ORIGINAL RESEARCH

Does Climate Variability Impact COVID-19 Outbreak? An Enhanced Semantics-Driven Theory-Guided Model

Monidipa Das[1] · Akash Ghosh[2] · Soumya K. Ghosh[3]

Received: 8 December 2020 / Accepted: 30 August 2021 / Published online: 9 September 2021

Abstract

COVID-19, a life-threatening infection by novel coronavirus, has broken out as a pandemic since December 2019. Eventually, with the aim of helping the World Health Organization and other health regulators to combat COVID-19, significant research effort has been exerted during last several months to analyze how the various factors, especially the climatic aspects, impact on the spread of this infection. However, due to insufficient test and lack of data transparency, these research findings, at times, are found to be inconsistent as well as conflicting. In our work, we aim to employ a semantics-driven probabilistic framework for analyzing the causal influence as well as the impact of climate variability on the COVID-19 outbreak. The idea here is to tackle the data inadequacy and uncertainty issues using probabilistic graphical analysis along with embedded technology of incorporating semantics from climatological domain. Furthermore, the theoretical guidance from epidemiological model additionally helps the framework to better capture the pandemic characteristics. More significantly, we further enhance the impact analysis framework with an auxiliary module of measuring semantic relatedness on regional basis, so as to realistically account for the existence of multiple climate types within a single spatial region. This added notion of regional semantic relatedness further helps us to attain improved probabilistic analysis for modeling the climatological impact on this disease outbreak. Experimentation with COVID-19 datasets over 15 states (or provinces) belonging to varying climate regions in India, demonstrates the effectiveness of our semantically-enhanced theory-guided data-driven approach. It is worth noting that our proposed framework and the relevant semantic analyses are generic enough for intelligent as well as explainable impact analysis in many other application domains, by introducing minimal augmentation.

Keywords Semantic Bayesian analysis · Theory-guided approach · Climate variability · COVID-19

Introduction

The novel coronavirus disease 2019 or COVID-19 has become a serious health hazard throughout the globe. Because of its excessive infectivity, spreading capability, and ubiquitous nature, COVID-19 has been categorized as pandemic by the World Health Organization (WHO). The recent reports have already proved that relying on classic infection-control and public-health measures for tackling this pandemic are not sufficient [17, 23]. Consequently, several research initiatives have been undertaken across the globe to fight against this pandemic by leveraging the recent advancements in science and technology.

Analyzing the impact of geographic climate variations in modulating the COVID-19 outbreak is one of such current research concerns which has gained substantial attention. Numerous research publications in this context can be found in the literature. Nevertheless, owing to the diverse

This article is part of the topical collection "Computer Aided Methods to Combat COVID-19 Pandemic" guest edited by DavidClifton, Matthew Brown, Yuan-Ting Zhang and Tapabrata Chakraborty".

✉ Monidipa Das
monidipadas@hotmail.com

Akash Ghosh
akashkgp@gmail.com

Soumya K. Ghosh
skg@cse.iitkgp.ac.in

[1] Machine Intelligence Unit (MIU), Indian Statistical Institute (ISI), Kolkata, India

[2] Department of Computer Science and Engineering, Jadavpur University (JU), Kolkata, India

[3] Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Kharagpur, India

screening strategies, poor quality of collected data, and inadequate number of COVID tests, these research outcomes often become contradictory to each other. For example, a collaborative research of China-USA team has observed that low humidity, low temperature, and mild diurnal temperature range may promote this disease transmission [13], whereas another group of researchers from USA has found that the hot and humid climate does not help controlling the COVID-19 outbreak [2].

In our proposed impact analysis scheme, we aim at addressing such uncertainty issues by intelligent incorporation of theoretical knowledge from the domain of epidemiology and semantic knowledge from the domain of climatology. The theoretical guidance from epidemiological model helps our impact analysis to remain consistent with the underlying theory of the infectious disease spread. On the other side, the study of semantic relatedness in the datasets aids in reducing uncertainty at the time of learning causal relationships between disease development and climate variability. Unlike majority of the existing models, instead of considering the individual climate factors, we deal with the overall climatic pattern of the geographical areas. Though our present approach of semantics-driven theory-guided impact analysis is influenced from our recently introduced framework [9], the fundamental difference between the two lies in the following two aspects. Firstly, in contrast to [9], the semantic analysis in the presently proposed framework is probabilistically augmented to account for the real-life scenario where the same climate zone may be expanded over multiple spatial regions. Secondly, the impact analysis in our present scheme is performed with consideration to the expected values of relative-recovered cases and per million new infected/confirmed cases over the various regions, which eventually make our present scheme capable of providing a more realistic view.

Motivation In essence, our research is motivated by the fact that *epistemic uncertainty* [7] can be reduced using more information and knowledge regarding the relevant domain. For example, added insights from mathematical models can help in handling uncertainty that emerges because of our unawareness about the *basic tenet* of epidemiology [16]. Further, the use of additional data samples can also help in tackling uncertainty by reducing sampling error during inference generation. In this regard, we may exploit the semantic relatedness between data collected from different spatial regions [9]. For instance, as explained in Fig. 1, the climate class (BWk) of the spatial region-2 is quite similar to that of region-3 (BWh) and region-4 (BSk), since all these belong to arid type of climate. Accordingly, at the time of analyzing climatological impact on COVID-19 spread in region-2, we may utilize the data samples from region-3 and region-4 as well, along with the relevant measures of semantic relatedness. However, designing appropriate measure of semantic relatedness, especially when a region contains multiple climate types in its various parts, becomes a challenging task. Our earlier work [9] assumes that each spatial region belongs to strictly one climate zone. Accordingly, it cannot measure the semantic relatedness between the data collected from region-1 (climate type: BSh + Aw) and that collected from other considered spatial regions, as depicted in Fig. 1b, c. To overcome this limitation and to make the impact analysis model more appropriate for dealing with real-life scenario, in this work, we introduce the concepts of ‘regional semantic relatedness’ and ‘semantic generic index’, which aid in enhancing the prediction model developed in [9].

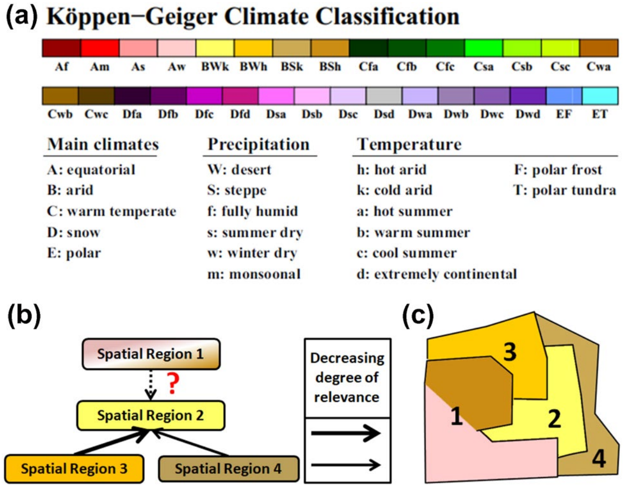


Fig. 1 Example showing semantic relatedness between climate type of spatial region-2 and that of the others, as per the commonality in climate type (denoted by colors)

Contributions It may be noted from the motivational example that, the primary challenges involved in this research are, firstly, to define suitable measure for quantifying semantic relatedness, and secondly, to develop a scheme for combining climate domain semantics as well as epidemiological knowledge into a data-driven model. In this work, we address both the aforementioned challenges by considering real-life scenario where a spatial region may partly belong to multiple climate zones. Accordingly, the key contributions of our work are as follows.

- Exploring effective means for representing climatological domain knowledge/semantics.
- Defining probabilistic measure of regional semantic relatedness (*regSR*) to account for the existence of multiple climate types within a spatial region.
- Developing an advanced version of data-driven framework that can offer theory-guided semantics-driven analysis of climatological impact on COVID-19 spread, while considering real-life scenario for regional distribution of climate.

- Proposing *Semantic-GI* as a semantics-driven generic index for analyzing the influence of climate variability on the regional outbreak of COVID-19.

The proposed model is validated using daily time series of COVID-19 data over 15 states (provinces) belonging to diverse climate regions in India. Our experimental findings indicate that dry (arid/semi-arid) climate zones, like BSh, BWh etc., are most susceptible for COVID-19 transmission. The temperate climate zones, e.g. Cwa, are also quite vulnerable as the daily relative-recovery[1] in these zones are comparatively lower than that in tropical climate zones. Our study also identifies *humid climate* to be a principal factor favoring the daily relative-recovery in India.

Incidentally, our proposed framework is not only applicable for the present purpose of assessing climatological impact on COVID-19 spread, but also it is potential enough for analyzing the impact of any other categorical factor on the possible transmission of COVID-19. For example, given the classification of the cities in terms of house rent allowance (HRA) grade, the similar framework can be used for analyzing the impact of population density on the COVID-19 transmission on sub-regional basis. More significantly, the proposed measures of *regional semantic relatedness* and *Semantic-GI*, are some generic notions, which can be successfully utilized for semantics-driven explainable analyses in diverse application areas of machine learning and computational intelligence.

The rest of the paper is structured as follows. Section "Problem Scenario" describes the overall problem scenario. Section "Methodological Overview" discusses on the methodological details of the proposed impact analysis model. Section "Experimental Evaluation" presents the experimental evaluation of our proposed model in comparison with state-of-the-art approaches. Section "Related Works" provides a summary of the various related works, and finally, we conclude in Section "Conclusions".

Problem Scenario

Given the daily statistics of COVID-19 case count, including *confirmed cases*, *recovered cases*, and *active cases* over a set of regions with known climate patterns, the prime goal of this research is to explore any possible correlation between the various climate types and the development of COVID-19 in the regions. Accordingly, we aim at answering the following research questions (**RQs**).

- **RQ1:** Do the climate patterns of the regions have any correlation with the daily statistics of confirmed/ recovered case counts?
- **RQ2:** In case such correlation is identified, which climate type(s) help(s) increasing/decreasing the infected/ recovered case count the most?

By the term "climate pattern" here we indicate spatial distribution of various climate classes, including *equatorial/ tropical* (e.g. 'Am', 'As' etc.), *arid* (e.g. 'BWh', 'BWk' etc.), and *temperate* (e.g. 'Cwa', 'Cfa' etc.) etc., as defined in [12] (refer Fig. 1a).

Methodological Overview

An overview of process flow within our proposed framework is shown in Fig. 2. Primarily, the overall process is comprised of five major activities: (i) representing climate domain semantics, (ii) estimating semantic relatedness on regional basis,(iii) modeling probabilistic relationship between COVID development and regional climate types, based on enhanced semantics-driven theory-guided analysis, (iv) conducting predictive analysis for COVID-19 cases, and (v) assessing climatological impact on the disease outbreak. The various steps in this regard are discussed in the subsequent subsections. Both the research questions (**RQ1** and **RQ2**) are primarily answered in the final step.

Representing Climate Domain Semantics

The objective here is to represent the climatological domain knowledge in the form of a semantic graph/network which can be utilized in successive step to reduce epistemic uncertainty of learning probabilistic relationship between regional variability of climate types and COVID development.

Formation of Semantic Network

The typical semantic network used in our proposed model is shown in Fig. 3. The network is generated by combining semantic hierarchies rooted at the various major concepts related to climate pattern, namely *precipitation* (*P*), *main climate* (*MC*), and *temperature* (*T*), as mentioned in Köppen Geiger climate classification [12]. More specific concepts (such as 'summer dry'(s), 'desert' (W), 'hot arid' (h), 'cold arid' (k), etc.), which can be directly related to any of these root concepts, are represented by the intermediate nodes in the network. Finally, the exact climate classes (such as As, BWh, Cwa, etc.) which are indirectly related to multiple root

[1] The 'relative recovery' is defined as the relative count of recovered case with respect to the confirmed case count on the same day. Accordingly, relative recovery = $\frac{\text{(daily recovered case count)}}{\text{(daily confirmed case count)}}$, and hence, it is not exactly the same as the 'recovery rate'.

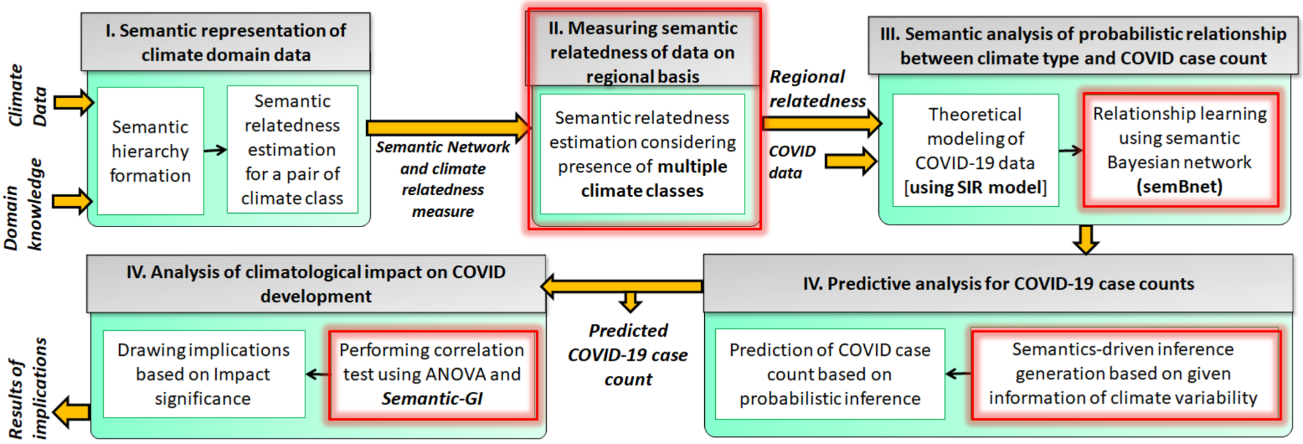

Climate Data
Domain knowledge
I. Semantic representation of climate domain data
Semantic hierarchy formation
Semantic relatedness estimation for a pair of climate class
Semantic Network and climate relatedness measure
II. Measuring semantic relatedness of data on regional basis
Semantic relatedness estimation considering presence of multiple climate classes
Regional relatedness
COVID data
III. Semantic analysis of probabilistic relationship between climate type and COVID case count
Theoretical modeling of COVID-19 data [using SIR model]
Relationship learning using semantic Bayesian network (semBnet)
IV. Predictive analysis for COVID-19 case counts
Prediction of COVID case count based on probabilistic inference
Semantics-driven inference generation based on given information of climate variability
Predicted COVID-19 case count
IV. Analysis of climatological impact on COVID development
Drawing implications based on Impact significance
Performing correlation test using ANOVA and Semantic-GI
Results of implications

Fig. 2 Proposed framework: overall process flow. [The red boxes indicate our major contributing steps.]

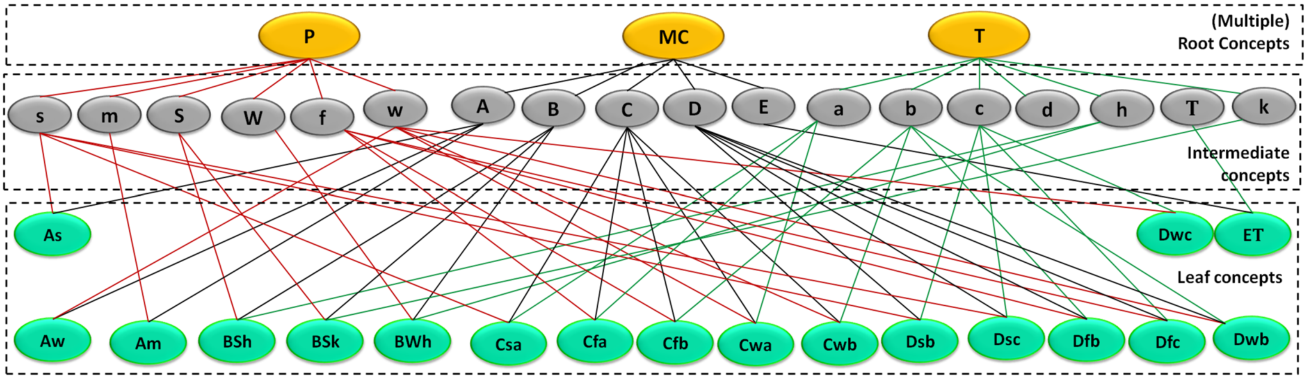


Fig. 3 Semantic network corresponding to the climatological concept defined in Fig. 1a [9]

concepts via intermediate ones, are represented by the leaf nodes in the semantic network (refer Fig. 3).

Measuring Semantic Relatedness

To measure the semantic relatedness (SR) between any pair of climate classes (say CT_1 and CT_2), the conceptual relationship as represented through the semantic network is utilized [9]. Formally, SR can be presented as follows.

$$SR\left(CT_1, CT_2\right) = e^{-\lambda\cdot\max_l}\cdot\frac{e^{\delta\cdot D\cdot\max_d}-e^{-\delta\cdot D\cdot\max_d}}{e^{\delta\cdot D\cdot\max_d}+e^{-\delta\cdot D\cdot\max_d}}. \qquad (1)$$

Here, $max_l = \max(l_1, \ldots, l_{\mathcal{R}})$ represents the maximum of the shortest path lengths $l_1, l_2, \ldots, l_{\mathcal{R}}$ between CT_1 and CT_2 via each of the R root concepts; $max_d = \max(l_1, \cdots, l_{\mathcal{R}})$ represents the maximum of the subsumer depth $d_1, d_2, \ldots, d_{\mathcal{R}}$ relevant to CT_1 and CT_2, measured in regards to each of the $\mathcal{R}$ root concepts. λ and δ are scaling parameters that help adjusting the contribution of $\max_l$ and $\max_d$, respectively. D represents the *degree of conceptual overlap*, defined as follows [9].

$$D = \begin{cases} 2, & \text{if } CT_1 \text{ and } CT_2 \text{ overlap in terms of } MC \\ 1, & \text{if the overlap is in terms of } P \text{ and/or } T \\ 0, & \text{if } CT_1 \text{ and } CT_2 \text{ have no overlapping concept.} \end{cases} \qquad (2)$$

Definition 1 *Semantic relatedness* $SR(CT_i, CT_j)$ is a quantitative measure of commonality between climate type CT_i and CT_j, computed with respect to the semantics of main climate as well as temperature and precipitation pattern. To be noted, given the underlying semantics in the form of hierarchical representation, the SR measure can also be applied on any pair of concepts from domains beyond climatology.

Measuring Semantic Relatedness on Regional Basis

In this stage we aim at determining the semantic relatedness of data collected from various spatial regions belonging to different climate types. As per the semantic relatedness measure defined above, the semantic relatedness between data collected from a spatial region r under climate type CT^r and that collected from regions under climate type CT_j, can be estimated as $regSR\ (r, CT_j) = SR(CT^r, CT_j)$. Note that, here we use subscript to denote specific climate type and superscript to denote the spatial region to which the climate type is associated with. However, it is quite common that a large spatial region is covered with multiple climate zones in its various parts. That means, CT^r can be a set of climate types, and therefore we may define it as $CT^r = \left\{CT_1^r, CT_2^r, \dots\right\}$. In such case, the semantic relatedness must also consider the percentage of area covered by each climate type within the region. The higher the covered area percentage the more the probability of the data being collected from the particular climate zone. Accordingly, to tackle the issue of presence of multiple climate types within a region, we enhance the regional semantic relatedness measure in following manner.

$$\begin{aligned} &regSR(r, CT_j)\\ &\quad= \frac{1}{(n-1)} \sum_{r'(\neq r)} \max\left(P\left(CT_i^r\right) \cdot P\left(CT_j^{r'}\right)\right.\\ &\qquad \left.\cdot SR\left(CT_i^r, CT_j^{r'}\right)\right), \forall CT_i^r \in CT^r, \end{aligned} \tag{3}$$

where $P(CT_i^r)$ denotes the probability of presence of a climate type $CT_i\ (i = 1, 2, \dots)$ in a region r, $P(CT_j^{r'})$ denotes the probability of presence of a climate type CT_j in region $r'(\neq r)$, n is the total number of spatial regions (including r). The $P(CT_i^r)$ can be estimated as the area of climate type CT_i covered in per unit area of the region r. Similarly, $P(CT_j^{r'})$ can be estimated as the area of climate type CT_j covered in per unit area of the region r'.

Definition 2 *Regional semantic relatedness* $regSR\ (r, CT_i)$ is a quantitative measure of semantic relatedness between the data collected from region r and that collected from regions having climate type CT_i. The $regSR(r, CT_i)$ assumes that the region r may be associated with multiple climate types which may or may not include CT_i. Moreover, here, the data collected from CT_i may be associated with multiple regions.

Semantic Analysis of Relationship Between Climate Type and COVID Case Development

This step aims at modeling the causal influence of climate variability over the dynamics of new confirmed cases, active cases, and recovered cases of COVID-19. In this context, we employ semantically enhanced Bayesian network [6], where the Bayesian model handles the uncertainty by its *probabilistic analysis* for learning causal relationship. The incorporation of climate domain semantics further helps in tackling the uncertainty by reducing sampling error during inference generation step. The Bayesian network structure as used in our model is shown in Fig. 4. This directed acyclic graph primarily represents the causal dependency among climate type (CT), confirmed case (CC), recovered case (RC) and active case (AC). To generate this dependency structure, we employed structure learning based on a combination of HPC (hybrid parents and children) and pairwise mutual information algorithms [10, 14].

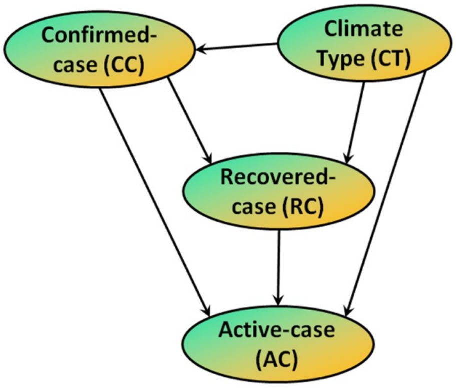


Fig. 4 Causal dependency between climate type and COVID-19 case counts

Semantically-Enhanced Bayesian Modeling of Causal Relationships

Given the causal dependency graph, the conditional probability distributions for confirmed case count (CC), recovered case count (RC) and active case count (AC) in any spatial region r for each timestamp can be obtained through semantic Bayesian analysis as follows.

$$P^{\dagger}\left(CC|CT^r\right) = \psi \cdot \left(\sum_j regSR\left(r, CT_j\right) \times P\left(CC|CT_j\right)\right) \tag{4}$$

$$\begin{aligned} &P^{\dagger}\left(RC|CC, CT^r\right)\\ &\quad= \psi \cdot \left(\sum_j regSR\left(r, CT_j\right) \times P\left(RC|CC, CT_j\right)\right) \end{aligned} \tag{5}$$

$$\begin{aligned} &P^{\dagger}\left(AC|CC, RC, CT^r\right)\\ &\quad= \psi \cdot \left(\sum_j regSR\left(r, CT_j\right) \times P\left(AC|CC, RC, CT_j\right)\right), \end{aligned} \tag{6}$$

where,

$$P\left(CC|CT_j\right) = \frac{1}{\sigma_{CC}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{CC-\theta_{0j}}{\sigma_{CC}}\right)^2} \tag{7}$$

$$P\left(RC|CC,CT_j\right)=\frac{1}{\sigma_{RC}\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{RC-(\theta_{2j}\cdot CC+\theta_{1j})}{\sigma_{RC}}\right)^2} \tag{8}$$

$$P\left(AC|CC,RC,CT_j\right)=\frac{1}{\sigma_{AC}\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{AC-(\theta_{5j}\cdot CC+\theta_{4j}\cdot RC+\theta_{3j})}{\sigma_{AC}}\right)^2}. \tag{9}$$

In Eqs. 4–9, ψ denotes the normalization constant, CT_j denotes the j-th climate type, and $regSR(r, CT_j)$ denotes the semantic relatedness between COVID data collected from region r and those collected from regions having climate type CT_j; the σ_{CC}, σ_{RC}, and σ_{AC} are the standard deviations corresponding to confirmed case count, recovered case count, and active case count, respectively; the $\left\{\theta_{0j}\right\}$, $\left\{\theta_{1j}, \theta_{2j}\right\}$, and $\left\{\theta_{3j}, \theta_{4j}, \theta_{5j}\right\}$ are parameters regulating the means of confirmed case count, recovered case count, and active case count, respectively. These can be computed by employing maximum likelihood analysis of Expectation Maximization (EM) algorithm [11].

Theoretical Analysis for Data Sample Generation

It is interesting to note here that every epidemic/pandemic, like COVID-19, has some particular temporal development pattern which is governed by several other factors including susceptible population size, contagiousness or transmissibility of the disease, and so on. Accordingly, if the disease infected case counts predominantly show upsurge during the initial phase, then the purely data-driven techniques, including the dynamic Bayesian models, cannot properly guess the declining trend of the infected case count in long run. It is therefore necessary to provide theoretical guidance to the data-driven models so that the learnt parameters remain consistent with the underlying physics of epidemiological development.

To incorporate theoretical knowledge, first, we utilize the kinetic scheme as defined by Kermack-McKendrick SIR Model [22]. This is an epidemiological model that can mathematically express (refer Eqs. 10–12) the dynamic interaction among susceptible (S), infected (I), and recovered/removed (R) fraction of population in a region. Subsequently, we follow this system of differential equations to generate training samples of COVID-19 cases ensuring that the parameters learnt through our semantic Bayesian analyses are compatible with the underlying theory of epidemiology.

$$\frac{\mathrm{d}S}{\mathrm{d}t}=-\beta SI \tag{10}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t}=\beta SI-\gamma I \tag{11}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t}=\gamma I. \tag{12}$$

In Eqs. (10)–(12), β and γ indicate the effective contact rate and the mean recovery rate, respectively, and t indicates the time. Given the new confirmed case count (CC) and the new recovered case count (RC) as recorded on every t basis, $R(t)$ can be estimated as $cusum(RC)$, and $I(t)$ can be estimated as ($cusum(CC)-$ $cusum(RC)$) for each t, where $cusum()$ is the function to compute cumulative sum [9]. As per the SIR model, $S(t)+I(t)+R(t)$ is assumed to remain constant for all t and the sum is equal to the population size of the region. It may please be noted here that, to be consistent with the terms used by Mathematical Association of America (MAA), we have sometimes used the word "Recovered" to indicate 'R', which though includes recovered as well as death case count.

Typical pattern of temporal development of COVID-19 cases, as obtained by employing SIR model, is shown in Fig. 5a. Our framework utilizes these theory-governed temporal distributions of active cases, new confirmed cases, and new recovered cases (refer Fig. 5b, c) to produce appropriate training samples that can eventually help the parameter learning process to remain congruous with the physical understanding of COVID-19 spread. Note that, despite the availability of various enhanced versions of SIR model, we choose the basic one for our epidemiological modeling, since the recent findings [19] demonstrate that this simplest version can also adequately model COVID-19 dynamics.

Predictive Analysis for COVID-19 Cases

After the causal relationships are learnt, the semantically-enhanced theory-guided model can be used for inferring the COVID-19 case counts (confirmed case count, active case count and recovered case count) given the evidence on climate variability in a region r. For this purpose, our framework employs semantic Bayesian inference generation [6, 8] in following manner.

$$P^{\dagger}(CC|CT^r)=\sum_{RC}\sum_{AC}\left\{P(CT^r).P^{\dagger}(CC|CT^r).\right.$$
$$\left.P^{\dagger}(RC|CC,CT^r).P^{\dagger}(AC|CC,RC,CT^r)\right\} \tag{13}$$

$$P^{\dagger}(RC|CT^r)=\sum_{CC}\sum_{AC}\left\{P(CT^r).P^{\dagger}(CC|CT^r).\right.$$
$$\left.P^{\dagger}(RC|CC,CT^r).P^{\dagger}(AC|CC,RC,CT^r)\right\} \tag{14}$$

Here, $P(CT^r)$ denotes the marginal probability of presence of various climate types in region r, and this can be estimated as $\sum_i P(CT_i^r)$, where $CT_i^r \in CT^r$ (refer Section "Measuring Semantic Relatedness on Regional Basis"). To be noted, this semantically enhanced inference generation for COVID-19 cases overcomes uncertainty issue, emerging due to sample scarcity, and also maintains the theoretical guidelines which are utilized at the time of estimating $P^{\dagger}(RC|CC, CT^{r})$, $P^{\dagger}(AC|CC, RC, CT^{r})$ etc., in the parameter learning phase.

Assessing Climatological Impact on COVID-19 Outbreak

This is the ultimate step which aims at assessing whether the regional climate variability has any correlation with the patterns of confirmed and recovered case development for COVID-19. Since, instead of the continuous climatic factors, our framework deals with the categorical values of climate types, we use ANOVA test [21] for analyzing the correlation with climate variability.

Additionally, we propose *Semantic-GI* as a generic index (correlation measure) so as to utilize the underlying semantics from climate domain. The *Semantic-GI* is measured as follows.

$$\text{Semantic-GI} = \frac{n \cdot \sum_{r=1}^{n} \sum_{q=1}^{n} sw_{rq} \cdot (x_r - x')(x_q - x')}{\left(\sum_{r=1}^{n} \sum_{q=1}^{n} sw_{rq}\right) \cdot \left(\sum_{r=1}^{n} (x_r - x')^2\right)}, \tag{15}$$

where, n is the total count of study regions considered. These belong to diverse climate zones, x' is the mean of COVID case counts (confirmed case count or recovered case count) per million individual over all the regions, x_r is the COVID case count per million individual at a particular spatial region r, and sw_{rq} is the *semantic weight* between region r and region q. The *semantic weight* between any pair of regions r and q associated with climate classes CT^r and CT^q, is defined as follows.

$$sw_{rq} = \begin{cases} 1, \mathit{if}\ r == q \\ \max\left(P(CT_i^r) \cdot P\left(CT_j^q\right) \cdot SR\left(CT_i^r, CT_j^q\right)\right) \\ \forall CT_i^r \in CT^r, CT_j^q \in CT^q, \text{otherwise.} \end{cases} \tag{16}$$

In case each of the considered spatial regions (r and q) are associated with single climate type, i.e., if CT^r and CT^q contain only one element each, then the Eq. 16 is simplified as follows: $sw_{rq} = SR(CT^r, CT^q)$.

It can be interpreted from Eq. 15 that, similar to the Semantic-I measure [9], our presently proposed *Semantic-GI* is also founded on the concept of *semantic auto-correlation*. However, our computation of semantic weight (sw) wisely takes into account the presence of multiple climate types within each spatial region (refer Eq. 16), which makes our model more appropriate to deal with real-world scenario. The proposed *Semantic-GI* can be used to analyze whether the COVID-19 case counts associated with the various regions with semantically related climate type form a cluster pattern or a disperse pattern. In specific, a positive value of *Semantic-GI* indicates a cluster pattern, whereas a negative value of *Semantic-GI* indicates a disperse pattern. In the same way as defined for Moran's index of spatial auto-correlation [5], the significance of *Semantic-GI* value can be quantified through Z-test. If Z-score > 2.58, it can be claimed with 99% confidence that there is a cluster pattern, whereas if Z-score < -2.58, with the same level of confidence it can be claimed that there is a disperse pattern. Otherwise, the pattern is random. Together, the ANOVA test and the *Semantic-GI* test help answering the **RQ1**.

To resolve the **RQ2**, our framework draws more specific conclusions regarding the influence of climate variability. Accordingly, the *Semantic-GI* analysis is followed by comparative study of semantically averaged case count (daily relative-recovered case count and daily new confirmed case count) for each climate type. The semantically averaged case count ($sAvg_i$) corresponding to a climate type CT_i is measured with consideration to the presence of multiple climate types within each spatial region, in following manner.

$$sAvg_i = \alpha_1 \cdot \sum_j \left(SR(CT_i, CT_j) \times regsAvg_j\right), \tag{17}$$

where $\alpha_1 = \frac{1}{\sum_j SR(CT_i, CT_j)}$ is the normalization constant, and $regsAvg_j$ is the semantically averaged case count corresponding to a climate type CT_j over all the considered region r. The $regsAvg_j$ can be mathematically presented as follows, where x_r indicates the latest count of case (daily new confirmed case or relative-recovered case) per million people in r-th spatial region, and $\alpha_2 = \frac{1}{\sum_r regSR(r, CT_j)}$ is the normalization constant.

$$regsAvg_j = \alpha_2 \cdot \sum_r \left(regSR(r, CT_j) \times x_r\right). \tag{18}$$

Once the $sAvg_i$ for each climate type CT_i is estimated, these can be graphically plotted to draw conclusion on which particular climate type has higher/lower impact on the development of COVID-19. The qualitative estimate of 'higher' or 'lower' can be decided based on the expected value (refer Table 2). For the confirmed case, the expected value is computed considering all the spatial regions, irrespective of the climate types. On the other side, for the relative-recovered case, the expected value becomes 1, indicating $\frac{\text{daily recovered case count}}{\text{daily confirmed case count}} = 1$.

The various symbols used in this section (Section "Methodological Overview") are summarized in Table 1.

Table 1 Symbols and notations used in the Section "Methodological Overview"

Notation	Meaning
β	Effective contact rate (theoretical)
γ	Mean recovery rate (theoretical)
AC	Daily active case count
CC	Daily new confirmed case count
CT	Climate type
CT_i	i-th Climate type
CT^r	Set of climate types associated with region r
CT_i^r	i-th climate type associated with region r
D	Degree of conceptual overlap
I	Infected fraction of regional population
n	Total number of spatial regions under study
P	Probability distribution corresponding to standard Bayesian network
$P^\dagger$	Probability distribution corresponding to semantic Bayesian network
$\mathcal{R}$	Number of roots in the semantic network
R	Recovered/removed fraction of regional population
RC	Daily new recovered case count
$regSR(r, CT_i)$	Regional semantic relatedness between data collected from region r and that collected from that from regions having climate type CT_i
$regsAvg_i$	Semantically averaged case count relevant to climate type CT_i over all the considered region r
S	Susceptible fraction of regional population
$SR(CT_i, CT_j)$	Semantic relatedness between climate type CT_i and CT_j
$sAvg_i$	Semantically averaged case count (new confirmed or new recovered case) relevant to climate type CT_i
sw	Semantic weight
sw_{rq}	Semantic weight between region r and q
x_r	COVID case count (new confirmed or new recovered case) per million individual associated with region r

Experimental Evaluation

This section evaluates our proposed impact analysis framework with consideration to the COVID-19 spread scenario in India, which is presently found to be one of the most adversely affected countries in the world.

Dataset and Study Area

The experimentation is carried out using the daily data[2] over COVID-19 case count. This includes active cases, new confirmed cases, and recovered cases over 15 different states in India (Fig. 6). All these states belong to variants of climate zones. Moreover, a single state may have multiple climate zones in its various parts which can be prudently handled by our model. The details of all the considered states are presented in Tables 3 and 4 which show that the considered states are associated with 6 different climate types/classes: Am, As, Aw, BSh, BWh, and Cwa [12]. The semantic relatedness (*SR*) among these climate types, and the regional semantic relatedness (*regSR*) with various climate types, have been calculated as per our proposed approach and the same are summarized in Tables 5, 6. The entire experiment is carried out considering the daily time series of active, confirmed, and recovered case count from the mid of March 2020 to the mid of November 2020.

Baselines and Experimental Set-Up

Since, in literature there has not been an agreed gold standard in terms of assessing climatological impact on COVID-19 outbreak, we primarily *compare only the prediction power* of our enhanced semantics-driven approach, with the existing linear regression (LR) [1] and nonlinear regression (NLR) [13] models. We also consider our recently introduced semantically-enhanced th-eory-guided model (SETG) [9] as one of the baselines, since our presently proposed enhanced semantics-driven theory-guided model is inspired from SETG.

The proposed enhanced semantics-driven theory-guided predictive analysis and all the baselines are executed in

[2] Data source: https://covid19india.org.

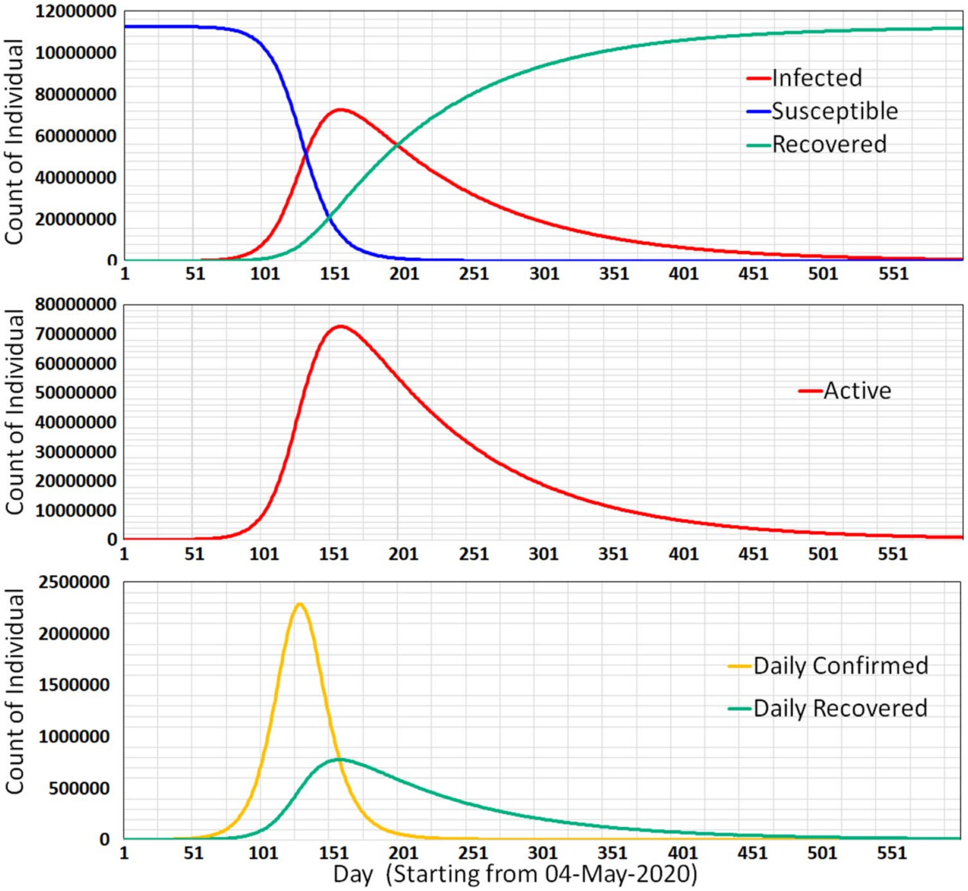


Fig. 5 Typical example of theoretically derived temporal pattern of COVID cases in Maharashtra, India: **a** overall development of susceptible, recovered, and infected cases, **b** development patterns of new confirmed and new recovered cases, **c** development pattern of latest active cases

Table 2 Qualitative estimate of COVID-19 case development with respect to expected value

Quantitative estimate of COVID-19 case development	Qualitative estimate
< 50% of Expected value	Low
50–100% of Expected value	High
100–200% of Expected value	Very high
> 200% of Expected value	Extremely high

R-tool[3] (version 4.0.0) in Windows 64-bit OS (3.1 GHz CPU processor and 4 GB RAM). The SIR-based theoretical modeling of COVID-19 and the structural learning of the Bayesian network have been conducted using 'SimInf' and 'bnlearn' packages of the R-tool.

Performance Metrics

The prediction performance has been measured in terms of root mean squared error (RMSE) and mean absolute error (MAE) [20], as defined below.

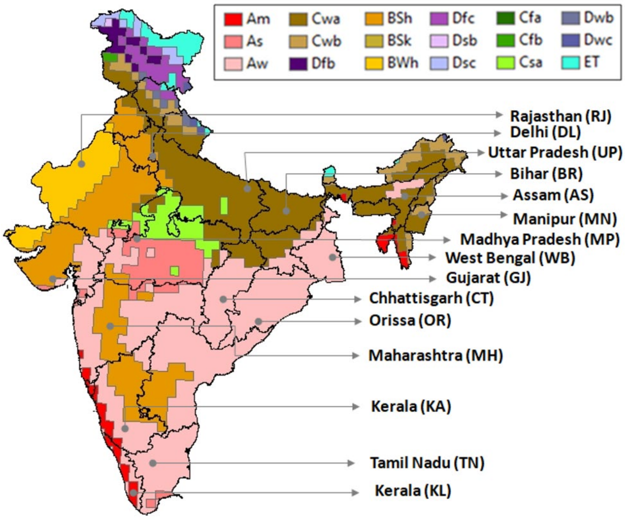


Fig. 6 Various Indian states (along with state codes) considered in the present case study. The color codes follow the Köppen–Geiger classification of regional climate [12]

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{V}_{o_i} - \mathcal{V}_{p_i}\right)^2} \tag{19}$$

[3] https://cran.r-project.org/bin/windows/Rtools/.

Table 3 Summary of the considered states in India

State Name	Code	Location	Major climate class	Population [18]
Assam	AS	N-E	Cwa	31,205,576
Bihar	BR	E	Cwa	104,099,452
Chhattisgarh	CT	Central	Aw	25,545,198
Delhi	DL	N	BSh	16,787,941
Gujarat	GJ	W	BSh, BWh, Aw	60,439,692
Karnataka	KA	S–W	Aw, BSh	61,095,297
Kerala	KL	S	Am	33,406,061
Madhya Pradesh	MP	Central	As	72,626,809
Maharashtra	MH	W	BSh, BWh	112,374,333
Manipur	MN	N–E	Cwa	2,855,794
Orissa	OR	E	Aw	41,974,218
Rajasthan	RJ	W	BWh, BSh	68,548,437
Tamil Nadu	TN	S	Aw	72,147,030
Uttar Pradesh	UP	N	Cwa	199,812,341
West Bengal	WB	E	Aw	91,276,115

N north, *S* south, *E* east, *W* west

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\mathcal{V}_{o_i} - \mathcal{V}_{p_i}\right|. \tag{20}$$

where, $\mathcal{V}_{o_i}$ indicates the count of COVID-19 case (confirmed case or recovered case) which is actually observed on the i-th day of prediction, and the $\mathcal{V}_{p_i}$ is the corresponding predicted value. In our experimental study, the prediction is made for **succeeding 2 months**, based on the daily observed data till 17-Sep-2020. ***Further,*** apart from the comparative study with respect to prediction performance, we also perform impact analysis using ANOVA test as well as using proposed *Semantic-GI* test.

Results and Discussions

The results of comparative prediction performance are presented in Tables 7, 8 and in Figs. 7, 8, whereas the summary of impact analysis is presented through Table 9 and in Figs. 9, 10. Our interpretations from the results are discussed below.

Table 4 Area of different major climate types per unit area in various Indian states

State code	Major climate types					
	Am	As	Aw	BSh	BWh	Cwa
AS	0	0	0	0	0	1
BR	0	0	0	0	0	1
CT	0	0	1	0	0	0
DL	0	0	0	1	0	0
GJ	0	0	0.28	0.56	0.16	0
KA	0	0	0.56	0.44	0	0
KL	1	0	0	0	0	0
MP	0	1	0	0	0	0
MH	0	0	0.7	0.3	0	0
MN	0	0	0	0	0	1
OR	0	0	1	0	0	0
RJ	0	0	0	0.5	0.5	0
TN	0	0	1	0	0	0
UP	0	0	0	0	0	1
WB	0	0	1	0	0	0

Table 5 Semantic relatedness (*SR*) between various climate types

Climate Type	Am	As	Aw	BSh	BWh	Cwa
Am	1	0.6	0.6	0	0	0
As	0.6	1	0.6	0	0	0
Aw	0.6	0.6	1	0	0	0.4
BSh	0	0	0	1	0.6	0
BWh	0	0	0	0.6	1	0
Cwa	0	0	0.4	0	0	1

Table 6 Regional semantic relatedness (regSR) with data collected from various climate zones in India

State code	Major climate types					
	Am	As	Aw	BSh	BWh	Cwa
AS	0	0	0.3	0.1	0.1	1
BR	0	0	0.3	0.1	0.1	1
CT	0.6	0.6	0.8	0.2	0.1	0.4
DL	0	0	0.1	0.6	0.5	0
GJ	0.3	0.3	0.3	0.6	0.6	0.1
KA	0.3	0.3	0.5	0.4	0.4	0.3
KL	1	0.6	0.5	0.2	0.2	0
MP	0.6	1	0.5	0.2	0.2	0
MH	0.4	0.4	0.6	0.5	0.4	0.2
MN	0	0	0.3	0.1	0.1	1
OR	0.6	0.6	0.8	0.2	0.1	0.4
RJ	0	0	0.1	0.5	0.5	0
TN	0.6	0.6	0.8	0.2	0.1	0.4
UP	0	0	0.3	0.1	0.1	1
WB	0.6	0.6	0.8	0.2	0.1	0.4

Comparative Study of Predictive Analytics

Tables 7, 8 evidently shows that the proposed semantically-enhanced theory-guided model has better prediction potential compared to other baselines. In respect of both MAE and RMSE, the proposed model is able to outperform all the other baselines in predicting recovered and confirmed case counts for 10–11 states from amongst the total 15 Indian states considered. More importantly, the superiority of the proposed model is prominent in case of those states (e.g. MH: Maharashtra) where the confirmed case counts are substantially high in recent days. This is so, primarily because our proposed model has implanted mechanism of following epidemiological development theory, which is completely ignored by the considered LR and NLR approaches. Accordingly, it can be well anticipated that for wide-ranging prediction over future several months our proposed model would be more appropriate than the others. Though SETG works as per theoretical guidance as well, our presently proposed model outperforms SETG with average 11% improvement (reduction) in prediction error.[4]. This demonstrates the effectiveness of enhancing our model by introducing the concept of *regional semantic relatedness* that can handle the presence of multiple climate types within each spatial region.

Our predicted count of daily new confirmed cases and new recovered cases from 18-Sep-2020 to 16-Nov-2020 are graphically presented in Figs. 7, 8. It is evident from the figures that the predicted values from our proposed semantics-driven theory-guided model match well with the observed values for both daily confirmed case and daily recovered case count, given the climate type of the region.

Impact Analysis Based on ANOVA Test and *Semantic-GI*

As mentioned earlier, in our proposed framework, the impact analysis is conducted using both statistical ANOVA test and *Semantic-GI* test (refer Table 9). The ANOVA test is performed to analyze the significance of correlation between the daily count of confirmed/recovered cases and the variability of the climate types. Besides, the *Semantic-GI* test is conducted to analyze the same with regard to the semantic relatedness in the climate types, where a region may contain more than one type of climate in its different parts. To handle the uncertainty issue, both the tests are carried out on the data predicted by our enhanced semantics-driven theory-guided model.

As indicated by small F-values and large p-values obtained from the ANOVA test (refer Table 9), both confirmed and recovered case counts are not significantly correlated with the climate variability over the various states or provinces.

However, with respect to the *Semantic-GI* measures and the associated Z-scores (refer Table 9), we can infer that the daily counts of both confirmed cases and recovered cases have significantly high semantic correlations with the

[4] By the term "improvement" here, we indicate relative percentage improvement, or in specific, percentage reduction in produced error in comparison with the baseline considered. For example, if the SETG model (as baseline) produces error value of e_{setg} and the pro-

Footnote 4 (continued)

posed model produces error value of e_{prop}, then the relative improvement is $\left(\frac{(e_{setg}-e_{prop})*100}{e_{setg}}\right)\%$.

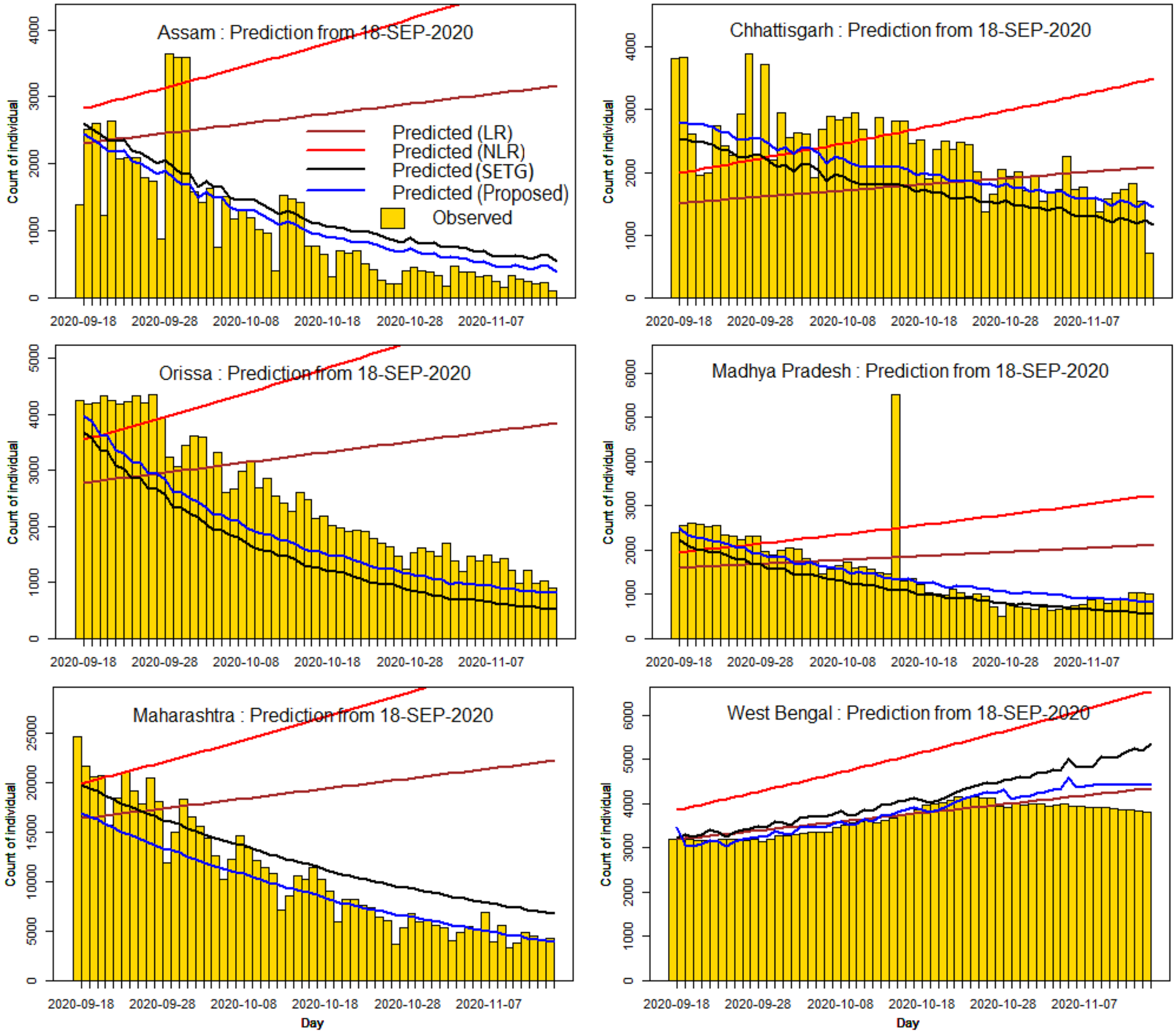


Fig. 7 Observed vs. Predicted count of daily new confirmed COVID-19 cases in some specific states in India

climate variations. As per the very high Z-scores, it can be claimed with 99% confidence that the semantically similar climate zones have very similar statistics for the daily confirmed/recovered case count. This answers the **RQ1**.

Subsequently, to answer **RQ2**, i.e. to understand the impact of region-specific climate, we perform in-depth analyses considering semantically averaged confirmed/recovered case count ($sAvg_i$) for each climate type separately (refer Fig. 9). From the figure we notice that, compared to the expected value (indicated by the red line), the daily confirmed case counts in BSh (hot semi-arid) and BWh (hot arid) type climate zones are substantially ($\approx$ 150%) high. However, confirmed case counts in humid-subtropical (Cwa) and tropical (Am, As, Aw) zones are not so high. This forms a significant cluster pattern which is also reflected by positive *Semantic-GI* with high Z-score. It can therefore be inferred that the arid/dry climate, attributed by very low humidity or little precipitation, is more vulnerable for COVID infection. A contrasting scenario can be observed for COVID-19 recovered case. For example, the daily relative-recovery in the tropical/equatorial climate zones (Am, As, Aw etc.) is extremely high, whereas that in the hot arid (BWh) and hot semi-arid (BSh) climate zones is quite low ($\approx$ 20%), compared to the expected value. Thus, the humidity is found to show a significant positive correlation with the recovery. In other words, the higher the humidity, the more the relative-recovery from COVID-19. Furthermore, as can be noted from the Fig. 9, the temperate climate zones (e.g. Cwa) are also quite vulnerable for severe COVID transmission, since unlike the tropical climate, the

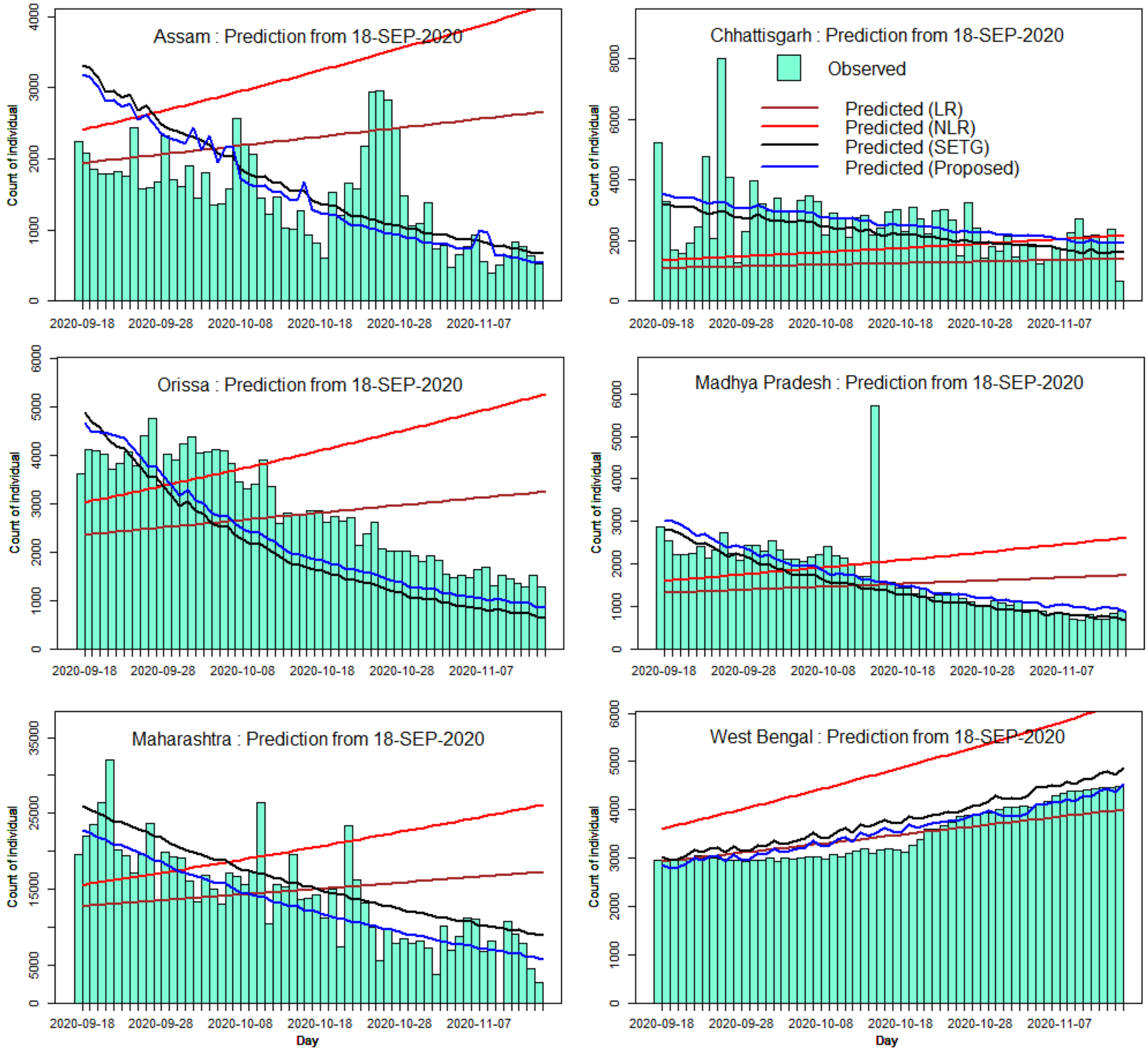


Fig. 8 Observed vs. Predicted count of daily new recovered COVID-19 cases in some specific states in India

relative-recovery in temperate zones is not too high (only 60% above the expected value) while the confirmed cases per million individual in these zones are prominently (around 85%) higher than the expected value. **Hence, based on the extent of the vulnerability for COVID-19 transmission, we can arrange our studied climate zones as follows:** $Tropical\{Am, As, Aw\} < Temperate\{Cwa\} < Arid/Semi\text{-}arid \{BSh, BWh\}$. This answers the **RQ2**.

The semantic weight matrix as used in our case study is depicted in Fig. 10a and the same is represented in terms of semantic neighborhood graph (with consideration to 7 selected states) in Fig. 10b, to help in better interpretation. It can be interestingly noted from this graph that though a pair of states (e.g. GJ: Gujarat and DL: Delhi) may not be treated as neighbors from spatial perspective, they can still become semantic neighbors of each other, if their semantic weight, i.e. the degree of semantic relatedness in their climate types, is non-zero.

Significance of Our Research Outcomes

Overall, our research provides insights into climatological impact on infection as well as recovery from the novel coronavirus disease. Our semantically enhanced theory-guided analyses reveal that the regions which belong to dry climate are most susceptible for infection on everyday basis. At the

Table 7 Comparative study of performance regarding confirmed case count prediction

State Code	MAE of the models				RMSE of the models			
	LR	NLR	SETG	Proposed	LR	NLR	SETG	Proposed
AS	1935.49	2983.94	440.8	**326.32**	2121.11	3282.61	570.9	**522.29**
BR	1773.8	2772.72	603.78	**569.64**	1870.92	2922.27	665.69	**630.88**
CT	738.05	903.13	568.78	**401.88**	893.23	1162.11	667.28	**497.48**
DD	1315.01	1113.54	1288.03	**1280.08**	1879.93	**1399.39**	1862.16	1800.84
GJ	464.32	921.1	429.31	**289.05**	550.02	1027.05	485.64	**336.73**
KA	4003.84	7513.58	1755.75	**1628.61**	5105.5	9073.32	2311.23	**2029.25**
KL	4311.25	3196.26	1717.29	**1347.9**	4694.13	3678.28	2077.71	**1728.74**
MP	807.47	1333.6	454.8	**251.62**	983.83	1562.59	726.15	**595.01**
MH	10462.63	17578.48	2382.59	**1693.54**	12089.25	20020.84	2745.28	**2234.12**
MN	82.05	**58.05**	98.83	69.97	101.15	**71.78**	113.32	91.5
OR	1512.37	2782.41	896.68	**593.08**	1746.95	3263.77	936.71	**652.03**
RJ	**265.99**	516.86	1077.98	980.06	**319.19**	663.79	1176.32	1072.63
TN	4291.93	6791.28	2026.75	**1957.42**	4800.4	7433.35	2136.09	**2063.38**
UP	3479.5	6217.64	**387.49**	436.22	3839.44	6721.56	**507.25**	552.22
WB	**211.54**	1633.57	547.88	338.47	**329.62**	1769.06	762.67	600.8

The bold values indicate the best prediction performances attained by any model, for a given state/province

Table 8 Comparative study of performance regarding recovered case count prediction

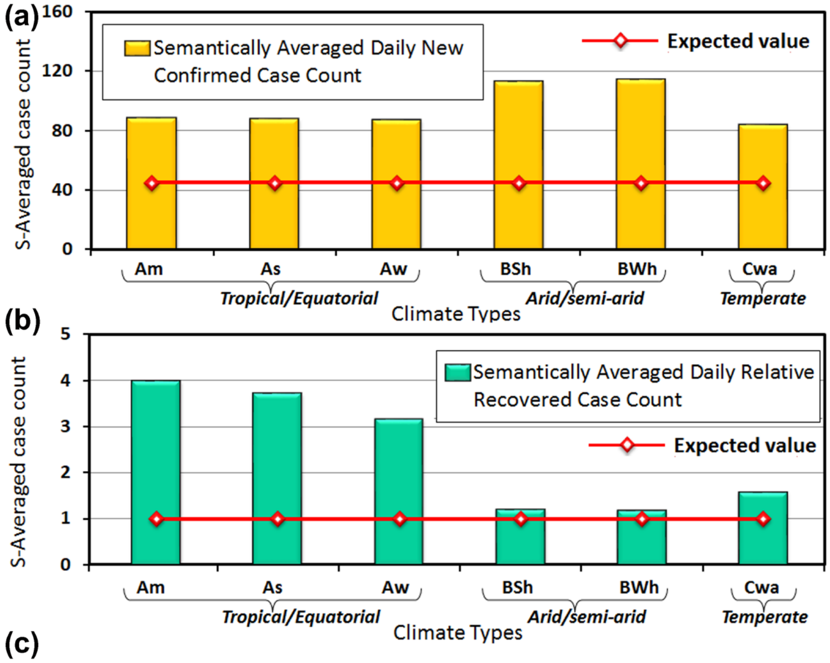
State Code	MAE of the models				RMSE of the models			
	LR	NLR	SETG	Proposed	LR	NLR	SETG	Proposed
AS	1012.45	1934.22	603.1	**582.24**	1238.17	2219.91	791.35	**777.55**
BR	1599.22	2730.12	743.59	**743.41**	1694.27	2879.52	839.75	**837.1**
CT	1555.54	1151.43	748.36	**666.94**	1901.66	1577.51	1052.13	**975.87**
DD	1316.05	1044.28	1102.08	**1035.32**	1768.04	1560.41	1577.37	**1490.44**
GJ	411.4	771.3	465.47	**391.87**	504.67	888.87	520.13	**495.34**
KA	**2290.76**	4599.57	2894.49	3202.58	**3173.07**	5989.39	3422.7	3790.83
KL	4361.27	3448.6	1785.37	**1521.97**	4704.23	3803.99	1959.68	**2057.84**
MP	751.67	974.18	385.68	**251.31**	965.65	1147.33	683.96	**612.9**
MH	6491.79	9877.93	3615.85	**2762.47**	7887.57	12082.24	4426	**3824.2**
MN	94.44	**78.11**	104.12	101.61	169.29	**147.42**	172.06	168.37
OR	1170.91	1865.89	939.2	**738.9**	1330.54	2263.87	987.66	**785.23**
RJ	441	**374.96**	1018.03	1070	524.02	**518.2**	1118.78	1173.75
TN	3555.96	6338.53	**2375.31**	2428.69	4049.04	6957.31	**2582.13**	2628.19
UP	2259.24	4216.58	**472.14**	506.21	2588.59	4916.67	**633.66**	658.96
WB	238.73	1458.76	325.47	**150.51**	269.53	1494.36	362.45	**196.82**

The bold values indicate the best prediction performances attained by any model, for a given state/province

same time, we also find that the daily relative-recovery in dry regions is quite unfavorable. Accordingly, there remains huge scope to more effectively control the pandemic scenario in India by not only imposing stronger isolation measures but also improving the health-care facilities in the dry/arid regions (e.g. Maharashtra, Rajasthan, Delhi, Gujarat etc.). Though the promising recovery pattern in tropical/equatorial regions indicates that an intense isolation/quarantine measure can enough help controlling the COVID-19 pandemic in the Indian states like West Bengal, Orissa, Tamil Nadu, Chhattisgarh, etc., the upcoming winter season (during December, January, February) can become vulnerable, since winter is dry in these states. To combat COVID-19 outbreak in India, additional care must also be given towards strengthening the health-care infrastructures in the states having temperate climate, such as Assam, Bihar etc., since the relative-recovery in temperate climate is found to be quite low compared to the infection.

To be noted, our proposed impact analysis based on *Semantic-GI* is more meaningful than that achieved with respect to ANOVA test. This is so, because the consideration of semantic knowledge effectively handles the uncertainty

Fig. 9 Assessment for specific impact of climate variability on daily confirmed/recovered case count



CLIMATE TYPE	MAIN FEATURES	DAILY CONFIRMED CASE	DAILY RELATIVE-RECOVERED CASE
Am	Tropical climate, hot temperature, significant annual precipitation	High (≈ 95%)	Extremely High (>200%)
As	Tropical climate, hot temperature, abundant annual precipitation	High (≈ 95%)	Extremely High (>200%)
Aw	Tropical climate, hot temperature, abundant annual precipitation	High (≈ 95%)	Extremely High (>200%)
BSh	Hot semi-arid climate, little precipitation	Very High (≈ 150%)	Low (≈ 20%)
BWh	Hot desert climate (arid), little precipitation	Very High (≈ 150%)	Low (≈ 20%)
Cwa	Temperate climate, monsoon-influenced, hot-summer	High (≈ 85%)	High (≈ 60%)

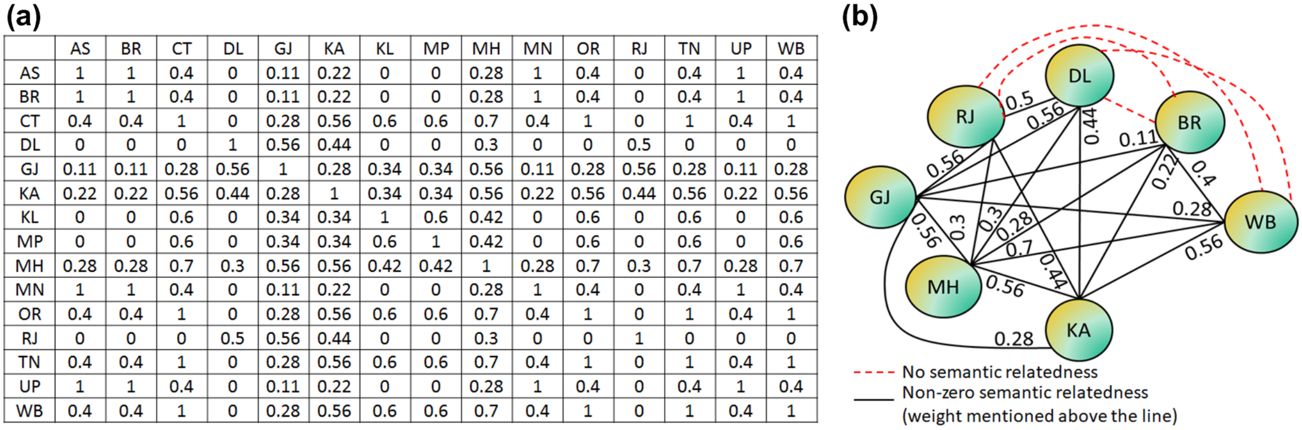
Table 9 Summary of correlation analyses: climate variability vs. development of COVID-19 cases (confirmed and recovered)

COVID-19	ANOVA Test		Semantic-*GI* Test	
Cases	*F* value	Pr(> *F*)	Semantic-*GI*	*Z*-score
Confirmed	1.91	0.18	0.067	2.580
Recovered	1.80	0.20	0.052	2.580

in the data which emerges due to unawareness about the other influencing factors that may also affect the COVID-19 spread within each state or province. The consideration of data from multiple states having semantically related climate, can indirectly neutralize the impact of these unknown or hidden factors to certain extent. Thus, the *Semantic-GI* test helps us achieving more robust outcome of impact analysis.

Related Works

The recent researches regarding climatological effect on COVID-19 outbreak dynamics can be split into two separate groups based on the respective research conclusions. In the research works by the first group, at least one of the



(a)

	AS	BR	CT	DL	GJ	KA	KL	MP	MH	MN	OR	RJ	TN	UP	WB
AS	1	1	0.4	0	0.11	0.22	0	0	0.28	1	0.4	0	0.4	1	0.4
BR	1	1	0.4	0	0.11	0.22	0	0	0.28	1	0.4	0	0.4	1	0.4
CT	0.4	0.4	1	0	0.28	0.56	0.6	0.6	0.7	0.4	1	0	1	0.4	1
DL	0	0	0	1	0.56	0.44	0	0	0.3	0	0	0.5	0	0	0
GJ	0.11	0.11	0.28	0.56	1	0.28	0.34	0.34	0.56	0.11	0.28	0.56	0.28	0.11	0.28
KA	0.22	0.22	0.56	0.44	0.28	1	0.34	0.34	0.56	0.22	0.56	0.44	0.56	0.22	0.56
KL	0	0	0.6	0	0.34	0.34	1	0.6	0.42	0	0.6	0	0.6	0	0.6
MP	0	0	0.6	0	0.34	0.34	0.6	1	0.42	0	0.6	0	0.6	0	0.6
MH	0.28	0.28	0.7	0.3	0.56	0.56	0.42	0.42	1	0.28	0.7	0.3	0.7	0.28	0.7
MN	1	1	0.4	0	0.11	0.22	0	0	0.28	1	0.4	0	0.4	1	0.4
OR	0.4	0.4	1	0	0.28	0.56	0.6	0.6	0.7	0.4	1	0	1	0.4	1
RJ	0	0	0	0.5	0.56	0.44	0	0	0.3	0	0	1	0	0	0
TN	0.4	0.4	1	0	0.28	0.56	0.6	0.6	0.7	0.4	1	0	1	0.4	1
UP	1	1	0.4	0	0.11	0.22	0	0	0.28	1	0.4	0	0.4	1	0.4
WB	0.4	0.4	1	0	0.28	0.56	0.6	0.6	0.7	0.4	1	0	1	0.4	1

Fig. 10 Semantic weight matrix for the considered states (**a**), and graphical illustration for semantic neighborhood (**b**)

climatic factors, such as humidity, minimum temperature, average temperature, etc., has been identified to be significantly correlated with COVID-19 pandemic, whereas, the second group of researches has not found any such evidence in this regard.

Regarding the first group of research, the works of Pani et al. [15], Bashir et al. [3], Liu et al. [13], Auler et al. [1], Ward et al. [25], and Tosepu et al. [24] are worth mentioning. Interestingly, though in a generic sense all these research works notice some association between COVID outbreak and climatic variables, the specific results are not very identical. For example, using Spearman and Kendall rank correlation tests, Pani et al. [15] have found the temperature, dew point, and humidity to be significantly and positively associated with COVID-19 transmission. Contrarily, by employing Spearman correlation measure, Ward et al. [25] have noticed a significant negative association between relative humidity and novel coronavirus transmission. Moreover, they have found no association with temperature. The works of Bashir et al. [3], and Tosepu et al. [24], who identified average temperature to be one of the climatic factors influencing COVID-19 spread, are therefore, contradicting with the work of Ward et al. [25]. The outcomes of the relevant researches done by Liu et al. [13] and Auler et al. [1] are also quite inconsistent. In the former work, primarily using nonlinear regression model, the authors noted that low humidity, low temperature, and mild diurnal temperature range were possibly favorable for COVID-19 transmission. Contrarily, in the latter case, based on a combination of linear regression and multivariate statistical analysis, the authors noticed that higher mean temperatures and average relative humidity can also favor the transmission. The key limitation in these works remain in their purely data-driven approaches that ignore the physical understanding of infectious disease dynamics. Though our previously introduced SETG model [9] takes into account the theoretical principles of epidemic development, it has its own limitations in real-world application scenario, since it does not take into account the presence of multiple climate types within a region.

The second group of researches are primarily based on either theoretical models or data-driven models with nonlinear analysis. For example, with the help of pandemic simulation using SIRS (Susceptible-Infected-Rec-overed-Susceptible) model, Baker et al. [2] found that the summer weather would not substantially limit pandemic growth. This research observation also conforms to the findings of Zhu et al. [26] and Briz et al. [4], who employed generalized additive model and approximated Bayesian inference technique to serve the purpose. However, recent research also indicates that these results are highly sensitive to uncertainty underlying the data.

Existing work vs. Proposed Approach As per the findings, our semantically-enhanced theory-guided research primarily belongs to the first group. However, in contrast to majority of those works, we consider the theoretical guidance as well. Moreover, our enhanced semantics-driven theory-guided analysis primarily utilizes the *overall climate patterns* of the regions, rather than considering individual climate factors. Based on our research outcomes, we find that the *dry/arid* and *semi-arid* climate zones are most vulnerable for the increasing infection from COVID-19, followed by the *temperate* climate zones. Our observation on arid/semi-arid climate and temperate climate are supported by the works of Liu et al. [13] and Auler et al. [1], respectively. Additionally, the present work also reveals a significantly positive correlation of humidity with the daily relative-recovery from this disease, which eventually can help making administrative decisions to effectively control COVID-19 transmission on regional basis.

Conclusions

Motivated by the semantically-enhanced theory-guided framework as introduced in [9], in this paper, we have proposed an improved data-driven model to provide a more realistic analysis of how regional climate pattern impacts on the COVID-19 outbreak. Novelty of this work is primarily embedded in the following three aspects: (1) introducing the concept of "regional semantic average" to account for the relatedness of data from the same climate zone expanded over multiple spatial regions; (2) enhancing interpretation of the causal relationship between climate variability and COVID case development, considering semantic relatedness of the data on regional basis; and (3) upgrading the impact analysis with consideration to the expected values of relative-recovered cases and per million new infected/confirmed cases over the various regions. Consideration of regional semantic relatedness at the time of learning causal relationship between climate variability and COVID-19 outbreak not only helps to deal with the underlying uncertainty but also enables us to better assess the climatological impact on the development of infected and recovered cases of the disease on regional basis. Moreover, the theoretical guidance from the epidemiological model helps our model in attaining a generalizable solution. At the end of the study we find that both arid/semi-arid and temperate climate are evidently susceptible to COVID-19 transmission. We also observe that humid climate positively influences the recovery from this novel corona virus disease in India.

Ample scopes remain in further upgrading the framework with added knowledge on genetic aspects of the virus, and also, in exploring the impact of other factors. It may be noted that, though our proposed framework has been illustrated with respect to analyzing impact of climate variability on COVID-19 outbreak, it can also be extended easily for semantics-driven theory-guided analyses in various other domains, including bio-medical science, material science, quantum chemistry etc., by incorporating appropriate domain knowledge.

Acknowledgement We would like to acknowledge support from INSPIRE Faculty Fellowship Research Grant [DST/INSPIRE/04/2019/001670 to M.D.] by the Department of Science and Technology, Govt. of India.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Auler A, Cássaro F, da Silva V, Pires L. Evidence that high temperatures and intermediate relative humidity might favor the spread of COVID-19 in tropical climate: a case study for the most affected Brazilian cities. Sci Total Environ. 2020;729:139090.
2. Baker RE, Yang W, Vecchi GA, Metcalf CJE, Grenfell BT. Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic. Science. 2020;369:315–9.
3. Bashir MF, Ma B, Komal B, Bashir MA, Tan D, Bashir M, et al. Correlation between climate indicators and COVID-19 pandemic in New York, USA. Sci Total Environ. 2020;728:138835.
4. Briz-Redón Á, Serrano-Aroca Á. A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. Sci Total Environ. 2020;728:138811.
5. Das M, Ghosh SK. Measuring Moran's I in a cost-efficient manner to describe a land-cover change pattern in large-scale remote sensing imagery. IEEE J Sel Top Appl Earth Observ Remote Sens. 2017;10(6):2631–9.
6. Das M, Ghosh SK. semBnet: a semantic Bayesian network for multivariate prediction of meteorological time series data. Pattern Recognit Lett. 2017;93:192–201.
7. Das M, Ghosh SK. Reducing parameter value uncertainty in discrete Bayesian network learning: a semantic fuzzy Bayesian approach. IEEE Trans Emerg Top Comput Intell. 2019;5(3):361–72.
8. Das M, Ghosh SK. Enhanced Bayesian network models for spatial time series prediction. Cham, Switzerland: Springer; 2020.
9. Das M, Ghosh SK. Analyzing impact of climate variability on COVID-19 outbreak: a semantically-enhanced theory-guided data-driven approach. In: Proceedings of the 8th ACM India Joint International Conference on data science and management of data. 2021. p. 1–9. https://www.isical.ac.in/~monidipa_t/Papers/SETG.pdf. Accessed 15 Aug 2021.
10. Gasse M, Aussem A, Elghazel H. A hybrid algorithm for Bayesian network structure learning with application to multi-label learning. Expert Syst Appl. 2014;41(15):6755–72.
11. Gupta N, Ari S, Panigrahi N. Change detection in landsat images using unsupervised learning and rbf-based clustering. IEEE Trans. Emerg. Top. Comput. Intell. 2019;5(2):284–97.
12. Kottek M, Grieser J, Beck C, Rudolf B, Rubel F. World map of the Köppen-Geiger climate classification updated. Meteorol Zeitschrift. 2006;15(3):259–63.
13. Liu J, Zhou J, Yao J, Zhang X, Li L, Xu X, He X, Wang B, Fu S, Niu T, et al. Impact of meteorological factors on the COVID-19 transmission: a multicity study in China. Sci Total Environ. 2020;726:138513.
14. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinform. 2006;7:S7.
15. Pani SK, Lin NH, RavindraBabu S. Association of COVID-19 pandemic with meteorological parameters over Singapore. Sci Total Environ. 2020;740:140112.
16. Park K. Park's text book of preventive and social medicine. Jabalpur, India: Banarsidas Bhanot Publishers; 2015.
17. Pham QV, Nguyen DC, Hwang WJ, Pathirana PN, et al. Artificial intelligence (ai) and big data for coronavirus (COVID-19) pandemic: a survey on the state-of-the-arts. IEEE Access. 2020.
18. Planning Commission, G.o.I.: Census 2011 (Final Data) - Demographic details, Literate Population (Total, Rural and Urban). planningcommission.gov.in. Planning Commission, Government of India. 2019.
19. Postnikov EB. Estimation of COVID-19 dynamics "on a back-of-envelope": does the simplest SIR model provide

quantitative parameters and predictions? Chaos Solitons Fract. 2020;135:109841.

20. Rustam F, Reshi AA, Mehmood A, Ullah S, On B, Aslam W, Choi GS. Covid-19 future forecasting using supervised machine learning models. IEEE. Access. 2020;8:101489–99.
21. Storlie CB, Lane WA, Ryan EM, Gattiker JR, Higdon DM. Calibration of computational models with categorical parameters and correlated outputs via Bayesian smoothing spline anova. J Am Stat Assoc. 2015;110(509):68–82.
22. Thomas DM, Sturdivant R, Dhurandhar NV, Debroy S, Clark N. A primer on COVID-19 mathematical models. Obesity. 2020;28(8):1375–7.
23. Ting DSW, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. Nat Med. 2020;26(4):459–61.
24. Tosepu R, Gunawan J, Effendy DS, Lestari H, Bahar H, Asfian P, et al. Correlation between weather and COVID-19 pandemic in Jakarta, Indonesia. Sci Total Environ. 2020;725:138436.
25. Ward MP, Xiao S, Zhang Z. Humidity is a consistent climatic factor contributing to SARS-CoV-2 transmission. Transbound Emerg Dis. 2020;67(6):3069–74.
26. Zhu Y, Xie J. Association between ambient temperature and COVID-19 infection in 122 cities from China. Sci Total Environ. 2020;724:138201.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.