ISSN 2959-6157

Discover the Factors of Longevity by Causal Discovery with PC Algorithm and Neural Networks

Shengtao Ding[1, *]

[1]Worcester Academy, 81 Providence, Worcester, 01604, MA, USA
*Corresponding author: dingshengtaodst@ldy.edu.rs

Abstract:

The study title has been to study the multilayered factors that govern a human prospect of living through the application of the PC Algorithm (Graph-Based Algorithm), logistic regression analysis, and the neural networks in combination. To realign the research purpose, a vast array of data can be used, which included socioeconomic status information alongside health and environmental aspects carefully chosen for their likely influence on how long people can live. This article will explain the longevity determinants by examining not only the conditions but also the environmental and health indicators and their combinatory effects. The PC algorithm can produce a directed acyclic graph (DAG) to see how elements like healthcare spending, education, and CO_2 emissions are correlated with human longevity and how they influence each other. Moreover, the research takes other examples by analysis: researchers can examine hypothetical events. Neural networks were employed to depict the interrelationship among these variables which subsequently provided a user-appropriate understanding of their sudden joint influence. Consequently, a hypothesize was purposed that the potential interventions for the elongation of human life. Policy-making researchers can gain a better representation of the situation of health inequality and the new recommended solutions to tackle this problem and increase life expectancy by implementing complex analysis tools to blows away the criminal aspects of human longevity.

Keywords: PC algorithm; Neural network; causal discovery; counterfactual inference; structural equation modeling

1. Introduction

Longevity is influenced by genetic, environmental, lifestyle and other factors that shape human life expectancy and the human lifespan. Knowing these cornerstones and their interdependencies is crucial for outlining tactics to increase longevity. The connection between longevity studies and modern statistical and cause-effect research methods for these detailed relationships provides a deeper understanding of the processes behind healthy human aging. One of the most powerful tools in this area of research is the Peter-Clark(PC) algorithm, a constraint-based method designed to uncover causal structures from observational data. This method proceeds by testing for variable independence in a systematic way, which will result in the formation of a Directed Acyclic Graph (DAG) reflecting the reticulation of a causal relation [1]. Such analysis comes forward as a more efficient approach in high-dimensional conditions of life-course analysis of different factors altogether, for the fact that multiple factors should be evaluated simultaneously for the effects on lifespan are observable [2].

In longevity studies, multiple logistic regression might be used. Its main purpose is to calculate the overall risk of several events that may either be attenuated or increased by different risk factors, including age, health behaviors, and socio-economic status, which are present in the data set. However, employing the causal discovery methods, which may be later coupled with logistic regression analysis, can not only establish correlations between variables, but also help not only in establishing cause-effect relations critical for research oriented at finding out the determinants of a person's age [3, 4].

Neural networks, which achieve interdisciplinary status in longevity studies, can model complex, non-linear relationships among various factors. These models can analyze complex patterns contained in big data sets, which are often insensitive to traditional approaches currently implemented. Collision between neural networks, concerned with causality interaction, and causal discovery methods, including algorithms like PC, will result in a broad under-

standing of the contribution mechanisms to limitedness [4].

Causal theory and the Structural Equation Method (SEM) constitute valuable instruments for the examination of such scenarios. The structural equation modeling method, which is among the central significance in longevity research, permits the evaluation of many tied elements within an equation framework and for measuring direct-dependent and indirect dependencies precisely. Hence, a systemic view is attained, as researchers can now visualize the crossovers among different elements and the combined effects on lifelong mortality [5]. Meek's rule is again employed in the PC algorithm background to refine these causal models. Meek's rule additionally supports that the corresponding DAG remains acyclic and in line with the data, ensuring an accurate representation of the causal model. Refining these wrong or exaggerated assumptions and bolstering them by integrating them with the background knowledge will require rigorous reasoning and computing power to fix possible mistakes [6, 7].

Additionally, counterfactual inference allows researchers to explore hypothetical scenarios, such as the likely outcome should one in a given group change his or her diet plan. On the other hand, by altering some factors, the prospective analysis will generate precise predictions on what may affect the existence of the disease and, therefore, aid in proactive undertakings for health improvements. This study aims to integrate these advanced methodologies—covering causal discovery with the PC algorithm and the application of SEM-to create a synthesis life expectancy model. Beyond that, it not only creates a more in-depth comprehension of the genetic and environment-related factors leading to age and diseases but also contributes to designing more targeted strategies to increase life span and improve health quality in the elderly [8, 9].

2. Methods

2.1 Data Source

The data for this study were sourced from the Kaggle dataset titled "Life Expectancy & Socioeconomic (World Bank)." This dataset demonstrates an in-depth outline of the life expectancy at birth and many other variables associated with socioeconomic standpoint and environmental issues dating from 2000 to 2019. Among the key variables include the degradation of health expenditure in terms of GDP, education expenditure compared to GDP, jobless figures, quality of services, the global environmental performance index, etc. The data received are authentic and reliable from well-known sources, such as the World Bank Open Data and Our World in Data.

This dataset helps present the factors for life expectancy globally, especially in the parts of the world where social and economic disparities are more significant, such as Sub-Saharan Africa. The dataset thoroughly examines how various factors like economic condition, social context, and ecological factors contextualize the concept of life expectancy. This research seeks to identify the principal predictors of life expectations and provide recommendations to policymakers for effectively implementing health and social policies.

2.2 Variable Selection and Explanation

This research looks at the factors that affect life expectancy as viewed from an economic, social, and environmental aspect and summarized in Table 1. Health care expenditures accounted for the percentage of GDP are the health boost of development processes, which if well correlated with health of a nation's citizens. Economic growth is very crucial with Education Expenditure as a percentage of GDP, which leads to the improvement and development of human capital. This is because people with a high education level are linked with healthily ways of living.

The Unemployment Rate reflects economic instability and, in turn, limits access to healthcare, which adversely influences one's life expectancy. CO_2 emissions represent environmental risks, and in particular, they may be associated with several health conditions being detrimental to health and well-being, which significantly affects longevity. The Prevalence of Undernourishment measures not only hunger but also poor nutrition, which is a health-affecting factor as well. Sanitation signifies the availability of improved sanitation facilities and is also essential in disease prevention.

It was initially a feasible choice for the Corruption Perception Index, but it can no longer be applied, because the proportion of missing data is over 70%. Other characteristics, consisting of sanitation, education expenditure, undemourishment prevalence, and elevated unemployment, presented missing values from 10% to 40%. The missing values were inferred as the median for shuffled variables and the mean as for mere data wooed. Life expectancy was a less persistent year-to-year variable, making it suitable for the mean imputation.

In short, the provided data had been scrutinized and thoroughly cleaned up to be able to draw reliable conclusions on life expectancy determinants.

Table 1. List of Variables

Variable	Logogram	Range
Life Expectancy World Bank	x_1	Average years a newborn is expected to live. (40-85 years typically)
Prevelance of Undernourishment	x_2	Percentage of the population with insufficient food intake.(0%-100%)
CO_2	x_3	Total carbon dioxide emissions produced by a country.(0 - several billion tons)
Health Expenditure %	x_4	Proportion of GDP spent on healthcare.(0% - 20% typically)
Education Expenditure %	x_5	Proportion of GDP spent on education.(0% - 10% typically)
Unemployment	x_6	Percentage of the labor force that is unemployed.(0% - 100%)
Sanitation	x_7	Percentage of the population using improved sanitation facilities.(0% - 100%).
Longevity	Y	Overall measure of life expectancy in the study.(typically 40-85 years).

2.3 Method Introduction

In this study, a set of data analysis techniques will be pursued to delve into the cores of death rate predictors whose focus will be on economic, social, environmental and health concerns. The practices comprise of descriptive statistical treatment, multiple linear regression, logistic regression, a focus on the causal discovery algorithms (for instance, the PC algorithm), Meek's rule implementation, selection of neural network as well as counterfactual inference. At the start, descriptive statistics were used to outline the variables' fundamental details (such as mean, median, mode), unmask data distribution, and flag any skew or outliers. The said analysis laid down the groundwork for the following steps in modeling. Subsequently, more robust multiple linear regression models were fitted to measure the correlation between the dependent and independent variables, thus, allowing for the estimation of the factors responsible for the difference in life expectancy [10]. Regarding the continuous objects and categorical life expectancy influencing factors, the logistic regression was utilized for significant purposes. These structures are efficacious in the principles of the binary or the multinomial, for instance, classifications such as the decent hygiene, poverty, and their impacts on the lifespan of a specific individual.

The PC algorithm was instrumental in generating a Directed Acyclic Graph (DAG) depicting the causal processes affecting the dependent variables. To ensure the consistence of Meek's rule, such relations were selected, providing the angle accepted in the graph. Neural networks enabled the development of Structural Equation Models (SEMs), that captured how each type of member affected the life expectancy. They help persons understand the interactive outcome between factors. Use of counterfactual inference method encouraged the assessment of theoretical scenarios and the probable impacts on life expectancy reflecting health policies and economic conditions adjustments. This methodology becomes recurrence and resource allocation throughout the policy and the adequacy of dissemination of information. Hence, on the whole, the research is intended to extract some most vital determinants influencing longevity and present scientific evidence which the policy-makers can utilize in a bid to decrease additional health issues and lengthen the life span of the population.

3. Results and Discussion

3.1 Initial Causal Graph

Subsequent graph demonstrates the first PC algorithm output providing the initial impression of the relationship structure between the variables (Figure 1). In this instance, Life Expectancy (World Bank), having no edges pointing outwards (X1), means that this variable is a major determinant of the other variables; however, it is not determined by anything else. The observed effect, thus, serves as the dependent variable determined by available information at hand inside the execution data set.

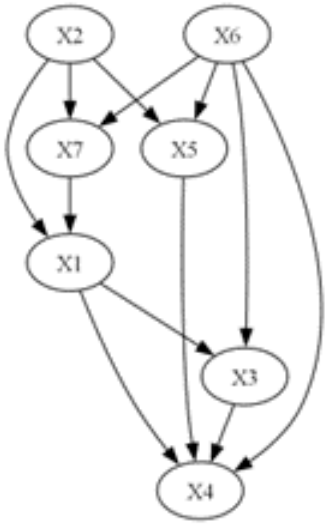

Fig. 1 Initial Causal Graph

3.2 Refined Causal Graph

The original causal graph was improved by holding in mind other hypotheses related to the contemporary environment. It appears that the impact is depicted in Figure 2, which shows the direction of influence between the variables as these edges are introduced. These trends, however, especially with CO_2 emissions (X3), Health Expenditure % (X4), and Education Expenditure % (X5), do not point in the same direction with Unemployment (X6). On the other hand, these findings imply that there is unlikely to be a correlation between these two variables of interest, which further support the theory that the type of unemployment rate found in the economy is often affected mostly by short-run factors rather than long-term socio-economic conditions.

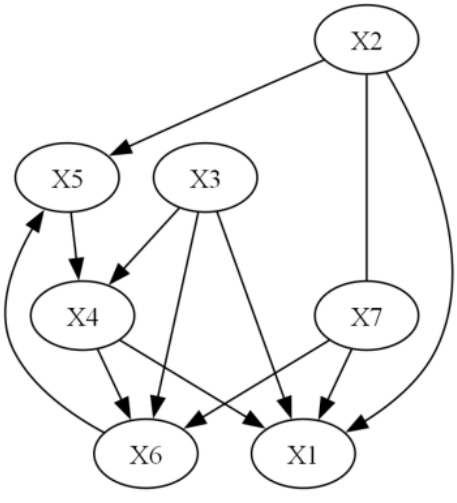

Fig. 2 Enhanced Causal Graph with Background Knowledge

3.3 CO_2 Emissions Impact

Besides, another insight, which can be obtained from analysis, was that education expenditure % (X5), health expenditure % (X4), and CO_2 emissions (X3) do not have a direct bearing on unemployment rates (X6), which was revealed by the absence of directed edges in the graphical causal models (Figure 3). This latter one possibly denotes a setting in which these economic characteristics are not the variables which have direct complicated effects, and the like.

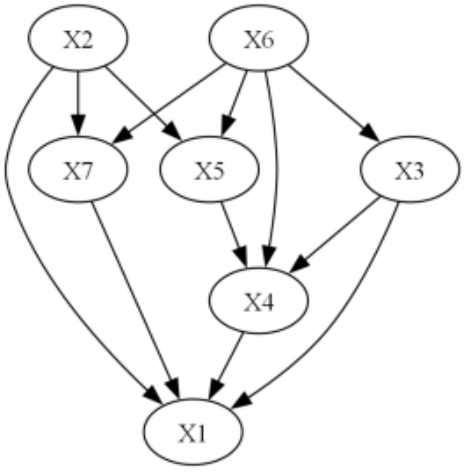

Fig. 3 Specific Influences on CO_2 Emissions

3.4 Final Causal Structure

On one hand, it is disclosed that in the causal graph CO_2 emissions (X3) have no arrows coming in, meaning that within this model they are not affected (causally) by any of the other variables looked at, which are health expenditure, education expenditure, and unemployment. Such absence of incoming edges is pointing at CO_2 emissions, which can be treated as an exogenous factor in the model, meaning that they are driven not by the dataset but by outside forces. This is visioned as the core statement because CO_2 emissions are not determined by the socio-economic variables that can be included in the study, rather the driving forces of global industry, power generation, international economy, and other broader trends (Figure 4). This belief agrees with a notable fact that CO_2 emissions in terms of contextual policy, foreign trade, and technology used in energy production, which are outside the control of the investigated variables mentioned above.

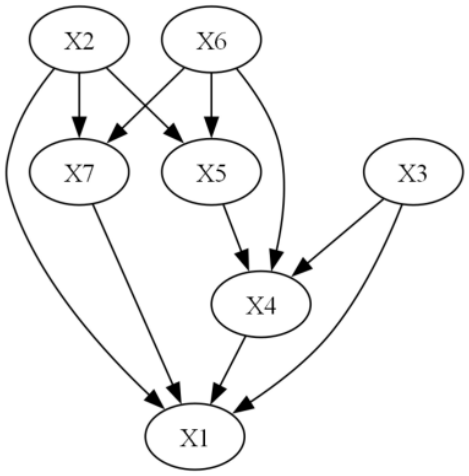

Fig. 4 Final Causal Structure

3.5 Neural network Analysis

This research is aimed at assessing the role of different aspects with the help of neural network algorithms on the life expectancy of people. The neural network model was trained over 20,000 epochs, and loss function value averaged 0.2778, a finding that indicated that the model was fitted well and was capable of producing reliable results. The study found out some important impacts on life expectancy that was quantified in the number of years (Figure 5).

To be more specific, there was a big positive impact on life expectancy resulting from the increasing level of CO_2 emissions, which is likely the presence of omitted variables or model misalignment. On the contrary, a rise in the unemployment rate could be linked to a decrease of around 7.81 years in life expectancy, which proves the lack of financial stability is a key factor affecting health. Higher medical expenditure corresponded to an emphasis on life expectancy by 4.83 years, which highlighted the important role of funding on health care.

By contrast, the severity of undernourishment led to a reduction in life expectancy of 4.48 years. This means that despite the contribution of food insecurity to an extending lifespan, it still does exist. Another significant finding was that education expenditure had a positive impact on life expectancy by increasing it by 2.40 years, indicating that the direct relationship between education and health extends into the future.

Lastly, introduction of improved sanitation instead of increasing average life expectancy of the population by 1.28 years resulted in a decrease. This might be caused by a number of reasons and hence further inquiry should be completed.

These values illustrate the manner in which a corresponding rise of 10% in certain values relates to year adjustments in life expectancy, the unit of measure being years. Therefore, the enormously positive coefficients imply that some sort of omitted variable is at play because the number does not make sense. Unlike this, factors like jobless and undernourished conditions are risk factors for shortening lives.

By calculating the impacts of 10% changes in the values of these variables on life expectancy, the outcomes show that a straightforward view on whether the shifts in the indicators cause big variations in the life expectancy or not. The CO_2 example, which has an extensive sphere of effect on people, reveals that such policies have broader consequences on health.

This study exhibits a detailed picture for which variables play roles at different levels and to different extents in living a longer life. The results will be further integrated and linked under the whole causative framework to reshape the policy proposition with the ultimate goal to improve life expectancy. Note that the neural network results in the accompanying figure describe the fixed average losses and special pieces of the model results.

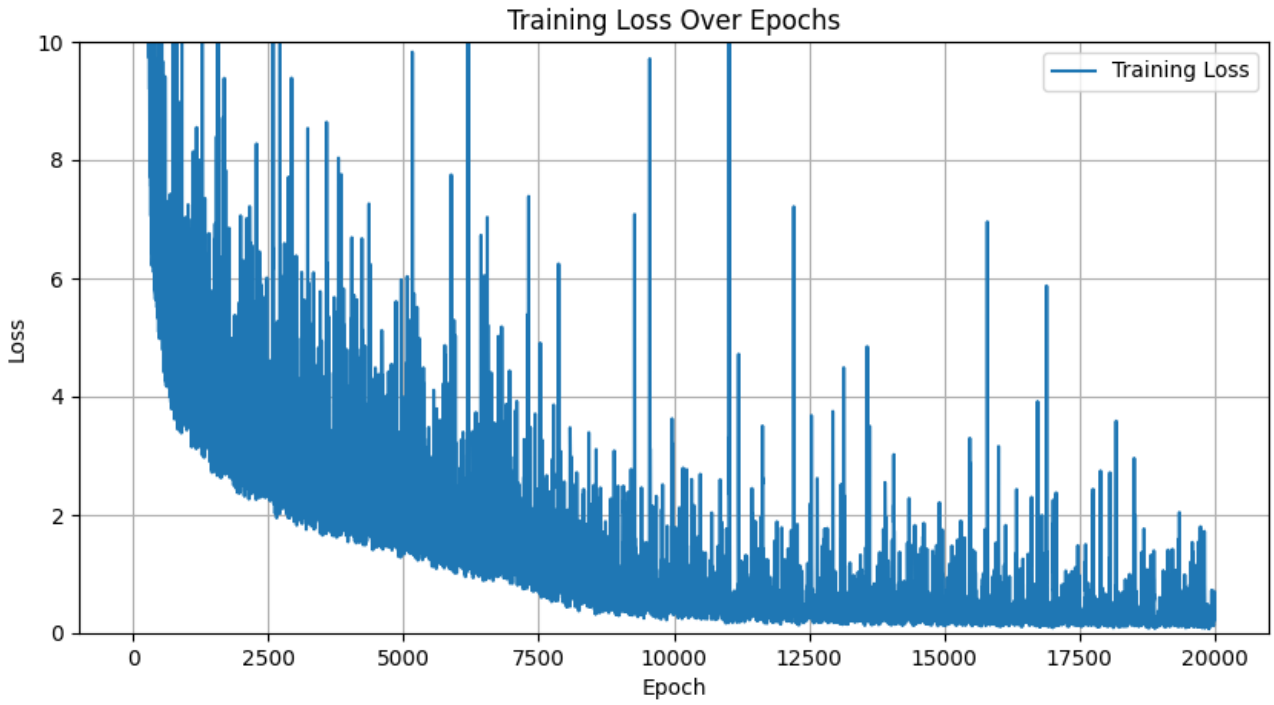

Fig. 5 Training Loss Over Epochs

3.6 Interpretation of the Results

The connectedness of the PC algorithm outputs with the neural network analysis offers the key factors affecting longevity, which are further illustrated with numbers that point out this relationship. The analysis is considered as providing CO_2 life expectancy as the only sole factor affecting it by about 248,720.28 years. The true value would be close to this exaggerated effect. Perhaps, this is because of unaccounted factors or a problem with the model's specifications. In general, it would be logically expected that increasing CO_2 emissions will have either negative or minimum effective, hence shows that environmental factors may act independently or interact with other variables in a very complex manner.

In contrast, health expenditure contributes positively to the magnitude of life expectancy, that is, by 4.83 years, which proves that individual investment in healthcare prolongs life. The increase in education expenditure similarly introduces an additional life expectancy of 2.40 years, demonstrating the long-term benefits that education has on people's health and awareness. Life expectancy, which suggests that unemployment, tends to shorten average individual life by 7.81 years, is another issue that manifests as a consequence of economic downturns. The increase in the incidence of undernourishment by 4.48 years of life expectancy times wider underlines food insecurity as possibly one of the biggest risk factors for early mortality.

On the contrary, it can be detected that with the availability of clean water and sanitation facilities, the life expectancy is decreased by 1.28 years, and the reason may be data issues or interactions with other variables which need to be studied more. Firstly, these results are precise and make clear how the different determinants function on the longevity. The results from both the PC algorithm and neural network analysis reveal these patterns, which are not out of common sense despite some of the knowledge gaps being brought into the focus. Buoyancy is not solely optimistic how CO_2 emissions outweigh the prescribed effect, thinking that it makes the assumption of the model with the other vital factors missing, and this needs to be explored in details. Besides, the positive effect of health spending, education, and employment levels, which are present in science, the negative aspects with reference to under-nutrition and sanitation, however, show the necessity for policy-making directed at health and specifically targeting the challenges that impede life expectancy.

The figures referenced (Graphs 1~4) correspond to the DAGs generated in this study, which visually depict the causal relationships identified by the PC algorithm. To be specific, X1 in the graph represents the 'Life Expectancy World Bank'; X2 in the graph represents the 'Prevelance of Undernourishment'; X3 in the graph represents 'CO_2'; X4 in the graph represents 'Health Expenditure %'; X5 in the graph represents the 'Education Expenditure %'; X6 in the graph represents the 'Unemployment'; X7 in the graph represents the 'Sanitation'.

4. Conclusions

In conclusion, this study successfully reinforces the effectiveness of the utilized causal discovery methods, such as the PC algorithm, in conjunction with the traditional ones to analyze the complex constellation of factors determining longevity. The critical factor of determinants, which consists of factors of health spending and education, environmental conditions, and socioeconomic status, proves the necessity of multicausal inquiry in achieving reasons for lifespan. Synthesizing neural networks and counterfactual reasoning, in this case, broadens the research possibilities akin to applying non-linear regression and exploring possible outcomes based on multiple conditions. They also provide paths for researchers to look into preventive health measures to help people have a longer lifespan. One can adapt health programs to tackle those determinants, bringing about improved health status and longer lives to those at the forefront of the struggles. Therefore, future research needs to develop daily improvements to these models and simultaneously assess other variables that might contribute to the longevity of life. The latter will stimulate the development and improvement of public global health.

Acknowledgments

I would like to express my deepest appreciation to my mentors and advisors, especially Dr. Peter Kempthorne, for his guidance on causal discovery methods and neural networks, which had a significant impact on this research. I also thank my peers for their constructive suggestions and discussions during the research process. Furthermore, I would also like to thank the institutions whom provided the database for this study, like Kaggle and World Bank, which helped a lot when carrying out this research. I cannot forget, finally, my family and friends for the unconditional support they gave me at every stage of the journey.

References

[1] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search (2nd ed.). MIT Press, 2000.

[2] Zhang K, Hyvärinen A. A general framework for discovering causal relations in time series data. Journal of Machine Learning Research, 2018, 19: 1-30.

[3] Pearl J. The Book of Why: The New Science of Cause and Effect. Basic Books, 2018.

[4] Siontis G C M, et al. Logistic regression was as good as machine learning for predicting major chronic diseases. PubMed, 2020.

[5] Bächtiger A, et al. Causal Inference in Statistics: A Primer. MDPI, 2021.

[6] Kosinski M. Why use logistic regression instead of a neural network? Cross Validated. Epidemiology, 2019.

[7] Greenland S, et al. Causal diagrams for epidemiologic research. Epidemiology, 1999, 10(1): 37-48.

[8] Peters J, Janzing D, Schölkopf B. Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press, 2017.

[9] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer, 2019.

[10] Rubin D B. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association, 2005, 100(469): 322-331.