results are summarized in Table 7.2c.

### Table 7.2c. Module 2 – Comparison of Cluster Memberships

| Confirmed Cases | High Level | Mid Level | Low Level |
|----------------|------------|-----------|-----------|
| Based on demographics | Module 1: Cluster 1 | Module 1: Cluster 2 | Module 1: Cluster 3 |
| Based on actual spread | Module 2: Cluster 3 | Module 2: Cluster 2 | Module 2: Cluster 1 |
| Change in Avg. Ranking | 12.54 → 6.12 | 16.09 → 20.5 | 28.22 → 23 |
| States “As Expected” (n=14, 42% of the sample) | California, Florida, Illinois, New York | Colorado, Kansas, Minnesota, Washington, Wisconsin | Alabama, Arkansas, Mississippi, South Carolina, Tennessee |
| States that “Do Worse” (n=7, 21% of the sample) | From Low to High: Louisiana | From Mid to High: Connecticut, Massachusetts, New Jersey | From Low to Mid: Oklahoma, Missouri, Kentucky |
| States that “Do Better” (n=12, 36% of the sample) | From High to Mid: Ohio, Oregon, Pennsylvania, Indiana, Michigan | From Mid to Low: Virginia, Maryland, Iowa | From High to Low: Arizona, Texas, Georgia, North Carolina |
It is worth noting that eight members of High Prevalence’ Cluster 3 are among the top 11 states (top-12 if counting Louisiana that is not a part of our data set) in terms of the Number of Confirmed Cases.

Step 7: DTI
By performing classification DTI, we can identify the top-level splits responsible for the separation of the sample into the sub-groups. The variables identified were Incidence Rate and Deaths. We present the results in the Table 7.2d below.

| Cluster | Condition |
|---------|-----------|
| Cluster 1 (Low Prevalence - Lower level of spread) | Incidence Rate < 819.9 |
| Cluster 2 (Mid Prevalence -Middle level of spread) | Incidence Rate ≥ 819.9 AND Deaths < 3403.5 |
| Cluster 3 (High Prevalence -Higher level of spread) | Incidence Rate ≥ 819.9 AND Deaths ≥ 3403.5 |

Table 7.2d. Module 2 – Top Variables that differentiate the Module 2 Clusters

The results suggest, specifically in regard to the difference between Cluster 2 and Cluster 3, that there are context-specific inefficiencies that we may want to look at. We’ll do in the next step.

Step 8: DEA
For this step we created and ran three DEA models.

The first model allows us to assess a relative efficiency of each state, as well as the average for each cluster, in “converting” the demographic risk factors into cases. The model that we used is “(Population Density, Population with Disability, Population Poverty) → Number of Cases”. We ran input-oriented DEA (initial conditions are controlled and Number of Cases is to be manipulated) under assumption of variable return to scale. Due to the nature of output-oriented DEA, which aims at maximization of the output, we inverted the output variable (e.g., Number of Cases) by subtracting, for each state, the actual number of cases from 400000. This allows us to run a model that “rewards” decision making units with the smaller, rather than larger, number of cases.

The purpose of second and third DEA models was to test the relative efficiency of the states in terms of utilization of the available medical resources, which were represented by two proxy variables: Total Active Patient Care Physicians, Rate per 100,000 and Number of Hospitals. The second DEA model has as its output variable the Incidence Rate, which is the Number of Confirmed Cases per 100,000 persons. The third DEA model has an output variable Mortality Rate, which is a Number of Recorded Deaths divided by the Number of Confirmed Cases. The results of this analysis are provided in Table 7.2e below.

| DEA Model | Group          |
|-----------|---------------|
|           | High Risk     | Mid Risk | Low Risk |
| Clustering is based on Demographic Risk Factors | Cluster 1 | Cluster 2 | Cluster 3 |
Table 7.2e. Module 2 – Average Relative Efficiency Scores

The results demonstrate the presence of relative inefficiencies for all three clusters, but we would like to also know whether or not performance of the clusters changed over time, and this is the purpose of the next step.

**Step 9: DEA MI**

We used July’s COVID data to run the analysis in the previous step, and in order to investigate changes in the scores of the relative efficiency for each cluster we would have to use the data for multiple time periods. Originally, we considered the following DEA model:

\[(Population\ Density,\ Testing\ Rate,\ Incidence\ Rate) \rightarrow Mortality\ Rate.\]

However, while \(Testing\ Rate\) and \(Incidence\ Rate\) show a low level of correlation with \(Mortality\ Rate\) (0.30 and 0.36, respectively), \(Testing\ Rate\) and \(Incidence\ Rate\) are highly correlated (0.88), thus, we removed \(Testing\ Rate\) from the model. By having data available for April, May, June, July and August we were able to construct 4 time periods (April-May, May-June, June-July, July-August). Additionally, because DEA rewards DMUs with higher levels of outputs per given level of inputs, we needed to convert \(Mortality\ Rate\) in such way, that the states with the lower, and not higher, mortality rates will be rewarded. We converted the output by subtracting the actual reported rate from 10 (the highest original level is 6.10 for Michigan, and the lowest is 1.92 for Tennessee). The summarized results are presented in Table 7.2f below.
It is worth noting that all three risk groups exhibited a significant decline in efficiency (i.e. MI < 1) during the first period of April-May, and for the High Prevalence group the decline was significantly steeper than for the Mid and Low Prevalence groups. However, the consequent three periods have shown a significant improvement in the levels of relative efficiency during May-June and June-July (i.e. MI > 1), followed by a decline in July-August.

Table 7.2f. Module 2 – Average Malmquist Index (MI) Scores

|       |       |       |
|-------|-------|-------|
| June-July | 1.45  | 1.31  | 1.05  |
| July-August | 0.98  | 0.81  | 0.89  |
| Average    | 1.06  | 0.99  | 0.89  |

Table 7.2f. Module 2 – Average Malmquist Index (MI) Scores

It is important to note that a DEA model is not a “true” production model in the sense, let us say, that a recipe is, where \((\text{wheat flour, water}) \rightarrow \text{bread}\). While we can say that flour and water cause bread, and that more flour and more water would cause more bread to be made, we cannot assert, based on DEA alone, that population density and incidence rate cause mortality rate. This is where a decision maker may consider using MR- to investigate if, in fact, the relationships between the inputs and outputs of a DEA model are causal. We illustrate such application of MR to our readers as it applies to our case.

We can assess effectiveness of the states in terms of their fighting the disease by testing the following regression model:

\[(\text{Population Density, Incidence Rate}) \rightarrow \text{Mortality Rate}\].

Table 7.2g. Statistical Analysis of the \((\text{Population Density, Incidence Rate}) \rightarrow \text{Mortality Rate}\) link

Based on the results of MR (see Table 7.2g) we can see that during the first three months of the pandemic it is Incidence Rate that impacts Mortality Rate, but during the later period (e.g., July and August) it is Population Density that has a statistically significant impact on Mortality Rate. One of the possible interpretations is that during the beginning of a pandemic the efforts of medical practitioners should be on getting the number of incidents under control (possibly via increase in testing), while during the developed
stage of a pandemic the efforts should be allocated towards reducing the density of the population (perhaps by implementing quarantines and social distancing measures).

Also, as we indicated above, *Testing Rate* and *Incidence Rate* are highly correlated - this, however, does not imply the presence of causal relationships. But it is the knowledge of the presence of causal relationships that helps a decision maker in fighting the spread of a disease. For example, it is important to pose, and to answer, the following questions:

- Does an increase in *Testing Rate* result in a greater *Incidence Rate*?
- Does an increase in *Incidence Rate* results in an increase in *Mortality Rate*?

Consequently, it is of interest to investigate the following sequence of causal links:

*Testing Rate* → *Incidence Rate* → *Mortality Rate*

We do so via 2 stage OLS, where the first model explores the *Testing Rate* → *Incidence Rate* link (see Table 7.2h), and the second model investigates *Incidence Rate* → *Mortality Rate* link (see Table 7.2i). The reader may observe that for both links, the strength and statistical significance of each link vary depending on the group (or cluster) and the given month. For example, while in April *Testing Rate* has a statistical significant relationship with *Incidence Rate* for all clusters and the complete sample, in August no such corresponding relationship existed for *Low* and *Mid Prevalence* the clusters though it still existed for the *High Prevalence* cluster and the complete sample.

| April       | Module 2 Cluster | R²   | Coefficient | P-value |
|-------------|------------------|------|-------------|---------|
| Low         | 0.9610           | 0.4825 | 0.0000      |
| Mid         | 0.6036           | 0.3996 | 0.0049      |
| High        | 0.9120           | 0.2248 | 0.0001      |
| Complete Sample | 0.7764       | 0.3908 | 0.0000      |
| May         | Low              | 0.1749 | 0.1031      | 0.1549  |
| Mid         | 0.1282           | 0.0667 | 0.2530      |
| High        | 0.7936           | 0.3640 | 0.0029      |
| Complete Sample | 0.7328       | 0.2923 | 0.0000      |
| June        | Low              | 0.0700 | 0.033       | 0.3821  |
| Mid         | 0.1473           | 0.0599 | 0.2180      |
| High        | 0.8542           | 0.2087 | 0.0010      |
| Complete Sample | 0.7318       | 0.1537 | 0.0000      |
| July        | Low              | 0.1036 | 0.0206      | 0.2834  |
| Mid         | 0.0605           | -0.0334 | 0.4408     |
| High        | 0.5774           | 0.0759 | 0.0286      |
| Complete Sample | 0.5318       | 0.0809 | 0.0000      |
| August      | Low              | 0.1357 | 0.0180      | 0.2153  |
| Mid         | 0.0328           | -0.0236 | 0.5726     |
Table 7.2h – Statistical Analysis of Testing Rate → Incidence Rate link

| Month | Module 2 Cluster | R²   | Coefficient | P-value |
|-------|------------------|------|-------------|---------|
| April | Low              | 0.2866 | 0.0022   | 0.0594  |
|       | Mid              | 0.0021 | 0.0002   | 0.8990  |
|       | High             | 0.0484 | 0.0020   | 0.6007  |
|       | Complete Sample  | 0.1321 | 0.0017   | 0.0376  |
| May   | Low              | 0.4692 | 0.0079   | 0.0097  |
|       | Mid              | 0.0224 | 0.0012   | 0.6420  |
|       | High             | 0.4773 | 0.0020   | 0.0577  |
|       | Complete Sample  | 0.3413 | 0.0025   | 0.0004  |
| June  | Low              | 0.4383 | 0.0073   | 0.0136  |
|       | Mid              | 0.3711 | 0.0032   | 0.0354  |
|       | High             | 0.6557 | 0.0027   | 0.0148  |
|       | Complete Sample  | 0.3966 | 0.0029   | 0.0001  |
| July  | Low              | 0.3518 | 0.0078   | 0.0326  |
|       | Mid              | 0.0466 | 0.0008   | 0.5001  |
|       | High             | 0.3641 | 0.0045   | 0.1131  |
|       | Complete Sample  | 0.1471 | 0.0021   | 0.0275  |
| August| Low              | 0.0064 | 0.0007   | 0.7942  |
|       | Mid              | 0.0072 | 0.0001   | 0.7921  |
|       | High             | 0.0744 | -0.0015  | 0.5133  |
|       | Complete Sample  | 0.0027 | 0.0001   | 0.7733  |

Table 7.2i: Statistical Analysis of Incidence Rate → Mortality Rate link

Step 10: DEA MI- sources of change, EC vs TC

The purpose of this step is to inquire into the sources of the change in the averaged scores of the relative efficiency of the three groups of states. By decomposing the overall change in efficiency (MI) into two of its components, EC and TC, we could gain insights into the drivers of change.

The TC component is associated with the increased availability of a technology. This could be, in the case of our study, increased availability of masks, respirators, gowns, ventilators, and other medical equipment. This also could be associated with the increase in the number of the available hospital beds and medical personnel. Simply put, this component signifies growth driven by the increased availability of the resources.

For example, let us consider a scenario of a student who spends 4 hours in front of a computer to study for a test to get a grade of 80. If we give this student a better computer (higher resolution, faster processor, more RAM, etc.), and, as a result, the student receives a grade of 85 after studying for test for 4 hours, then this growth is driven by change in technology.
The EC component is representative of the improved utilization of the available resources. In the context of our inquiry this could be associated with the implementation of new policies, rules, procedures, and practices according to which the available technology (e.g., personnel, beds, ventilators, etc.) is utilized. By referring to the scenario of the student who studies 4 hours to get a grade of 80, if the student receives the grade of 85 all things being equal—study time is 4 hours, and the computer is not changed, then this change is due to the EC component.

Ideally, we would like to see the change in efficiency that is balanced— if a student receives a new computer she improves her score because of the better technology AND because she becomes better at using it. In the case of COVID-19, we would like to see the decreased Mortality Rate being due to the increased availability of the needed resources AND due to increased efficiency of utilization of the additional resources. Tables 7.2j and 7.2k presents data on the EC and TC components and the dominant cause of the change in relative efficiency across pairs of months.

With respect to the demographic risk factors (see Table 7.2j below), for each of the 3 groups, there were improvements in efficiency (i.e. MI > 1) only in the May-June & June-July periods. For these periods, for the High and Mid Prevalence groups the improvements were primarily due to improved technology (i.e. TC > EC), while for the Low Prevalence group the situation is mixed as for the May-June period the improvement can be attributed to the EC (i.e. EC > TC), while for the June-July period the improvements were primarily due to improved technology (i.e. TC > EC). These results indicate that for the High Prevalence group though for the June-July period there was a decrease in EC, the improvement in TC was sufficient to increase in the corresponding overall efficiency (i.e. MI > 1). However, for the July-August period, while there were improvements in EC component there were not sufficient improvements in the TC component to increase in the corresponding overall efficiency.

| Risk Group | Mid | Low |
|------------|-----|-----|
| **Period** | **MI** | **EC** | **TC** | **MI** | **EC** | **TC** | **MI** | **EC** | **TC** |
| **April-May** | 0.33 | 0.60 | 0.59 | 0.42 | 0.60 | 0.65 | 0.47 | 0.74 | 0.66 |
| **May-June** | 1.38 | 1.13 | 1.19 | 1.38 | 1.16 | 1.19 | 1.22 | 1.18 | 1.12 |
| **June-July** | 1.32 | 0.84 | 1.58 | 1.16 | 0.74 | 1.53 | 1.21 | 0.94 | 1.30 |
| **July-August** | 0.82 | **1.30** | 0.67 | 0.98 | **1.40** | 0.73 | 0.85 | **1.00** | 0.82 |
| **Average** | 0.96 | 0.97 | **1.01** | 0.99 | 0.98 | **1.03** | 0.94 | 0.97 | 0.98 |

Table 7.2j. Clustering based on Demographic Risk Factors: Comparison of EC vs TC

With respect to the actual spread of the disease (see Table 7.2k below), for each of the 3 groups, there were improvements in efficiency (i.e. MI > 1) only in the May-June & June-July periods. For these periods, for
the **High** and **Low Prevalence** groups the improvements were primarily due to improved technology (i.e. \( TC > EC \)), while for the **Mid threat** group the situation is mixed as for the **May-June** period the improvement can be attributed to the **EC** (i.e. \( EC > TC \)), while for the **June-July** period the improvements were primarily due to improved technology (\( TC > EC \)). Interestingly it is the **Mid Prevalence** group that shows improvement in **EC** over 3 of the 4 periods, though sometimes not accompanied by sufficient improvements in **TC** (i.e. \( TC < 1 \)). These results indicate that for the **High Prevalence** group though for the June-July period there was a decrease in **EC**, the improvement in **TC** was sufficient to increase in the corresponding overall efficiency (i.e. \( MI > 1 \)). However, for the **July-August** period, while there were improvements in **EC** component there were not sufficient improvements in the **TC** component to increase in the corresponding overall efficiency. Interestingly the **Mid and Low Prevalence** groups showed improvements in the **EC** component in the over each of the last 3 periods.

| Prevalence Group | High          | Mid          | Low          |
|------------------|---------------|--------------|--------------|
| **Period**       | **MI**        | **EC**       | **TC**       | **MI**        | **EC**       | **TC**       |
| April-May        | 0.31          | 0.51         | **0.63**     | 0.41          | **0.69**     | 0.58         | 0.45          | 0.66         | 0.67         |
| May-June         | **1.49**      | 1.19         | **1.24**     | **1.42**      | 1.21         | 1.16         | **1.17**      | 1.02         | **1.13**     |
| June-July        | **1.45**      | 0.86         | **1.67**     | **1.31**      | 1.19         | **1.37**     | **1.05**      | 1.07         | **1.49**     |
| July-August      | 0.98          | **1.55**     | 0.64         | 0.81          | **1.03**     | 0.80         | 0.89          | **1.27**     | 0.72         |
| **Average**      | 1.06          | 1.03         | **1.05**     | 0.99          | **1.03**     | 0.98         | 0.89          | 1.01         | **1.00**     |

Table 7.2k. Clustering based on Actual Spread of the Disease: Comparison of EC vs TC

How could this information, presented in immediate two tables above, be used for a better decision making? Let us consider the guidelines offered by Centers for Disease Control and Prevention for making decisions on the allocation of ventilators to facilities (https://www.cdc.gov/coronavirus/2019-ncov/hcp/ppe-strategy/ventilators.html). The main factors are: 1) Assessment of need; Determination of facilities’ ability to absorb additional ventilators; 2) Ethical considerations to inform how this scarce resource is provided to facilities to save as many lives as possible; 3) Input from state and local leadership, legal and ethical experts, and 4) informed stakeholders. Of especial interest here is the second bullet point - determination of facilities’ ability to absorb additional respirators, where the decision is made based on the following sequence of steps:

- Identify facilities that may have capacity to care for critically ill patients who will need mechanical ventilation (from prior or current assessments).
- Quantify the number of additional ventilators each facility can realistically absorb.
  - Base this estimate on having enough trained and qualified staff, space, and necessary equipment needed for caring for additional patients on mechanical ventilation. (Zaza et al., 2016)
- Determine the population size that each hospital serves and assess the capacity of each facility to serve vulnerable and high-risk populations within this area.
- Consider whether each hospital serves as a referral hospital/regional hospital or serves a high-density population area, rural area, or underserved populations (ibid).

It is easy to see that the decision regarding the estimated quantity of the allocated respirators is based on the available context-specific resources, such as number of trained and qualified staff, space, and supporting equipment. However, as our results suggest, an additional factor to consider is the relative efficiency of the local context to utilize the requested resource- ventilators in this case. It is only expected, that the local context would demonstrate the increase in the $TC$ component of the overall change in efficiency is as a result of the allocation of additional ventilators. However, if we want to obtain a balanced growth in efficiency, then the local context must also exhibit adequate corresponding change in the $EC$ component, which cannot be automatically expected. Instead, it is quite possible (based on the example of the High Prevalence cluster during the June-July period) that the increase via change in technology (i.e. $TC > 1$) would be corresponded by the decrease via change in efficiency of utilization of the technology (i.e. $EC < 1$).

The important aspect of the shown above decision making is an implicit assumption of a constant return to scale (CRS). This is because “having enough trained and qualified staff, space, and necessary equipment” presumes, pretty much, a particular ratio that should exist in order to accommodate the currently present, as well as incoming, resource (e.g., ventilators). However, any socio-technical environment, firms, businesses, hospitals, schools, etc. are not perfectly scalable. Thus, we could expect the changes to the ratio specifying the requirements for a particular equipment, not only ventilators.

The DEA-based approach presented above allows for considering not only perfectly scalable constant return to scale (CRS), but also a more suitable variable return to scale (VRS). One of the benefits of such consideration is a more flexible allocation of the needed resources, distribution of which, at this point, has been noted to be uneven (Livingston et al., 2020).

Furthermore, it is worth noting that the capability of assessing the $EC$ and $TC$ components allows for the explicit consideration of the human component of the utilization of resources. This is important, because according to CDC human factors (e.g., respiratory therapists, staff operating the ventilators) serve as a bottleneck(https://www.nationalgeographic.com/science/2020/03/us-america-has-fraction-medical-supplies-it-needs-to-combat-coronavirus/) in applying physical resources to treatment of a disease. If we apply our DSS at the hospital level, then, knowing that there are approximately 20 technicians capable of operating ventilators per hospital (https://www.bls.gov/ooh/healthcare/respiratory-therapists.htm), we
should be able to obtain a context-sensitive representation of the relative efficiency of utilization of supplies and equipment allowing for more appropriate allocation of what is needed.

6. Conclusion

Delivery of a modern healthcare is an increasingly multidimensional undertaking that requires optimization of the provision of health services, be it in the context of emergency departments (Cabrera et al., 2011), or prescription of medicine (Sintchenko et al., 2008), or managing a Covid-19 pandemic (Mora et al., 2021). This increase in dimensionality of the problem results in the evolution of DSS used in healthcare in terms of their complexity (Safwan et al., 2016), thus requiring their reliance on increasingly more complex components (e.g., data mining and machine learning) (Shailaja et al., 2018). At this point, it has been reported that applications of DSS to optimize provision of healthcare in the context of Covid-19 was very limited and without incorporating such fundamental aspects as feasibility and health system considerations and consideration of the advanced methods of data analysis (Mora et al., 2021). For example, there were efforts to contribute via creating DSS targeting physical distancing (Adam et al., 2021) and available food supply (Blackmon et al., 2021). And while it might be possible that DSS developed for a different purpose and a different industry (e.g., bankruptcy prediction) could be adapted to address some pandemic related issues (Perboli & Arabnejad, 2021), it can also be advantageous to develop a target-specific healthcare DSS (Sutton et al., 2020) for dealing with the spread of the infectious disease. And this is the route we followed in our work.

Despite the emergence of DSS as one of the premier exemplars of IT in healthcare, it is primarily due to the complexity of the systems that their use is not widespread (Rajalakshmi et al., 2011, Wasylewicz & Scheepers-Hoeks, 2019). It is difficult to reconcile the desire for simplicity of DSS in healthcare with the call for the increase in their functionality, where the normal targets (e.g., quality, risk, productivity, etc.) are to be supplemented with pattern recognition and proactive decision making (Kohli & Piontek, 2008), but this is exactly what we attempted to do in this investigation, for the benefits are worth the effort (Latif et al., 2020).

A crucial component to a success of a DSS in fighting a pandemic is the data and their source, in our case reflecting Covid-19 data (Guidotti & Ardia, 2020, Wang et al., 2020, Chen et al., 2020). There are multiple global data repositories available (e.g., IHME (http://www.healthdata.org/covid/data-downloads), LANL-GR (https://covid-19.bsvgateway.org/), USC SIKJalpha (https://github.com/scc-usc/ReCOVER-COVID-19), Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) (https://github.com/CSSEGISandData/COVID-19), etc.) but the main focus of several investigators using the data was on forecasts and alternative scenarios of Covid-19 mortality as being critical inputs in fighting the pandemic (Friedman, 2021). Khan et al. (2021) noted that the applications of
AI in the fight against Covid-19 could be categorized as focusing on diagnosis, screening, prediction, and drug repurposing. However, none of previous Covid-19 research explicitly considered the issue of efficiency.

In this paper we presented and illustrated a conceptual model for an easy-to-construct modular DSS that would be useful for addressing a variety of public health questions related to the occurrence of a ‘new’ disease. Our research can be considered to be complementary to the previous Covid-19 studies, for it deals with identifying a preferential route of the spread of the disease, and with the assessment of the efficiency of utilization of the available healthcare resources once the disease penetrated a new location. Benefits of using our DSS conceptual model include:

- Provides for discovering naturally occurring groups based on the demographic risk factors
- Provides for identifying the sources of heterogeneity between the groups
- For each time period, provides for:
  - Estimating the relative efficiency of each group with for example the disease containment
  - Determining whether there are improvements in efficiency
  - Identifying whether the dominant cause (better technology, or better processes including better utilization of technology) for changes in relative efficiency in each group
- Provides for uncovering sample and group-specific non-obvious causal structures
- Identifying causal impact of the risk factors on the contraction of the disease
- Identification of High, Mid, and Low risk naturally occurring groups based on the efficiency and effectiveness in containing the spread of the diseases (and minimizing the mortality rate) with a correspondent suggestion for the improvements

REFERENCES
Avkiran, N. and Rowlands, T. (2008). How to better identify the true managerial performance: State of the art using DEA. Omega, 36(2), 317-324.

Babaei, G., & Bamdad, S. (2020). A multi-objective instance-based decision support system for investment recommendation in peer-to-peer lending. Expert Systems with Applications, 150, 113278.

Blackmon, L., et al. (2021). Rapid Development of a Decision Support System to Alleviate Food Insecurity at the Los Angeles Regional Food Bank amid the COVID-19 Pandemic. Production and Operations Management (2021), https://doi.org/10.1111/poms.13365.

Bollou, F. and Ngwenyama, O. (2008). Are ICT investments paying off in Africa? An analysis of total factor productivity in six West African countries from 1995 to 2002. Information Technology for Development, 14(4), 294-307.
Cabrera, E., Taboada, M., Iglesias, M. L., Epelde, F., and Luque, E. (2011). Optimization of Healthcare Emergency Departments by Agent-Based Simulation, Procedia Computer Science, 4, 1880-1889.

Çelebi, D., and Bayraktar, D. (2008). An integrated neural network and data envelopment analysis for supplier evaluation under incomplete information. Expert Systems with Applications, 35(4), 1698-1710.

Chen, E., Lerman, K. and Ferrara, E. (2020). Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set, JMIR Public Health and Surveillance, 6(2): e19273.

Collins, J., Ketter, W., & Gini, M. (2010). Flexible decision support in dynamic inter-organisational networks. European Journal of Information Systems, 19(4), 436-448.

Cooper, W.W., and Tone, K. (1997). Measures of inefficiency in data envelopment analysis and stochastic frontier estimation. European Journal of Operational Research, 99(1), 72-88.

Du, J., Liang, L., Chen, Y., and Bi, G. (2010). DEA-based production planning. Omega, 38(1-2), 105-112.

Dula, J.H., (2002). Data Envelopment Analysis (DEA). Handbook of Applied Optimization, P.M. Pardalos and M.G.C. Resende, (eds.), pp. 531--543, Oxford University Press, New York, 2002.

Eilat, H., Golany, B., and Shtub, A. (2008). R&D project evaluation: An integrated DEA and balanced scorecard approach. Omega, 36(5), 895-912.

Emrouznejad, A., and Shale E. (2009). A combined neural network and DEA for measuring efficiency of large scale datasets. Computers and Industrial Engineering, 56(1), 249-254.

Friedman, J., Liu, P., Troeger, C. E., Carter, A., Reiner, R. C., Barber, R. M., ... & Gakidou, E. (2021). Predictive Performance of International COVID-19 Mortality Forecasting Models. Nature Communications, 12(1), 1-13.

Guidotti, E. and Ardia, D. (2020). COVID-19 Data Hub. Journal of Open Source Software, 5(51), 2376.

Hirschberg, J.G., and Lye, J.N. (2001). Clustering in a data envelopment analysis using bootstrapped efficiency scores. Department of Economics – Working Paper Series 800, The University of Melbourne.

Kao, C. and Hung, H. (2009). Efficiency analysis of university departments: An empirical study. Omega, 36(4), 653-664.

Khan, M., Mehran, M. T., Haq, Z. U., Ullah, Z., & Naqvi, S. R. (2021). Applications of Artificial Intelligence in COVID-19 Pandemic: A Comprehensive Review. Expert Systems with Applications, 115695.

Khouja, M. (1995). The use of data envelopment analysis for technology selection. Computers and Industrial Engineering, 28(1), 123-132.

Kohli R. and Piontek F. (2008). DSS in Healthcare: Advances and Opportunities. In: Handbook on Decision Support Systems 2. International Handbooks Information System. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-48716-6_23.

Latif, S. et al. (2020). Leveraging Data Science to Combat COVID-19: A Comprehensive Review, in IEEE Transactions on Artificial Intelligence, 1 (1), 85-103, Aug. 2020, doi: 10.1109/TAI.2020.3020521.
Lemos, C.A.A., Lins, M.P.E., and Ebecken, N.F.F. (2005). DEA implementation and clustering analysis using the K-means algorithm. In Data Mining VI - Data Mining, Text Mining and their Business Applications, Transactions of the Wessex Institute.

Liu, J. and Lu, W. (2010). DEA and ranking with the network-based approach: a case of R&D performance. Omega, 38(6), 453-464.

Livingston, E., Desai, A., & Berkwits, M. (2020). Sourcing personal protective equipment during the COVID-19 pandemic. JAMA, 323(19), 1912-1914.

Lozano-Vivas, A. and Pastor, J. (2010). Do performance and environmental conditions act as barriers for cross-border banking in Europe? Omega, 38(5), 275-282.

Mora, M., Wang, F., Phillips-Wren, G., and Gomez, J. M. (2021). Evaluating analytics DSS for the COVID-19 pandemic through WHO-INTEGRATE EtD for health policy, Journal of Decision Systems, DOI: 10.1080/12460125.2021.1914292.

Morais P. and Camanho, A. (2011). Evaluation of performance of European cities with the aim to promote quality of life improvements. Omega, 39(4), 398-409.

Mostafa, M.M. (2009). A probabilistic neural network approach for modelling and classifying efficiency of GCC banks. International Journal of Business Performance Management, 11(3), 236-258.

Osei-Bryson, K. M., & Ngwenyama, O. (2014). Advances in Research Methods for Information Systems Research. New York, USA: Springer, 10, 978-1.

Parthasarathy, S., and Anbazhagan, N. (2008). Evaluating ERP projects using DEA and regression analysis. International Journal of Business Information Systems, 3(2), 140-157.

Perboli, G., & Arabnezhad, E. (2021). A Machine Learning-based DSS for Mid and Long-Term Company Crisis Prediction. Expert Systems with Applications, 174, 114758.

R Core Team. A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2017. © 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). https://www.R-project.org/

Rajalakshmi, K., Chandra Mohan, S., and Babu, S. D. (2011). Decision Support System in Healthcare Industry. International Journal of Computer Applications, 26(9), 42-44.

Ramanathan, B. and Yunfeng, J. (2009). Incorporating cost and environmental factors in quality function deployment using data envelopment analysis. Omega, 37(3), 711-723.

Safwan, E. R., Meredith, R. and Burstein, F. (2016) Business Intelligence (BI) system evolution: a case in a healthcare institution, Journal of Decision Systems, 25(1), 463-475.

Samoilenko, S. and Osei-Bryson, K.M (2008a). Strategies for Telecoms to Improve Efficiency in the Production of Revenues: An Empirical Investigation in the Context of Transition Economies. Journal of Global Information Technology Management, 11(4), 56-75.

Samoilenko, S. and Osei-Bryson, K.-M. (2007). Increasing the Discriminatory Power of DEA in the Presence of the Sample Heterogeneity with Cluster Analysis and Decision Trees. Expert Systems with Applications, 34 (2), 1568-1581.
Samoilenko, S. and Osei-Bryson, K.M. (2008b). An Exploration of the Effects of the Interaction between ICT and Labor Force on Economic Growth in Transitional Economies. International Journal of Production Economics, 115(2), 471-481.

Samoilenko, S. V. and Osei-Bryson, K. M. (2017). Creating Theoretical Research Frameworks Using Multiple Methods: Insight from ICT4D Investigations. CRC Press.

Samoilenko, S. and Osei-Bryson, K. M. (2013). Using Data Envelopment Analysis (DEA) for monitoring efficiency-based performance of productivity-driven organizations: Design and implementation of a decision support system. Omega, 41(1), 131-142.

Samoilenko, S., and Green, L. (2008). Convergence and productive efficiency in the context of 18 transition economies: Empirical investigation using DEA. Proceedings of the Southern Association for Information Systems Conference, Richmond, VA, USA March 13-15.

Shailaja, K., Seetharamulu, B., and Jabbar, M. A. (2018). Machine Learning in Healthcare: A Review, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, 910-914, doi: 10.1109/ICECA.2018.8474918.

Shao B.B.M. and Lin W.T. (2001). Measuring the value of information technology in technical efficiency with stochastic production frontiers. Information and Software Technology, 43(7), 447–56.

Shi, J. and Fung, R. Y. (2017). Decision support system for purchasing management of seasonal products: A capital-constrained retailer perspective. Expert Systems with Applications, 80, 171-182.

Shin, H.W. and Sohn S.Y. (2004). Multi-attribute scoring method for mobile telecommunication subscribers. Expert Systems with Applications, 26(3), 363-368.

Sintchenko, V., Coiera, E., and Gilbert, G. (2008). Decision support systems for antibiotic prescribing, Current Opinion in Infectious Diseases, 21(6), 573-579.

Tsolas, I. E., Charles, V., & Gherman, T. (2020). Supporting Better Practice Benchmarking: A DEA-ANN Approach to Bank Branch Performance Assessment. Expert Systems with Applications, 113599.

Wang, L. L., Lo, K., Chandrasekar, Y., Reas, R., Yang, J., Eide, D., Funk, K., et al. (2020). CORD-19: The covid-19 open research dataset. arXiv preprint arXiv:2004.10706.

Wasylewicz, A. and Scheepers-Hoeks, A. (2019). Clinical Decision Support Systems. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science [Internet]. Cham (CH): Springer; 2019. Chapter 11. Available from: https://www.ncbi.nlm.nih.gov/books/NBK543516/ doi: 10.1007/978-3-319-99713-1_11.

Wu, D. (2009). Supplier selection: A hybrid model using DEA, decision tree and neural network. Expert Systems with Applications, 36(5), 9105-9112.

Yu, M. and Lin, E. (2008). Efficiency and effectiveness in railway performance using a multi-activity network DEA model. Omega, 36(6), 1005-1017.

Zaza, S., Koonin, L. M., Ajao, A., Nystrom, S. V., Branson, R., Patel, A., ... & Iademarco, M. F. (2016). A conceptual framework for allocation of federally stockpiled ventilators during large-scale public health emergencies. Health Security, 14(1), 1-6.
Design of a Modular DSS for Public Health Decision-making in the Context of a COVID-19 Pandemic Landscape

Highlights

- Management of a pandemic could be enhanced by a rational public health decision making process
- Decision Support Systems (DSS) can facilitate the exploration of pandemic-related questions
- A public health decision-making process could be formalized and embedded within a modular DSS
- A methodology for designing a modular DSS for public health decision-making is presented
- The modular DSS would include established data analytic methods such a data mining methods and DEA