Network Learning on Open Data to aid Policy Making

Pooja Bassin

International Institute of Information Technology Bangalore (IIIT-B), 26/C, Hosur Rd, Electronic City Phase 1, Bengaluru, Karnataka 560100

Abstract

With the increasing proliferation of Big Data, Machine Learning, and Artificial Intelligence, there is increasing interest in designing an AI-based support system for supporting policy formulation and decision-making. We call such a system a Policy Support System or PSS. A PSS aims to characterize a policy based on its targets and indicators to aid policymakers in taking informed decisions not only based on the present state of affairs but also anticipate future scenarios pertaining to policy interventions. To solve numerous problems in social, economic, and environmental domains, Sustainable Development Goals (SDGs) were adopted by the United Nations in 2015 that intend to be achieved by 2030. The proposed PSS focuses on designing a set of Bayesian models to support policy interventions in the area of SDGs as a prototype implementation. Policy formulation is supported by modeling interventions and counter-factual reasoning on the models and assessing their impact on data storytelling. Two kinds of impacts are observed: (a) *downstream* impacts that track expected outcomes from a given intervention, and (b) *lateral* impacts, that provide insights into possible side-effects of any given policy intervention. The research objective is to build causal dependency models for different indicators to understand and analyze minimal-cost policy interventions for achieving the intended targets.

Keywords

policy support, Big Data, Bayesian network, open data

1. Introduction

Sustainable development goals, adopted by the United Nations in 2015, are designed to meet the urgent need of social, economic, political and environmental challenges confronting our world. The Brundtland Commission report,1987 [1], has given the most prevalent definition of sustainable development which has also been adopted by the UN. It states, "Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs."

The public policy think tank of the Government of India, Niti Aayog, measures the progress of SDGs at national and sub-national levels. This has led to the release of *SDG India Index* reports from 2018 onward to track the progress of 306 national SDG indicators [2]. The explosion in the availability of big data and multiple open data initiatives accompanied by artificial intelligence technologies have prompted various government-led measures to utilize data-driven approaches for planning and decision-making around sustainable development goals. Some example initiatives are as follows: National Agricultural Market(eNAM) is an electronic

ER'2022 Forum and Symposium, October 17-20, 2022, Online
pooja.bassin@iiitb.ac.in (P. Bassin)
0000-0002-0611-8734 (P. Bassin)

trading portal for farmers that facilitates live trading and marketing with better trade prices across various mandis of India [3, 4]. Tracking food prices online aids in monitoring food security (SDG 2) situations in real time. To monitor the sustainability of water supply, the Jal Jeevan ministry [5] plans to deploy smart sensors to water pumps to measure relevant aspects of water service delivery [6] (SDG 6). Feedback from such welfare measures taken by the governments prompts policy changes and encourages rolling out reformed policy initiatives.

So far, the steps taken by central and state governments are laudable but remain incomplete as piecemeal measures. The PSS aims to address this issue by creating a generic support system. Whichever policy is chosen, PSS characterizes its targets and indicators to meet its policy agenda. The design of the PSS architecture, Fig. 1, works not only on the chosen policy but seeks to touch upon all the associated components of the policy by implementing intervention modeling with a detailed explanation in section 3.1. In the case of SDGs, if an intervention is performed emphasizing on some aspects of SDG 1 (No poverty), its footprint on SDG 2 (Zero Hunger) can not be neglected. Hence, PSS aims at covering maximal elements related directly or indirectly to the policy under evaluation.

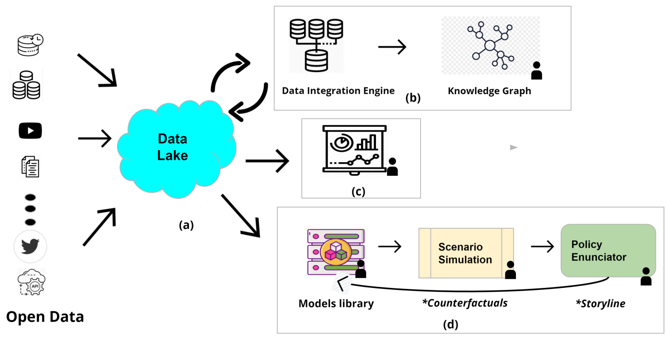

Figure 1: Policy Support System Architecture

Fig. 1 depicts the use of a data lake to collect and store data from varied data sources that further goes on for data cleaning, semantic resolution, canonicalization, analysis etc. [7]. The primary focus of PSS is the creation of *models library* with the aim to perform intervention modeling and counterfactual analysis. This is discussed in detail in section 3.1.

We understand that Decision Support Systems are built to facilitate decision-making within an overarching organizational framework, with elements of problem identification, and generation of reports based on the collection and analysis of data. DSS aids in picking up a decision among a set of alternatives specific to a domain. Organizational models are fairly well-engineered, but public administration needs to contend with complex inter-dependencies for which, well-engineered models may not exist. Building and managing models library is hence an integral part of policy support systems. PSS is also used to support intervention analysis and counterfactual reasoning, as well as in understanding how human beneficiaries react or adapt to a proposed policy. Policy Support Systems are also used in generating “nudges” that encourage behavioral change from the population toward desired outcomes.

2. Related Literature

The abundance of data and advances in AI-enabled technologies have expedited the policy formulation processes. Digital healthcare solutions and maintenance of Electronic Health Records(EHR) [8] , adoption of precision agriculture technologies [9], e-learning initiatives [10] such as Diksha [11], Swayam Prabha [12] are a few examples. The implementation of open data has democratized the system of accountability and transparency in governments and organizations. Increased participation of citizens[13] has resulted in extracting, understanding, and providing diverse inputs and suggestions to governments in public policy planning.

However, these methods are not sufficient when changes in one policy affect the outcomes in another. Hence, this drives the purpose of the creation of the Policy Support System which tracks changes, direct or lateral, across various policy scenarios. The focus of PSS is the development of a models library based on key indicators from policy instrument statements. Bayesian networks are used for answering probabilistic queries and in a policy scenario, they are suitable where we may ask policy-relevant questions to evaluate them in case of uncertainties. Background related to Bayesian networks is provided below because it is the fundamental tool for developing models library for PSS.

A *Bayesian Network* is defined as a directed acyclic graph(DAG) whose nodes represent variables of interest and the links represent informational or causal dependencies among the nodes [14]. Formally, Bayesian networks (BNs) are defined by a directed acyclic graph G = (V, A), where each node $v_i \in V$ corresponds to a random variable X_i a global probability distribution X with parameters θ is factorised into smaller local probability distributions, Θ_{X_i}, according to the arcs $a_{ij} \in A$ present in the graph. The network structure expresses the conditional independence relationships among the variables in the model through graphical separation, thus specifying the factorisation of the global distribution [15] :

$$P(X) = \prod_{i=1}^{N} P(X_i|\Pi_{X_i})$$

where, $\Pi_{X_i} = Parents\ of\ X_i$

Bayesian networks as a tool in the context of public policy decisions are well suited since they can handle subjective and objective data [16] with simplicity and are beneficial in decision-making in case of uncertainties.

3. Policy Support System

We propose to design a *Policy Support System* with the key feature of policy enunciation that aims to characterize a policy based on its targets and indicators to aid policymakers in taking informed decisions not only based on the present state of affairs but anticipate the likelihood of events that may entail. Policies are laid out by governments to address social concerns across multiple dimensions and domains including social, environmental, economic, law and legislature, etc. As a consequence, policy statements are defined in vague and elastic terms[17] to address the needs of the future. They are non-deterministic in nature prompting ever-changing uncertainties. To address uncertainties in the framework of policy implementation, we need to first understand the kinds of uncertainties that may arise and what mechanisms will be helpful in order to minimize the risks of the occurrence of precarious events. In the context of causal inference, [18] have stressed upon two types of uncertainties: factual uncertainty and causal uncertainty. Factual uncertainty may occur when based on the available incomplete information the facts established are speculative whereas causal uncertainty may occur when a variable may possibly be a strong cause of the event but it is not necessarily true in all cases. The Policy Enunciator as part of PSS shall attempt to model causal uncertainties among elements of a chosen policy indicator by employing Bayesian networks. This helps in understanding factors involved in attaining specified targets and indicators.

Bayesian networks are widely studied and implemented in the public policy landscape [19, 20, 21, 22]. Well-established Bayesian network tools [23, 24, 25] are used in the fields of agriculture, aerospace, finance, government, etc. [26]. The primary objective of utilizing Bayesian networks is the *explain-away* effect [27] that it encompasses. For policy-compliant behavior from the people, it is crucial to foster the trust that explains the reasoning behind policy decisions. Bayesian networks are well suited to model public-policy domains where predictions are a matter of livelihood [28]. They represent and reason with uncertainties which leads to the understanding of why and how the predictions were made.

The 17 SDGs are defined as the final overarching conceptual framework[1] adopted by the global leaders at the UN. The goals are further classified into 169 targets tracked by 232 unique indicators.[2] It is thus difficult to build a single Bayesian model for all 17 SDGs. The motivation to develop a library of Bayesian models is to focus on individual indicators of the intended target. This helps to put the focus on the systematic achievement of the targets and measure the progress of the goals. The "Karnataka Statistical Outlook Publication" report[3] is an example of how government compiles data from multiple departments that work largely independently to attain targets and indicators described within their scope.

3.1. Policy Enunciator System

The primary subsystem of the PSS is the *Policy Enunciator Subsystem* (PES). PES is meant for interaction with the end-user who is concerned with policy formulation. The primary design element of the PES is the *language* in which policy-related issues are enunciated.

To prepare a hand-crafted Bayesian network, the first step is to establish *dataframes*.[4] *Dataframes* are essentially divided into 3 categories by identifying independent variables/concepts, dependent variables/concepts, and key indicators from studies and reports relevant to the domain. This helps in setting up dependency arcs between the variables accessed from the PSS data lake. The PSS data lake is a repository of raw data coming from disparate data

[1] Sustainable Development Report 2022 From Crisis to Sustainable Development, the SDGs as Roadmap to 2030 and Beyond: https://www.sdgindex.org/

[2] Measuring progress towards the Sustainable Development Goals https://sdg-tracker.org/

[3] Karnataka at a Glance Department of Economics and Statistics https://kgis.ksrsac.in/kag/Downloads.aspx

[4] Decoding Malnutrition Analysis on Prevalence of Malnutrition in Children across Karnataka, Interim Report, Public Affairs Centre, July 2022

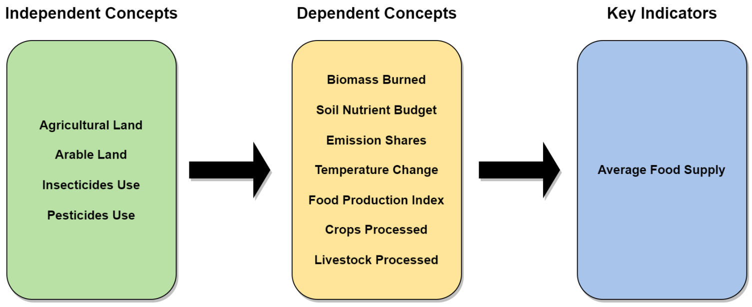

Figure 2: Dataframe Concepts

sources[5,6,7] with different extents of veracity and completeness. Data cleaning, canonicalization, and transformation are implemented in this layer so that the output from this layer is a set of structured and fairly clean datasets. The data lake is queried on the nature of data available from different data sources, their reliability, veracity, etc.

Fig. 2 illustrates "dataframes" from FAOSTAT[8]–the statistical division of UN Food and Agriculture Organisation(FAO). The framework is built by studying factors affecting food supply documented by FAOSTAT.

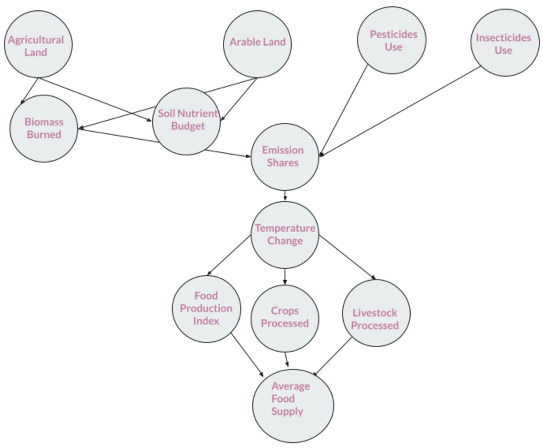

Figure 3: FAO Average Food Supply Model

With the dataframes in place, Fig. 3 shows a Bayesian network model depicting how various

[5] National Family Health Survey(NFHS 5), 2019-21, India Report https://dhsprogram.com/pubs/pdf/FR375/FR375.pdf

[6] Student Achievement Tracking System Department of Primary and Secondary Education, Karnataka https://sts.karnataka.gov.in/SATS/#

[7] Crop Cutting Experiments https://apps.karnataka.gov.in/app/24/en

[8] FAOSTAT: https://www.fao.org/faostat/

factors affect the average food supply in a region. FAOSTAT provides free access to food and agriculture data for over 245 countries for roughly 60 years. The linkages between the nodes in the DAG attempt to establish causal dependencies. The nodes represent elements from significant agriculture domains such as food production, land use, climate change, pesticides, etc. The target variable, average food supply, seeks to identify the food supplies available for human consumption in caloric value.

The Policy Enunciator represents policy related issues using a combination of two elements: *model ensemble* and *data story*. Fig. 4 schematically depicts the policy enunciator components. It comprises a *models ensemble* which is a subset of models from the models library that may have one or more variables in common. These variables act as confounding variables across different concerns. Confounding variables are shown in color and are coupled with their counterpart in another model.

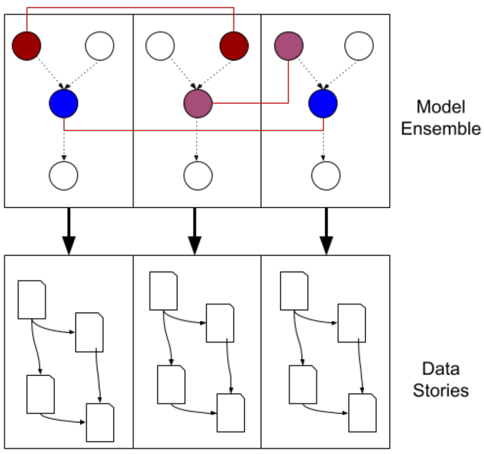

Figure 4: Policy Enunciator: Models Ensemble and Data Stories

A data story is specified by a template, which is populated from data available in the data lake and relevant model outputs. Set of interventions and counterfactual analysis around the target form a data story. Changes to one or more models, result in corresponding changes in data stories. The data-driven stories are constructed with the help of business intelligence tools. Narratives around the associations, intervention outcomes, and what-if analysis for rigorous policy-making are showcased to provide actionable insights to policymakers. A data story is also associated with a theme in the SDG ontology. Thus every proposed intervention can be tracked to different SDG themes that it may likely affect, and detailed using the corresponding data stories.

Fig.5 depicts a use case for the policy enunciation system. Here we consider a policy instrument, the Public Distribution System (PDS).[9] Policy instruments are a set of interventions to bring about change in one or more variables to achieve a set of goals. PDS looks into the distribution of food grains to poorer sections of societies at subsidized prices. Once the key

[9]Public Distribution System: https://dfpd.gov.in/pd-Introduction.htm

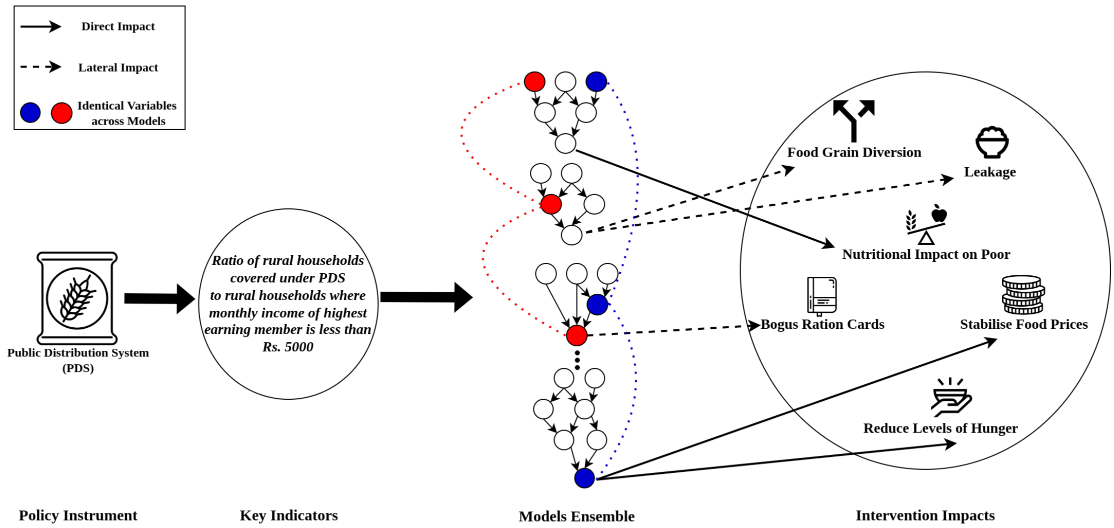

Figure 5: Policy Enunciation with Public Distribution System as a Policy Instrument Use Case

indicator of the given policy is identified, it results in the creation of a library of Bayesian models for the goal. Identical variables are represented using similar colors across models. Intervention in a variable in one model may also impact the outcomes of another model because of the presence of the variable in the model. These impacts may be direct or lateral as shown in the figure. Direct impacts such as stabilizing food prices [29] and improving the situation of hunger[30] may be visible due to policy interventions but may also lead to hidden impacts of food grain diversion to black markets[31] or open market[32], leakage of food grains[33] and creation of duplicate or bogus ration cards[34].

Once the models library is ready, policy enunciation is carried out by performing two kinds of operations: *intervention modeling* and *counterfactual analysis*.

Intervention modeling starts by converting a proposed policy instrument into one or more interventions in a Bayesian model, to set the value of variables to specific levels. For instance, a policy instrument that strives to reduce the use of pesticides in agriculture would affect the model shown in Figure 3, by setting the values of the nodes ”Pesticide Use” and “Insecticide Use” to low levels.

Counterfactual analysis, also called “what-if” analysis also performs similar interventions on models to set values to variables that may not necessarily be visible in the data. For example, taking the model from Figure 3 again, suppose that the values for “temperature change” recorded in the data were only minor changes. The model would have learned conditional probabilities of its effect on other downstream variables based on the data available. This model can then be used to set “temperature change” to a higher value, to see its impact on downstream variables.

In another example, Fig. 6, if we intervene in the number of agriculture loans disbursed, this may subsequently impact the probabilities of synthetic and organic fertilizer consumption, and soil health which in turn may impact the production of rice that hinders or boosts farmers’ income. This kind of impact can be seen within the model itself. In another case, considering Fig.

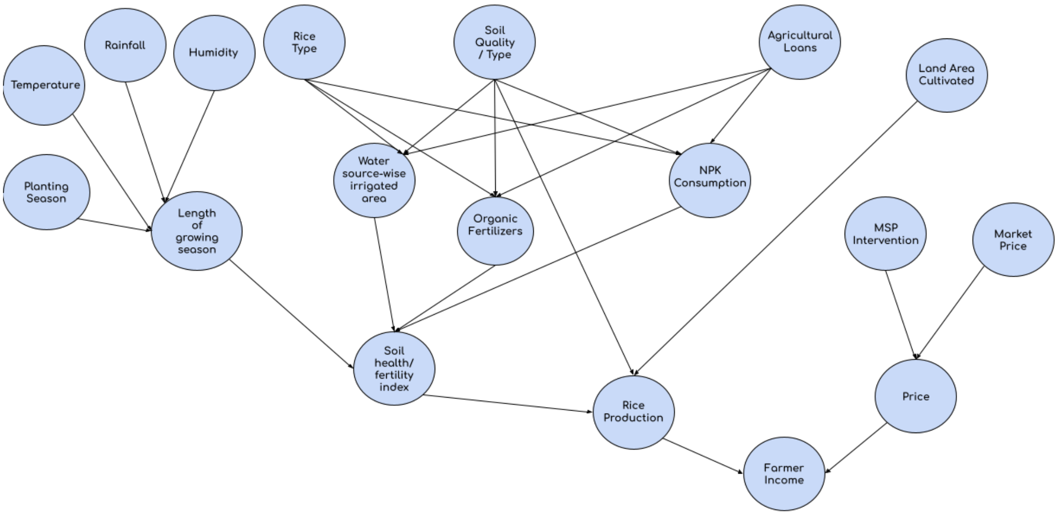

Figure 6: Rice Production Network

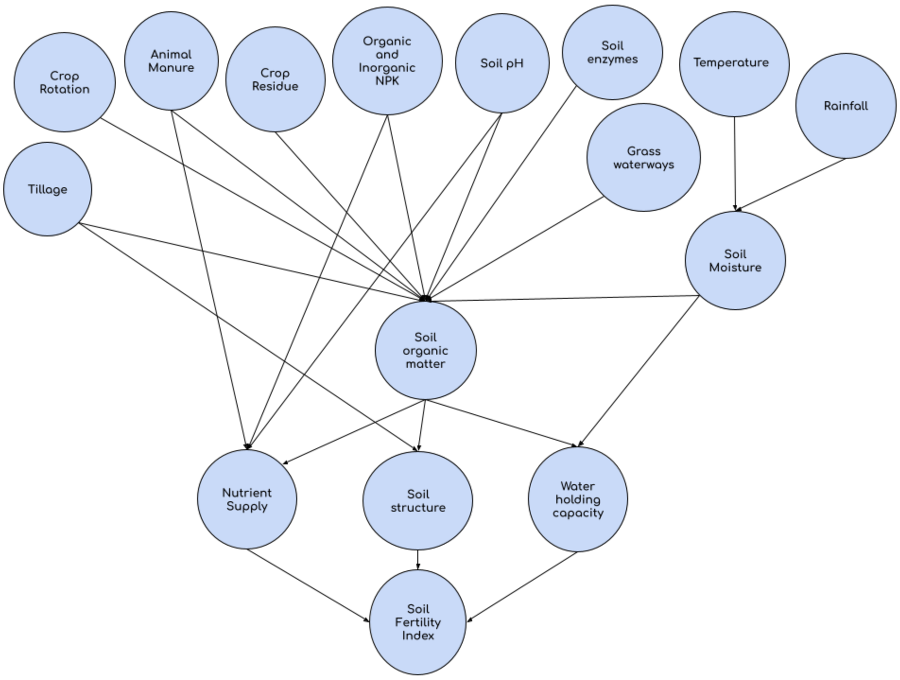

Figure 7: Soil Fertility Index Network

7 as part of our models library, when we intervene in shaping the grassed waterways channel, a downstream impact may be noted in the Soil Fertility Index Network where ultimately the values of the target variable *Soil Fertility Index* may change but this change may have a lateral impact in Fig. 6 as well. Predicting lateral impacts upon interventions is one of the primary

goals of Policy Enunciator. The more and richer the models we have for different aspects of governance, the better we will be able to track lateral repercussions.

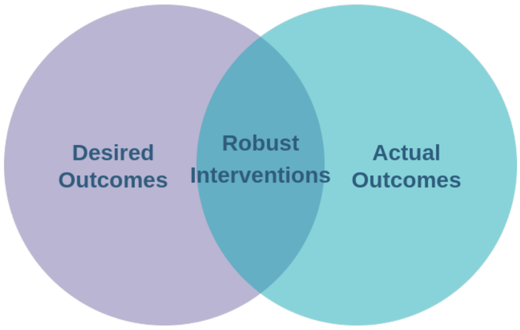

Figure 8: Desired vs Actual Outcomes and Maximization of Robust Interventions

The achievement of outcomes is needed to be maintained and maximized over time for attaining sustainability in practice. Fig.8 represents the intersection between the desired and actual outcomes. The intersection between the desired and actual outcomes are the robust interventions that have sustained the intervention impacts and brought out preferred policy changes. PSS aims at maximizing the robustness of the policy interventions.

Yes, it is challenging to evaluate PSS. The plan is to seek domain expertise to evaluate the correctness of the models which includes the construction of the dependency arcs, the interventions tested, the choice of target variables, and the generation of predictions and subsequent interpretations. Different scenarios with policy interventions and what-if analyses will be prepared for evaluation by the experts. With reference to the model in Fig. 6, a scenario-based question such as "If we ban *NPK Consumption* and use a high level of *Organic Fertilizers*, what could be the expected *Rice Production* levels?" is proposed for the expert in addition to multiple other questions. The evaluation forms for the experts will help in testing the credibility of these scenarios. We need to rely on the knowledge, experience, and judgment of the domain experts for the assessment of PSS.

4. Conclusion

Policy Support System incorporated with policy enunciation inspects real-world unpredictability that aids policymakers and planners to identify potential scenarios before implementing policies. PSS elucidates the consequences of the recommendations proposed using causal dependency graphs. In the context of open government data, the provenance of data becomes well-known making it sound and trustworthy for interpretation. Bayesian Networks provide a mechanism to model uncertainties under complex domains hence, representing expert knowledge with these models produces sufficient evidence to reason around the final course of action. Policy enunciation to track both intended and collateral consequences is the primary motivation for building the Policy Support System. The objective is to enhance the engagement of policymakers with the PSS towards capturing the total policy perspective to make judicious policy decisions.

Acknowledgments

I thank the Department of Planning and Statistics, Government of Karnataka, India, and our collaborating agencies, Centre for Open Data Research(CODR), Public Affairs Centre, Jigani, Bengaluru, India.

References

[1] G. H. Brundtland, Report of the World Commission on environment and development:" our common future.", UN, 1987.
[2] Sdg india index baseline report, 2018, 2018. URL: https://www.niti.gov.in/.
[3] India national agriculture market, 2022. URL: https://www.enam.gov.in/web/.
[4] Umang enam, 2022. URL: https://web.umang.gov.in/landing/department/enam.html.
[5] G. o. I. National Informatics Center, Ministry of Electronics & Information Technology, Jal jeevan mission, https://jaljeevanmission.gov.in/, 2022.
[6] Internet-connected sensors to monitor village water supply, https://www.thehindu.com/news/national/internet-connected-sensors-to-monitor-village-water-supply/article34210534.ece, 2021.
[7] P. Bassin, N. S. Parasa, S. Srinivasa, S. Mandyam, Big data management for policy support in sustainable development, in: International Conference on Big Data Analytics, Springer, 2021, pp. 3–15.
[8] N. Muni Kumar, R. Manjula, et al., Role of big data analytics in rural health care-a step towards svasth bharath, International Journal of Computer Science and Information Technologies 5 (2014) 7172–7178.
[9] P. Mondal, M. Basu, Adoption of precision agriculture technologies in india and in some developing countries: Scope, present status and strategies, Progress in Natural Science 19 (2009) 659–666.
[10] M. Singh, S. O. Adebayo, M. Saini, J. Singh, Indian government e-learning initiatives in response to covid-19 crisis: A case study on online learning in indian higher education system, Education and Information Technologies 26 (2021) 7569–7607.
[11] Diksha - digital infrastructure for knowledge sharing, 2022. URL: https://diksha.gov.in/.
[12] Swayam prabha - free dth channel for education, 2022. URL: https://www.swayamprabha.gov.in/.
[13] G. Wright, S. Abraham, N. Shah, Open government data study: India, Available at SSRN 2071605 (2012).
[14] J. Pearl, Bayesian networks (2011).
[15] M. Scutari, Understanding bayesian networks with examples in r, https://www.bnlearn.com/about/teaching/slides-bnshort.pdf, 2017. Department of Statistics, University of Oxford.
[16] L. Kiss, D. Fotheringhame, J. Mak, A. McAlpine, C. Zimmerman, The use of bayesian networks for realist evaluation of complex interventions: evidence for prevention of human trafficking, Journal of computational social science 4 (2021) 25–48.

[17] Public policy: Meaning and nature, https://egyankosh.ac.in/bitstream/123456789/19329/1/Unit-1.pdf, 2018.
[18] D. Yarlett, M. Ramscar, Uncertainty in causal and counterfactual inference, in: Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society, Routledge, 2019, pp. 956–961.
[19] D. Daniel, A. Diener, S. Pande, S. Jansen, S. Marks, R. Meierhofer, M. Bhatta, L. Rietveld, Understanding the effect of socio-economic characteristics and psychosocial factors on household water treatment practices in rural nepal using bayesian belief networks, International Journal of Hygiene and Environmental Health 222 (2019) 847–855.
[20] S. Lekar, D. Shumeiko, V. Lagodiienko, V. Nemchenko, Construction of bayesian networks in public administration of the economy, International Journal of Civil Engineering and Technology 10 (2019) 1379–1384.
[21] V. Lytvynenko, D. Nikytenko, M. Voronenko, N. Savina, O. Naumov, Assessing the possibility of a country's economic growth using dynamic bayesian network models, in: 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT), volume 1, IEEE, 2020, pp. 36–39.
[22] Bayes fusion, llc: Publications based on our software, 2022. URL: https://www.bayesfusion.com/publications/.
[23] Norsys software corp: Netica, 2022. URL: https://norsys.com/.
[24] Hugin expert: Capitalize on uncertainity using bayesian network technology, 2022. URL: https://www.hugin.com/.
[25] Bayes server: Bayesian network and causal artificial intelligence software, 2022. URL: https://www.bayesserver.com/.
[26] Norsys software corp: Commercial client list, 2022. URL: https://www.norsys.com/clients.htm.
[27] J. Pearl, D. Mackenzie, The book of why: the new science of cause and effect, Basic books, 2018.
[28] I. P. Derks, A. d. Waal, A taxonomy of explainable bayesian networks, in: Southern African Conference for Artificial Intelligence Research, Springer, 2021, pp. 220–235.
[29] S. Ray, I. A. Ray, et al., Role and effectiveness of public distribution system in assuring food security in india: An appraisal, Journal of Economics and Sustainable Development 2 (2011) 238–251.
[30] D. Sinha, Hunger and food security in the times of covid-19, Journal of Social and Economic Development 23 (2021) 320–331.
[31] A. Kishore, S. Chakrabarti, Is more inclusive more effective? the 'new style'public distribution system in india, Food Policy 55 (2015) 117–130.
[32] A. Gulati, S. Saini, Leakages from public distribution system (PDS) and the way forward, Technical Report, Working paper, 2015.
[33] A. Kumar, S. Ayyappan, Food security and public distribution system in india, Agricultural Research 3 (2014) 271–277.
[34] C. Chandankhede, D. Mukhopadhyay, A proposed architecture for automating public distribution system, in: 2017 International Conference on Computing, Communication and Automation (ICCCA), IEEE, 2017, pp. 935–939.