International Journal of Community Medicine and Public Health
Singh G et al. Int J Community Med Public Health. 2023 Jan;10(1):148-153
http://www.ijcmph.com
pISSN 2394-6032 | eISSN 2394-6040

DOI: https://dx.doi.org/10.18203/2394-6040.ijcmph20223288

Original Research Article

Development and use of open-source algorithms for space-time emerging hotspot analysis of routine dengue NVBDCP data in Punjab, India

Gurpreet Singh[1]*, Biju Soman[1], Gagandeep Singh Grover[2]

[1]Achutha Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, Kerala, India
[2]Department of Health and Family Welfare, Government of Punjab, India

Received: 18 November 2022
Revised: 01 December 2022
Accepted: 03 December 2022

***Correspondence:**
Dr. Gurpreet Singh,
E-mail: drgurpreet.md.afmc@gmail.com

ABSTRACT

Background: Understanding spatiotemporal epidemiology using open-source and reproducible algorithms add value to routine health information systems. Objectives were to estimate spatial clustering, identify spatial clusters and space-time hotspots of dengue.
Methods: Queen's contiguity neighborhood matrix and row-standardized spatial weights were used. Spatial clustering was estimated using Moran's I. Local Moran's I with sensitivity analysis at 0.01, 0.05, and 0.1 significance levels were performed. The space-time cube model was developed. Gi* statistic and seasonal Mann Kendal test identified persistent and intensifying, persistent, persistent and diminishing, emerging, oscillating, new, historical, and sporadic hotspot sub-districts. Analysis was carried out using R version 4.1.0.
Results: The expected Moran's value was -0.00671. Significant spatial clustering was observed annually in 2016-2018 ($p<0.01$, <0.01, and 0.04, respectively) and was most common in August, followed by July and November. High-high, high-low, low-low, and low-high sub-district clusters were identified between Aug-Dec from 2015-19. Sensitivity analysis highlighted the core and spread of spatial clusters. Faridkot and Muktsar blocks/ sub-districts were persistent and intensifying hotspots.
Conclusions: Spatial clusters were dynamic in space and time. The development of open-source algorithms provides a reproducible and scalable platform for future research and evidence for informed decision-making by public health managers.

Keywords: Spatial correlation analysis, Emerging hotspot analysis, Space-time cube, Routine data, Data science, Dengue

INTRODUCTION

Methods and applications in spatial data analytics have evolved in health sciences in recent decades. They have become an essential tool for epidemiologists for tracking infectious diseases, outbreak analysis, disease surveillance, emergency preparedness and response, environmental health, chronic disease prevention, and community health assessment.[1] These advances provide newer opportunities for public health administrators to plan, analyze, allocate resources, and manage health systems.[2]

Dengue is a viral mosquito-borne disease with more than half of the global population at risk of acquiring the infection through the bite of infected female *Aedes* mosquitoes.[3] It is highly prevalent in tropical and subtropical regions, and the Asian subcontinent

contributes substantially to the global burden. National vector borne disease control programme, India (NVBDCP) reported more than eight lakh dengue cases in the past decade, with a median annual incidence of 6.57 per lakh population. The incidence of dengue was highest in 2019 and 2017 (11.80 and 11.55 per lakh), and the highest median annual incidence of dengue was observed in the state of Punjab (24.49 per lakh). NVBDCP, the nodal programme for controlling vector-borne diseases in the country, maintains statistics on dengue occurrence across states as a part of routine health information systems (RHIS).[4]

There is no specific treatment, and it currently lacks large-scale effective vaccines to prevent and control the increasing dengue burden in India.[3] Understanding the Spatio-temporal patterns of dengue provides insights for estimating patterns of the disease and its association with risk factors in the local context, and thus has the potential for the development of forecasting models for optimizing resource allocation and planning effective vector control measures in low-and middle-income countries. Data science is an interdisciplinary science with utilities like exploratory data analysis, which enables data handling of routine datasets for creating analysis-ready tidy datasets through tools for wrangling, transformation, exploration, etc.[5,6] Further, open-source platforms provide scalable and reproducible algorithms for future use by researchers, epidemiologists, public health managers, and others.

Advancements in open-source geographical information system (GIS) platforms, data handling techniques, and computational resources enable the estimation of spatial clustering and identification of spatial clusters in space and time. Multiple statistical methods have been used for global estimates of spatial clustering depending on the type of data and disease under consideration. Most commonly used methods include Moran's I, Oden's Ipop, Geary's contiguity ratio, and Tango's excess event test for areal spatial data; Edward's K nearest neighbors, Ripley's K function, and Knox test for point spatial data. Similarly, as local indicators to identify spatial clusters, Local Moran's I, Getis-Ord's local Gi, and Gi* statistics for areal data and scan circles (based on Openshaw's, Besag-Newell, Turnbull, and Kulldorf's scan statistics) for point spatial datasets have been used.[7,8] Recently, emerging space-time hotspot analyses have been introduced to understand disease patterns. This looks at the space-time perspective as a modeled cube wherein spatial analysis is carried out for each time slice, and trend analysis of each spatial feature is carried out over time.[9] However, earlier studies have used proprietary software, and to the best of our knowledge, there is a lack of open-source algorithms for the same.

Therefore, the present study was carried out to estimate the spatial clustering of dengue in the state of Punjab, identify spatial clusters (sub-districts), develop an open-source algorithm for emerging space-time hotspot analysis, and provide empirical evidence using the dengue RHIS dataset.

METHODS

The study design of the present study was ecological study with a data science approach. The study included spatial autocorrelation analysis and space-time emerging hotspot analysis of secondary data. The anonymized line list of lab-confirmed dengue cases reported by NVBDCP, Punjab, from 2015-19, was analyzed. The line listing data included cleaned, standardized, and geocoded addresses. Point in polygon analysis was carried out, and population projections based on census 2011 were calculated to estimate monthly dengue incidence rates at the sub-district level. The spatial file of the sub-district (Block) multi-polygons was obtained from Punjab remote sensing authority. The inclusion criteria for the present study were lab-confirmed dengue cases reported by directorate health services. Those records where geocoded addresses and testing dates missing after pre-processing of raw data were excluded from the study. Total number of reported lab-confirmed cases during the study period was 64,454. After excluding cases with no location/ time details, 63,741 cases (98.8%) were included in study for analysis.

Spatial autocorrelation analysis

The neighborhood matrix was constructed based on areal units with contiguous boundaries. The neighborhood was defined according to the queen's contiguity wherein the sub-districts were neighbors when they had at least one shared boundary point. The row-standardized spatial weights were calculated. Annual and monthly spatial autocorrelation analyses were carried out. Moran's I statistic provided global estimates for spatial clustering, and Local Moran's I was calculated to identify spatial clusters. Sub-district categorized in LISA quadrants of high-high, high-low, low-low, low-high, and unclassified based on the dengue incidence in the sub-district under consideration and its neighbors as compared to the global estimates. Sensitivity analyses at 0.01, 0.05, and 0.1 levels of significance were carried out to determine the core and spread of spatial clusters at a given point in time.

Space-time emerging hotspot analysis

A space-time cube model was constructed wherein the space was defined as areal units (sub-districts) and time as monthly intervals. For each sub-district, a monthly time series was constructed, and trend analysis was performed using the seasonal Mann-Kendall trend test. A trend was said to be significantly positive when the z score was positive and the $p \leq 0.05$. For each month in the monthly time series, Getis Ord Gi* statistic was calculated for all sub-districts. The sub-districts were considered as hotspot/ coldspot for the respective month when the calculated z score was $\geq 1.96/\leq 1.96$, respectively. Based on the space-time patterns, sub-districts were categorized into eight categories.

The definitions for space-time classification were adapted from ArcGIS literature for emerging hotspot analysis.[10] A sub-district was categorized as a 'persistent' hotspot when it has been a hotspot for at least a month for four years between 2015-19, 'persistent and intensifying'/'persistent and diminishing' when a significant positive/ negative time trend was present in addition to above respectively. Sub-districts that were hotspots only in recent years (at least twice in 2017-19) were categorized as 'emerging,' those that were hotspots earlier but have ceased to be in recent years as 'historical', hotspots in 2019 but never earlier as 'new', hotspot only once as 'sporadic', a hotspot in on-off fashion as 'oscillating' and others as 'not categorized'.

Permissions and clearances

Present study was a part of a more extensive study being undertaken as a Ph.D. project by the first author. Ethical approval was obtained from institutional ethics committee (IEC/IEC-1653; IEC reg No. ECR/189/Inst/KL/2013/RR-16), and study has been registered on clinical trials registry of India (CTRI/2021/01/030245). Permission from the directorate of health services, Punjab and Punjab remote sensing Authority obtained. Detailed study protocol and algorithms for pre-processing datasets using reproducible open-source algorithms have been published elsewhere.[11-13]

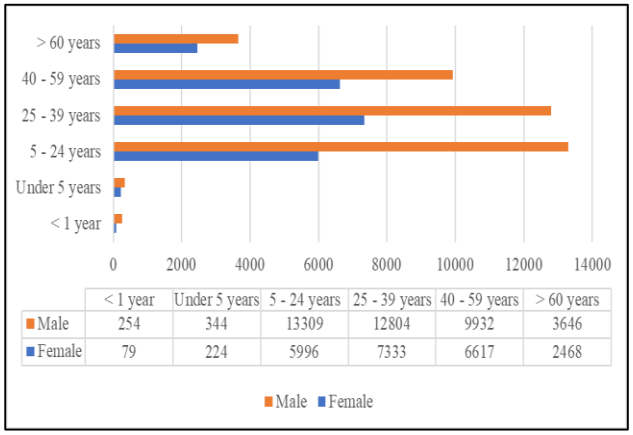
Statistical software

The analysis was carried out using R version 4.1.0 and included the use of tidyverse, lubridate, sf, spdep, rgeoda, and Kendall packages.

RESULTS

Age and sex distribution of reported cases

The age and sex distribution of reported cases are presented as Figure 1. Most cases were males (64%) and in the age group of 25-39 years (32%).



	< 1 year	Under 5 years	5 - 24 years	25 - 39 years	40 - 59 years	> 60 years
Male	254	344	13309	12804	9932	3646
Female	79	224	5996	7333	6617	2468

Figure 1: Age and sex distribution of reported dengue cases.

Neighborhood matrix

The neighborhood matrix features of the sub-district spatial file are represented in Table 1. The matrix was symmetrical without isolated areal units. There were 150 areal units, with the majority (37) sub-districts having five neighbors. The number of links varied from one to nine, and the mean number of links was 5.28.

Table 1: Neighborhood matrix features.

Neighborhood matrix features	Characteristics
Number of regions (sub-districts)	150
Number of nonzero links	792
Percentage nonzero weights	3.52
Average number of links	5.28
Minimum number of links	01
Sub-districts with minimum number of links	02 (Bamial and Sardulgarh)
Maximum number of links	09
Sub-districts with maximum number of links	02 (Ludhiana-I and Batala)

Spatial clustering of dengue

Global Moran's I was found to be statistically significant at multiple occasions during the study period. The space-time heatmap for the presence of spatial clustering is represented in Figure 2. It was observed that the spatial clustering was predominantly during the seasonal onset and waning periods for dengue occurrence in July-Aug and November, respectively. The duration of significant spatial clustering for a given annual timestamp was dynamic and varied across the years. The shortest duration of spatial clustering of dengue incidence was observed in 2019 (July-August), followed by the year 2015 (June-August and November) and 2016 (August-November). An on-off pattern of spatial clustering was observed in the year 2018.

Spatial clusters of dengue

LISA statistics provided evidence for the dynamic nature of both spatial clusters (high-high and low-low) and spatial outliers (high-low and low-high) across time. The sensitivity analysis enabled the identification of the core and spread of the clusters. The sensitivity analysis results at 0.01, 0.05, and 0.1 level of significance as a local significance map are represented in Figure 3. The patterns showed the highest density of high-high (Sirhind, Majri, and Sanaur), low-low (Batala, Ajnala, and Shahkot), low-high (Shambu Kalan, Khera, and Morinda), and high-low (Taran-Taran, Bathinda, and Verka) spatial clusters for dengue occurrence across sub-districts.

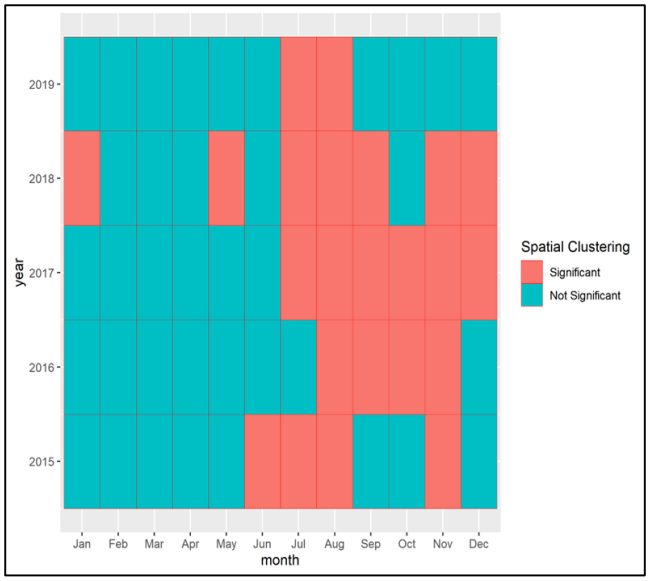


Figure 2: Space-time heatmap of global estimates for spatial clustering (Moran's I).

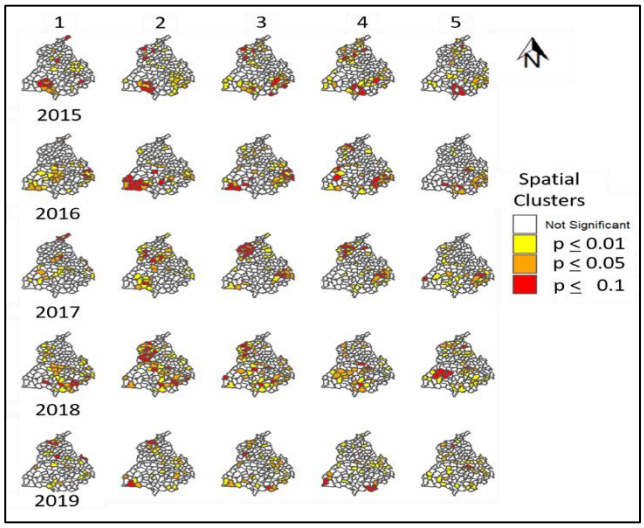


Figure 3: Local significance map for sensitivity analysis (Local Moran's I).

Space-time emerging hotspot analysis

Findings from space-time emerging hotspot analysis at monthly intervals are represented in Figure 4. It was observed that the numbers of hotspots were more persistent in the southwestern and south-eastern region; however, the sub-districts on the western border showed the presence of hotspots in recent years during the study period.

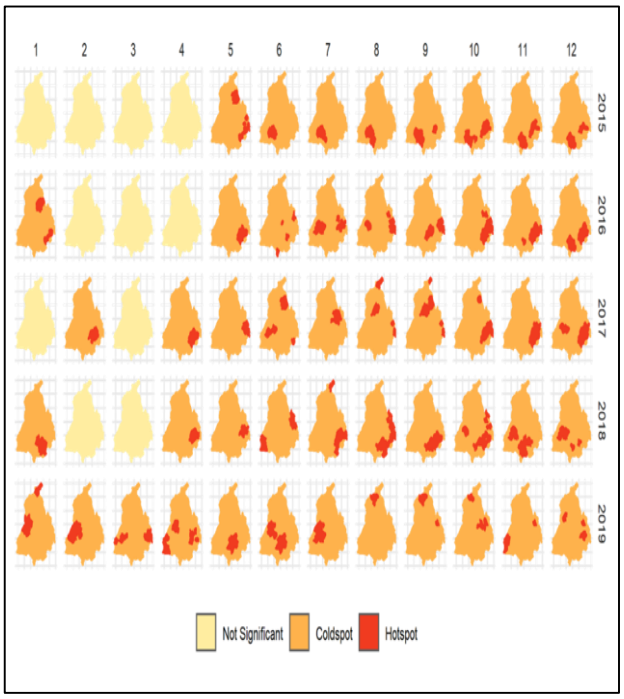


Figure 4: Space-time hotspot analysis grid (Gi* statistic).

The details of the identified hotspots are represented in Table 2. A mixed pattern of sub-districts belonging to all categories of emerging hotspot analyses were found to be existing in the state. The majority (n=27) of sub-districts were sporadic hotspots, followed by persistent hotspots (n=21), and new hotspots (n=21).

Table 2: Classification of sub-districts.

Hotspot category	Frequency, n=150 (%)	Sub-districts
Persistent and intensifying	02 (1.33)	Faridkot and Muktsar
Persistent	21 (14.00)	Amloh, Bagha Purana, Balachaur, Bassi Pathanan, Bhawanigarh, Bhikhi, Derabassi, Dhuri, Jaitu, Kharar, Khera, Kot Kapura, Majri, Morinda, Nabha, Patiala, Rajpura, Rampura, Samana, Shambu Kalan, and Sirhind
Persistent and diminishing	02 (1.33)	Ghanaur and Sanaur
Emerging	14 (9.33)	Abohar, Arniwal, Aur, Dhar Kalan, Fazilka, Gharota, Khamanon, Khuian Sarwar, Ludhiana II, Narot Jaimal Singh, Pathankot, Sangrur, Saroya, and Sujanpur
Oscillating	10 (6.67)	Adampur, Barnala, Bhogpur, Budhlada, Hoshiarpur I, Hoshiarpur II, Jhunir, Machhiwara, Mansa, and Maur

Continued.

Hotspot category	Frequency, n=150 (%)	Sub-districts
New	21 (14.00)	Batala, Dera Baba Nanak, Dhariwal, Dina Nagar, Fatehgarh Churian, Firozpur, Ghall Khurd, Gurdaspur, Guru Har Sahai, Jalalabad, Kalanaur, Khanna, Mahal Kalan, Makhu, Mamdot, Moga II, Naushera Pannuan, Patti, Sehna, Valtoha, and Zira
Historical	06 (4.00)	Bhunga, Jalandhar East, Jalandhar West, Malerkotla, Nawan Shahr, and Tanda
Sporadic	27 (18.00)	Ahemdagarh, Anandpur Sahib, Andana, Banga, Bathinda, Bhunarheri, Chamkaur Sahib, Dhilwan, Dirba, Garh Shankar, Goiniana, Kapurthala, Kot Bhai, Gidderbaha, Lehra Gaga, Nakodar, Nathana, Nurpur Bedi, Patran, Phagwara, Rupnagar, Sangat, Sardulgarh, Sher Pur, Sudhar, Sultanpur Lodhi, Sunam, and Talwandi Sabo

DISCUSSION

The present study documents evidence of the potential of routine health data to understand spatiotemporal patterns and the development of open-source algorithms for space-time emerging hotspot analysis using dengue line listing from NVBDCP, Punjab, as empirical datasets. Integrating data science approaches in routine health information systems is expected to help health authorities to enhance dengue preventive strategies and develop public health interventions targeted for the identified cluster areas.[2,14]

Adoption of geographic information systems accelerated rapidly in United States after the 1970s due to availability of open data-sharing platforms.[1] The launch of the open data initiative by the government of India, WHO-IDSP recommendations on increased use of GIS platforms, national digital health mission and other related initiatives in India provide opportunities for researchers and administrators to collaborate for understanding disease epidemiology and to incorporate advances in Spatio-temporal epidemiology, for resource allocation.[15-17]

The emergence of high-low spatial outliers at the seasonal onset was seen in predominantly urban areas surrounded by rural sub-districts, followed by the subsequent spread of high dengue incidence. Also, the dynamic shift of high dengue incidence across sub-districts was observed over time. In a study carried out in Delhi, India, Olivier et al described and compared a similar phenomenon to a 'forest fire' signature wherein the spread of dengue occurs rapidly and a local cluster before being burned out seed adjacent areas, even at a location beyond the flight range for the mosquito dispersal.[18] Similar findings in the present study strengthen evidence for institutionalization of GIS into health care to develop decision support systems. Also, such findings urge future research work incorporating non-health sector routine data for understanding nuances of associated environmental, climatic, socio-demographic and health system factors for risk-based resource allocation in health care which is a hierarchical system with multiple decision makers.[19]

It is essential to understand that the selection of the operational definition for creating the neighborhood matrix should be based on the transmission patterns and epidemiology of the disease under investigation.[8] The present study created a neighborhood matrix using Queen's contiguity. Dengue, being a mosquito-borne disease, and considering the mobility patterns of the population, the areal units sharing even a single boundary point should be considered.[7] This is in contrast to the approaches for modeling in studies in the domain of one-health. Compared to the disease transmission dynamics in plants, zoonoses, and other inter-sectoral areas, human mobility with increased connectivity and commutations for work and leisure, the selection of neighborhood matrices needs deliberate caution and consideration.[7,8]

The representation of the findings as a static figure provides a limited interpretation of space-time dynamic disease processes. Technological advancements need to be harnessed and applied in public health to become appropriate, affordable, and acceptable for benefits to the weaker sections of society for sustainable development. The development of interactive automated parameterized dashboards enables better understanding and fosters evidence-informed decision-making.[17] For the same, intersectoral collaborations between health care and information technology professionals is the need of the hour, evident across multiple sectors, including health care, and increasing exponentially.[17,18]

The study modeled data based on routinely reported cases of lab-confirmed dengue cases in the government setup. Sub-clinical and mild clinical manifestations are seen in up to 80% of dengue infections, and the same could not be captured in the present study.[3] However, the assumption of randomness for varied clinical patterns across sub-districts justifies estimating disease patterns using lab-confirmed line listing data. Future research on clinical patterns using serological studies is required and beyond the scope of the present study. Also, point pattern analysis could not be carried out in the study undertaken. This may be attributed to the lack of coordinates of case occurrence in the RHIS. The addresses were geocoded; however, the bounding boxes obtained for the geocoded data though found adequate for areal data analysis, and point pattern analysis was not recommended. Standardization of address components in RHIS is required to allow such inferences in future studies. Also, similar to the recent introduction of the integrated health

information portal by GoI, GIS integration into RHIS shall enable point pattern analysis in future studies based on the data science approach.

CONCLUSION

Spatial clustering was observed at the extremes of the seasonality pattern of dengue incidence. Spatial clusters were dynamic in space and time. The development of open-source algorithms provides evidence for informed decision-making by public health managers. Research on routinely collected data has the potential to provide insights into the data quality issues, spatio-temporal disease epidemiology, and identification of features/ variables for disease modeling in subsequent studies.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: The study was approved by the Institutional Ethics Committee

REFERENCES

1. Davenhall WF, Kinabrew C. Geographic Information Systems in Health and Human Services. In: Kresse W, Danko D, eds. Springer Handbook of Geographic Information. Cham: Springer International Publishing. 2022;781-805.
2. Garg PK. Geospatial Health Data Analytics for Society 5.0. In: Garg PK, Tripathi NK, Kappas M, Gaur L, eds. Geospatial Data Science in Healthcare for Society 5.0. Singapore: Springer Singapore. 2022;29-58.
3. World Health Organization. Dengue and severe dengue Factsheet. Available at: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue. Accessed on 27 Sept, 2022.
4. Directorate of National Vector Borne Disease Control Programme. Long Term Action Plan for prevention and control of Dengue and Chikungunya. 2007. Available at: https://nvbdcp.gov.in/Doc/Final_long_term_Action_Plan%20.pdf . Accessed on 27 Sept, 2022.
5. Wickham H, Grolemund G. R for data science: import, tidy, transform, visualize, and model data, First edition. Sebastopol, CA: O'Reilly. 2016.
6. Van der Aalst W. Data Science in Action. In: van der Aalst W, ed. Process Mining: Data Science in Action. Berlin, Heidelberg: Springer. 2016;3-23.
7. Pfeiffer DU, Robinson TP, Stevenson M, Stevens KB, Rogers DJ, Clements ACA. Spatial Analysis in Epidemiology. OUP Oxford. 2008.
8. O'Sullivan D, Unwin D. Geographic Information Analysis. Wiley. 2014.
9. Emerging Hot Spot Analysis: Finding Patterns over Space and Time. Azavea. 2017;15.
10. ArcGIS Pro. Emerging Hot Spot Analysis (Space Time Pattern Mining). Available at: https://pro.arcgis.com/en/pro-app/latest/tool-reference/space-time-pattern-mining/emerginghotspots.html. Accessed on 7 Oct, 2022.
11. Singh G, Mitra A, Soman B. Development and use of a reproducible framework for spatiotemporal climatic risk assessment and its association with decadal trend of dengue in India. Indian J Community Med. 2022;47:50.
12. Singh G, Soman B. Spatiotemporal Epidemiology and Forecasting of Dengue in the state of Punjab, India: Study Protocol. Spatial Spatio-temporal Epidemiol. 2021.
13. Singh G, Soman B, Mitra A. A Systematic Approach to Cleaning Routine Health Surveillance Datasets: An Illustration Using National Vector Borne Disease Control Programme Data of Punjab, India. arXiv:210809963 [cs] 2021;23.
14. Hung YW, Hoxha K, Irwin BR, Law MR, Grépin KA. Using routine health information data for research in low- and middle-income countries: a systematic review. BMC Health Services Res. 2020;20:790.
15. Open Government Data (OGD) Platform India. Open Government Data (OGD) Platform India. 2022. Available at: https://data.gov.in. Accessed on 29 Sept, 2022.
16. Directorate General of Health Services, India. Joint Monitoring Mission Rep. 2015.
17. National Digital Health Mission. Available at: https://ndhm.gov.in/. Accessed on 19 Feb, 2021.
18. Telle O, Vaguet A, Yadav NK. The Spread of Dengue in an Endemic Urban Milieu-The Case of Delhi, India. PLoS one. 2016;11:e0146539.
19. Yan Z, Haimes YY. Risk-based multiobjective resource allocation in hierarchical systems with multiple decisionmakers. Part I: Theory and methodology. Syst Engin. 2011;14:1-16.
20. Yigitbasioglu OM, Velcu O. A review of dashboards in performance management: Implications for design and research. Int J Accounting Information Systems. 2012;13:41-59.
21. Few S. Information dashboard design: the effective visual communication of data, 1st ed. Beijing; Cambride [MA]: O'Reilly, 2006.

Cite this article as: Singh G, Soman B, Grover GS. Development and use of open-source algorithms for space-time emerging hotspot analysis of routine dengue NVBDCP data in Punjab, India. Int J Community Med Public Health 2023;10:148-53.