Digital Detectives of Environment Tackling Cigarette Butt Pollution

İbrahim Topcu [1], Berk Anbaroğlu [1]

[1] Department of Geomatics Engineering, Hacettepe University, Ankara, Türkiye

Keywords: FOSS4G, cigarette butt, health policy, reproducible research, Web GIS, Green Deal

Abstract

The aim of this paper is to design and develop an openly accessible, web-based Crowdsourced Geographic Information (CGI) framework, referred to as the Digital Detectives of Environment (DiDE), to facilitate the collection of geo-located events. The framework incorporates three user roles: (i) citizens, (ii) experts, and (iii) supervisors. Citizens can browse relevant events without requiring authentication, while experts are responsible for collecting geographic data, including the optional attachment of photographs or videos. Supervisors, on the other hand, define and manage event types. Each event type is classified as either useful or harmful, which determines its visibility to citizens. The pilot implementation was conducted at the Beytepe Campus of Hacettepe University, focusing on four event types aligned with Green Deal actions: rubbish bins and recycling bins (useful), and cigarette butts and full rubbish/recycling bins (harmful). During a one-week data collection period, a total of 490 events were recorded by 37 students. The results reveal clear clustering patterns in both space and time. Temporally, a large proportion of the data was collected on the final day, indicating a tendency toward procrastination among participants. Spatially, events are concentrated in the southern part of the campus, where most facilities are located. This pattern is further supported by analyses using the *F* and *G* functions. In particular, cigarette butt events exhibit strong spatial clustering, with a mean nearest-neighbour distance of approximately 25 metres. This finding provides empirical support for the *broken windows* theory.

Project’s GitHub repo: https://github.com/ibrahimtopcu11/DiDe/tree/isprs26

Dataset: https://zenodo.org/records/19442148

1. Introduction

Cigarette smoking is one of the fundamental health issues of today. Beyond its well-documented impact on respiratory and cardiovascular health (Giovino et al., 2012), its effects spill over into environmental and social sphere, creating a burden that communities continue to struggle with. From the toxic chemicals released into the air to the billions of discarded cigarette butts contaminating streets, parks, and waterways. Consequently, they are to be the most common trash item collected worldwide (Sy, 2024). Addressing this issue requires not only medical intervention but also raising public awareness, initiating behavioural change, and developing innovative community-based initiatives that leverage web-based geospatial technologies.

An important social aspect of the struggle against smoking is the emergence of ‘smoke-free’ (SF) and ‘tobacco-free’ (TF) campuses. A recent study found out that employing such policies have positive effects in terms of smoking behaviour. Specifically, smoking initiation has decreased, while the number of people quitting smoking has increased. In addition to these health benefits, the environmental impact should not be under estimated, as the number of cigarette butts decreased after the implementation of such policies (Gnonlonfin, Geindreau, & Gallopel-Morvan, 2024). However, introducing such policies in a campus environment is often challenging. Previous research has identified several patterns of noncompliance (e.g. the inconvenience of traveling off-campus to smoke, or the tendency to smoke in concealed areas to avoid detection). Many of these themes have a strong spatial dimension: confusion about policy boundaries, poorly monitored campus zones, and the availability of hidden or semi-private outdoor spaces can all encourage noncompliant behaviour (Jancey et al., 2014). Therefore, understanding the spatio-temporal characteristics of cigarette butt pollution of a campus becomes crucial when designing and evaluating effective SF/TF policies.

The aim of this paper is to develop an open-source web-based geospatial framework that we refer to as *Digital Detective of Environment* (DiDE), and investigate its use in identifying the cigarette butt hot-spots and their relation to the locations of rubbish bins. The second section introduces the DiDE framework, and the third section introduces the pilot project carried out at Hacettepe University. The fourth section describes the quantitative analysis of the openly available dataset. Finally, section five concludes the paper, and presents the future research directions.

2. DiDE Framework

This section describes the open-source methodology adopted in the design and development of the DiDE framework. Five open-source software are required for establishing a running DiDE application. First, *PostgreSQL/PostGIS* is the preferred database management system, as it can be used to store the collected geographic and user data. Second, *NodeJS* has been chosen to run the server side. Third, *Bootstrap* is used on the front end to ensure mobile compatibility. Fourth, *QGIS* is used to setup the project for data analysis and easing the data collection on the field. Finally, *GitHub* is used for version control. Therefore, the DiDE framework is comprised of Free and Open Source for Geospatial (FOSS4G) software.

There are three types of users in the DiDE framework: supervisors, detectives, and citizens. Supervisors oversee the overall data-collection process by adding and updating feature types (e.g., cigarette-butt pollution, designated smoking zones, recycling-bin locations) and managing the detectives. Detectives serve as field workers who collect geospatial features and may attach photographs or videos to provide additional detail. Citizens represent the general public, who can be provided with useful geospatial information such as the locations of recycling bins or designated smoking areas. The Create-Read-Update-

Delete (CRUD) operations that each user type can perform on feature types and individual point observations are illustrated in Table 1.

Supervisor	✓	✗	✓*	✓✗	✗	✓✗	✓	✓
Detective	✗	✓	✓	✗	✓✗	✗	✓✗	✗
Citizen	✗	✗	✓✗	✗	✗	✗	✗	✗

Table 1. Database CRUD operations (Create, Read, Update and Delete) that each user type can realise on the DiDE framework.

Supervisors and detectives must login to the system in order to collect and manage the geospatial data, whereas citizens do not need to login. Once the project is live, they can browse the application, and observe the details of the posts that are relevant to them (e.g. location of the recycle bins) as illustrated in Figure 1. Their view can be dynamically arranged to include the details of events that are not directly relevant to them (e.g. cigarette butt pollution) during the setup of the project.

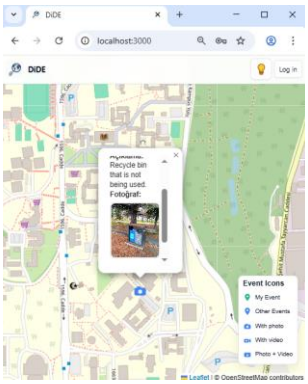


Figure 1. Citizens can display the geographic data and its details without logging in.

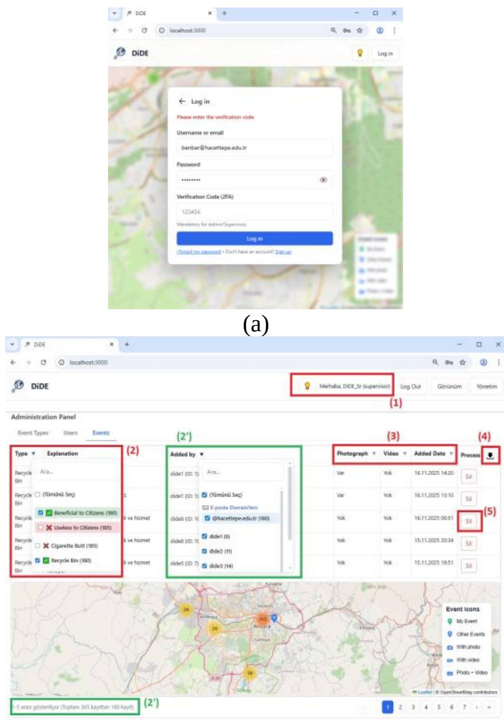
In order to provide additional security for supervisors, two-factor authentication is required during login as illustrated in Figure 3a. It is straightforward to associate a 2FA key with a supervisor, and all the steps required to add such a supervisor is described in the project's GitHub repo. Any number of supervisors can be associated with the project. These supervisors may correspond to the representatives of departments, or student champions who are committed to working on this project. This approach of having champions or expert citizens is common in citizen science projects to assure the quality of collected Crowdsourced Geographic Information (Gallagher et al., 2025; Marks, Mitchell, Rowbotham, Laird, & Smith, 2024). The main role of supervisors is to add the event types, represented as point features. In this way, detectives may only associate each data point to one of these event types. While introducing an event type, the supervisor must also specify whether it is useful to the community or not.

One of the advantages of the DiDE framework is the enablement of multimedia services on the desktop environment as well (Figure 2). Specifically, with the styling of widgets, it is possible to display the attached photos and videos in QGIS. Such multimedia interaction on QGIS is usually not possible.

(a) (b)
(c) (d)

Figure 2. The web-based nature of DiDE coupled with the advanced capabilities of QGIS widget styling, multimedia can be linked for photos (a, b), and videos (c, d).

Supervisors have extended anaylsis capabilities that can be outperformed on the web-based interface. The simulated environment consists of eight detectives and 365 records. The results of the real data collection activity will be described in the next section.



(a)
(b)

Figure 3. Supervisors can perform detailed analyses using dynamic filters, including feature type, detective ID, the presence or absence of multimedia (photos/videos).

The collected data could dynamically be filtered through their panel (Figure 3b). Specifically, out of the 365 posts, only 180 of them were to the benefit of citizens (i.e. recycle bin locations), which is illustrated in component 2 of Figure 3b. This selection, dynamically, adjusts the other filtering options as it can be seen in components labelled by 2’. The other attributes that offer filtering capability are detective ID, the presence/absence of photo/video, and the date each record was added (component 3). The supervisor can download the filtered data through component 4. Finally, supervisors have the right to monitor the collected data, and delete them as well as shown in component 5.

3. Testing DiDE

Beytepe Campus is the main and largest campus of the university, located in Ankara along the Eskişehir Road, approximately 13 km from the city center. Covering an area of nearly 6 million square meters, the campus is known for its wide green spaces. Beytepe Campus hosts a wide range of faculties, such as Engineering, Education, Science, Letters, Fine Arts, and Economics and Administrative Sciences, as well as several institutes and educational units, making it the central location for undergraduate and postgraduate education. The campus is designed as a self-contained environment, offering key facilities such as libraries, student housing, dining areas, sports facilities, and cultural centres, which support both academic activities and student life. Overall, Beytepe Campus stands out as a large and well-organised university campus that brings together education, research, and social life in one place. Consequently, it acts as a microcosm to study green deal policies and actions that are highly intertwined with sustainable development goals (Ahlawat, Bhardawaj, & Bhardwaj, 2024).

The DiDE framework is applied to the third year Geomatics Engineering students enrolled at the Geomatics Engineering Department of Hacettepe University as a Geographic Information Systems course activity. The data collection activity was carried out between 2 December 2025 commencing after course, and 9 December 2025 before the start of the course. Two useful (recycle bin and recycle bin), and two harmful event types (cigarette butt and full rubbish/recycle bin) were introduced. The students were instructed to collect at least 10 events and to distribute their data collection efforts across the campus to ensure broad spatial coverage. Once the data collection activity is over, they were asked to conduct spatial analysis and identify their DiDE partner through a systematic analysis of the dataset depending on their choice of spatio-temporal methods.

Before the data collection activity, the students were advised to strictly follow basic field safety rules, remain vigilant and avoid any potential traffic-related risks, and ensure that all work was conducted within the boundaries of the main campus area, as the Beytepe forests were considered out of scope. In addition, they were instructed that each observation must be carefully documented with a photograph, and that all photographs should be taken in a manner that respects privacy by avoiding the inclusion of any person’s face. The system enables users to provide descriptions to their observations. Even though this was not required, a substantial majority of the students also specified a description.

Characteristic	**Value**	
Number of detectives	37	
Total events	490	
Recycle bin		99
Rubbish bin		87
Cigarette Butt		234
Full rubbish/recycle bin		70
Number of events with a description	356	

Table 2. Overall characteristics of the collected data

Some students chose to continue their data collection activities beyond the required period. The view of the project, as of 30 March 2026, demonstrates the 200 useful events from the project’s web-site as illustrated in Figure 4, compared to 187 as illustrated in Table 2.

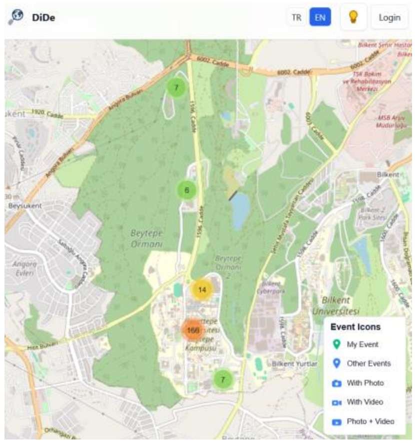


Figure 4. Visitors may view the useful events without registering to the web-site (https://dide.hacettepe.edu.tr/)

4. Analysis

This section presents the analysis of the collected dataset, which is openly available to enable reproducible research. Each post has a timestamp associated with it that is automatically obtained. First, the number of new users posting their first events by day, and the total number of posts on that day is illustrated in Figure 5. Among the 37 students, 6 submitted their first posts on the second day of the analysis period. This is intriguing as it was already the second day of analysis. The requirement to collect data on three separate days might have led to a peak on Thursday, 4 December 2025, as many students might have aimed to meet the requirement without coming to campus during the weekend (6 and 7 December), which correspondingly shows the lowest number of new users. Finally, nine students submitted their data only on the day before the deadline and therefore did not meet the requirement of collecting data on at least three different days. Consequently, highest engagement with the web-based system occurred just before the deadline on 8 December 2025.

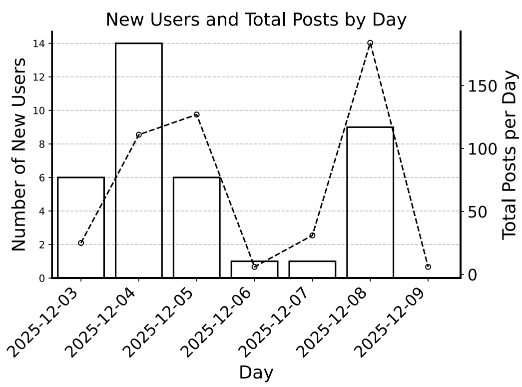


Figure 5. New users by start day

Another important requirement was meeting the minimum number of posts, whereby each student was expected to record at least 10 events while ensuring that their observations were spatially distributed across the campus as widely as possible. The requirement was to ensure that each post has at least 250 metres between them. This approach also aimed to help identify less visible or overlooked areas of pollution.

Ten students did not meet the minimum data collection requirement. As illustrated in Figure 6, these students also tended to cover relatively smaller areas of the campus which is measured by the area of the convex hull of their posts’ locations. In fact, only three of them demonstrated a notably sparse data collection pattern compared to the overall average.

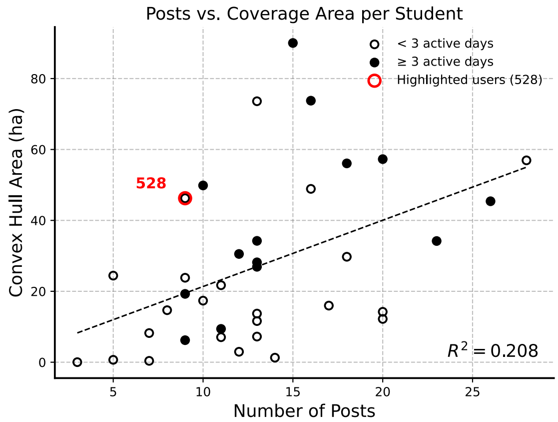


Figure 6. Number of posts collected and geographic coverage area (convex hull) per student

One student (ID: 528) collected data across two separate dates, yielding nine events whose convex hull covered an area of approximately 46 hectares (Figure 7). On 5 December, the westernmost point of the resulting convex hull was recorded by the student, together with three events situated within the hull between 10:14 and 10:31. Then, on 7 December, between 13:53 and 14:23, the remaining observations were collected, which delineate the northern extent of the campus and the eastern boundary of the convex hull.

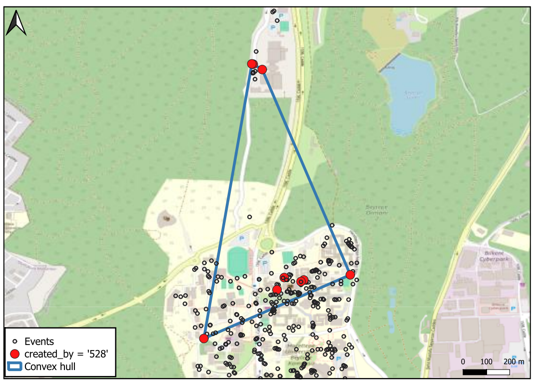
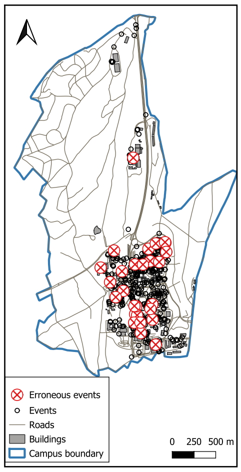


Figure 7. The convex hull of the events recorded by the student ‘528’. Majority of events are concentrated on the main campus area located on the southern part of the map.

Although the student provided explanations for all submitted posts, these were largely descriptions of locations, information already readily available through the web-based GIS system, which uses the smartphone’s location functionality. The user submitted four posts for ‘Recycle bin’, three for ‘Rubbish bin’, and one each for ‘Full rubbish/recycle bin’ and ‘Cigarette Butt’.

4.1 Erroneous Events

The web-based data collection framework does not assume an external dataset, such as building footprints or roads. This section assumes the availability of such data, which is obtained from OpenStreetMap, to identify those events that are inside a building footprint. The point-in-polygon analysis revealed that 29 events were indeed inside a building footprint as illustrated in Figure 8.



(a)

Event type	Count
Cigarette Butt	10
Full bin	7
Rubbish bin	7
Recycle bin	5

(b)

Number of students	Erroneous posts
7	2
15	1

(c)

Figure 8. Erroneous posts may occur due to GPS inaccuracies, as some points fall within building footprints (a). Panel (b) shows the distribution of the 29 events by event type, while panel (c) presents the number of students contributing erroneous posts.

The distribution of these events across event types and users appears random. A total of 22 students reported erroneous observations, indicating that the issue is unlikely to be attributable to individual mobile phones.

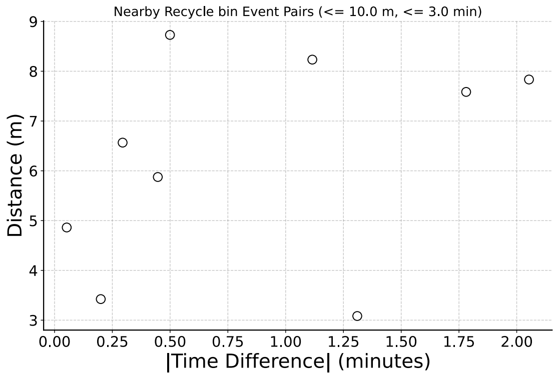
It is important to note that the campus is spatially heterogeneous, with the majority of units located in the southern part of the study area, as shown in Figure 8.

4.2 Identification of Repetitive Events

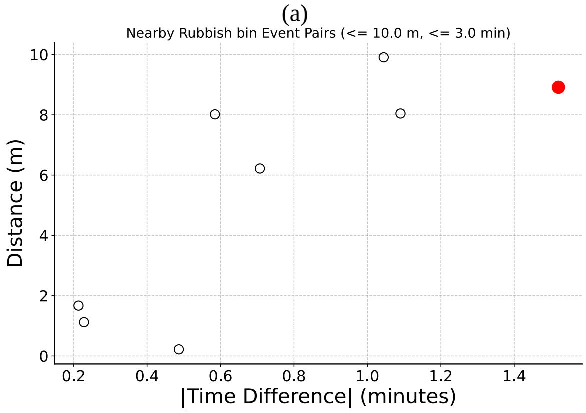
The students might not have carefully inspected previously submitted event images, which may have resulted in duplicate or repetitive event entries. This issue is expected to be more pronounced for static objects (e.g., rubbish bins and recycling bins). Therefore, this subsection investigates the extent of such repetitive event reporting.

To this end, a spatio-temporal analysis was conducted using a PostgreSQL/PostGIS database. Event pairs of the same type were identified through a self-join operation, where only pairs located within a predefined spatial proximity (10 m) were considered. For each pair, both the Euclidean distance and the temporal difference between their timestamps were computed. Subsequently, pairs with a time difference exceeding a specified threshold (3 minutes) were filtered out to focus on near-simultaneous observations.

The resulting event pairs were analysed using a scatter plot defined by spatial distance and temporal difference, enabling the investigation of the relationship between these two variables. It should be noted that there were six and one consecutive events for recycling bins and rubbish bins, respectively, that exceeded the 3-minute threshold, indicating that two distinct students collected the same event.



(a)



(b)

Figure 9. Student pairs may have recorded identical recycle bin (a) or rubbish bin (b) events due to participating in the data collection process together, which should be considered in data quality analysis. The highlighted point corresponds to event IDs 409 and 411.

The results for nearby rubbish bins should be interpreted with caution. In particular, some bins may be genuinely close to each other, as illustrated by the event pair with IDs 409 and 411, which corresponds to the highlighted nearby event in Figure 9b, they are indeed distinct events as can be observed from the photographs of the events illustrated in Figure 10.

(a)

(b) (c)

Figure 10. Nearby events may represent distinct occurrences. Overall view of the area (a), and photos associated with event IDs 409 (b) and 411 (c)

It should also be highlighted that the time gap between repeated event posts might also be substantial. For example, the recycle bin illustrated in Figure 11 have been observed by two students at different times of the day.

(a) (b)

Figure 11. Repetitive posts may result from different students reporting the same recycle bin, as shown in event ID 504 (a) and event ID 7 (b).

4.3 Nearest Neighbour Analysis

This subsection investigates the nearest neighbour analysis of the reported events by their type. Assume two events of type $t \in \{Recycle\ bin, Rubbish\ bin, Cigarette\ Butt,\ Full\ rubbish/recycle\ bin\}$, $\mathbf{s}_i^t$ and $\mathbf{s}_j^t$, the boxplot of nearest neighbour distances by event types are illustrated in Figure 12.

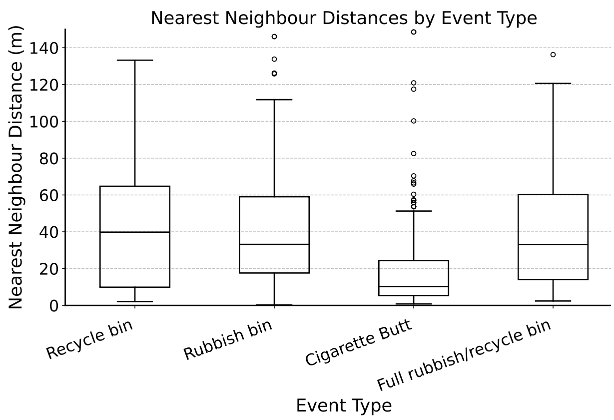


Figure 12. Boxplots of within-type nearest-neighbour distances for four event categories

Distances were computed in projected metric coordinates, with each point paired to its closest point of the same category. The *y*-axis is limited to 0-150 m to emphasise differences in the central distributions. The high density of cigarette butt pollution is evident. It should be noted that some posts might be considered to belong to two or even three different categories, as illustrated in Figure 13. Specifically, a student might have opted for ‘rubbish bin’ or ‘cigarette butt’ in Figure 13a, but the student selected ‘Full rubbish/recycle bin’ and added the description ‘The area is close to restaurants and cafeterias, so the trash bin is full.’. Similarly, in Figure 13b, the student might have chosen for ‘recycle bin’ but instead stated the event type to be ‘cigarette butt’. Consequently, students appear to prioritise the more notable event, with ‘Full bin’ ranking highest.

(a) (b)

Figure 13. Students prioritised the more 'notable' event. In (a), full rubbish (event_id = 63 & 13)

The results indicate clear differences in spatial structure across event types. Cigarette butt observations exhibit strong clustering, suggesting localised and repetitive human activity patterns. In contrast, recycling bins display a more dispersed and regular spatial distribution, consistent with planned infrastructure placement. Rubbish bins and full-bin events show intermediate behaviour, reflecting a combination of infrastructural layout and usage intensity.

4.4 Clustering Analysis using G and F Functions

The *G* function, also known as the refined nearest-neighbour function, summarises the cumulative frequency distribution of distances between events (O'Sullivan & Unwin, 2010). Instead of relying on a single mean value, *G* function examines the fraction of all nearest-neighbour distances that are less than a specified distance *d*. Formally, the *G* function is defined as in equation (1):

$$G(d) = \frac{\#(d_{min}(\boldsymbol{s_i}) < d)}{n} \quad (1)$$

where $d_{min}(s_i)$ represents the distance from an event *i* to its nearest neighbour, and *n* is the total number of events.

The shape of the *G* function reveals the spacing of the pattern. Specifically, for clustered patterns *G* increases rapidly at short distances, as many events are located very close to one another. On the contrary for dispersed patterns, *G* increases slowly initially, rising only when *d* reaches the range where most events are evenly spaced.

While the *G* function focuses on the events themselves, the *F* function measures the gaps between them (O'Sullivan & Unwin, 2010). It is calculated by selecting *m* random locations $\boldsymbol{p_i}$ anywhere within the study region $\boldsymbol{S}$ and determining the minimum distance from these random points to the nearest actual event. Formally, the ***F*** function is defined as in equation (2):

$$F(d) = \frac{\#[d_{min}(p_i, \boldsymbol{S}) < d]}{m} \quad (2)$$

The *F* function serves as a powerful diagnostic tool when compared to *G*. Specifically, in clustered patterns *F* rises slowly because large areas of empty space exist between clusters, meaning a random point is likely to be far from an event. On the contrary, in dispersed patterns, *F* rises quickly because the events are spread evenly throughout the area, making it unlikely for any random point to be very far from an event. The empirical observation illustrated in Figure 12 could further be validated by using the *G* and *F* functions, which is illustrated for $t = \{Cigarette\ Butt\}$ in Figure 14.

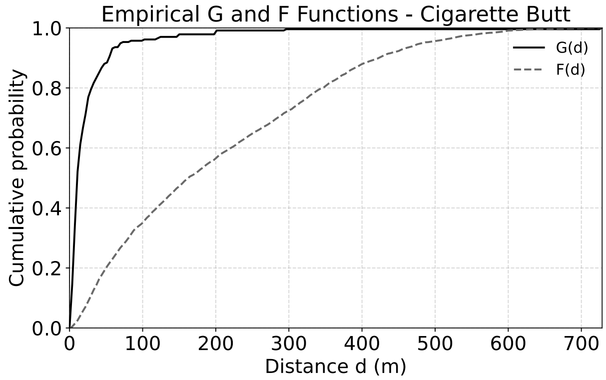


Figure 14. *G* and *F* functions for 'Cigarette Butt' events

Cigarette butt events exhibit strong spatial clustering with a mean nearest-neighbour distance of only 25.40 m. This indicates that essentially all cigarette butt events (>99%) have their nearest neighbour within 100 metres, demonstrating a highly concentrated spatial distribution pattern.

5. Conclusion

This paper presented an openly available web-based platform for the collection and analysis of Green Deal projects. The pilot study was conducted to gather four types of events. During a one-week data collection period, a total of 490 points were recorded by 37 students. The proposed platform is flexible and can be extended to a wide range of applications, particularly in the context of volunteered and crowdsourced geographic information projects.

Future work will focus on enhancing the developed web-based framework in several ways. In particular, measures will be implemented to prevent repetitive data collection. Additionally, the integration of GeoServer will facilitate the efficient dissemination of the collected geographic data.

Acknowledgements

The authors gratefully acknowledge the IT Department of Hacettepe University for providing the server infrastructure, with special thanks to Gökhan Güzel. The authors also thank the students who contributed to this project.

References

Ahlawat, Y., Bhardawaj, A., & Bhardwaj, R. (2024). Pick the butt up! Upcycling cigarette butts in road building material for sustainable road construction. IOP Conference Series: Earth and Environmental Science, 1382(1), 012006. https://doi.org/10.1088/1755-1315/1382/1/012006

Gallagher, R., Roger, E., Packer, J., Slatyer, C., Rowley, J., Cornwell, W., … Mesaglio, T. (2025). Incorporating citizen science into IUCN Red List assessments. Conservation Biology, 39(2), e14329. https://doi.org/10.1111/cobi.14329

Giovino, G. A., Mirza, S. A., Samet, J. M., Gupta, P. C., Jarvis, M. J., Bhala, N., … Asma, S. (2012). Tobacco use in 3 billion individuals from 16 countries: An analysis of nationally representative cross-sectional household surveys. The Lancet, 380(9842), 668–679. https://doi.org/10.1016/S0140-6736(12)61085-X

Gnonlonfin, E., Geindreau, D., & Gallopel-Morvan, K. (2024). What are the effects of smoke-free and tobacco-free university campus policies, and how can they be assessed? A systematic review. Journal of Epidemiology and Population Health, 72(2), 202520. https://doi.org/10.1016/j.jeph.2024.202520

Jancey, J., Bowser, N., Burns, S., Crawford, G., Portsmouth, L., & Smith, J. (2014). No Smoking Here: Examining Reasons for Noncompliance With a Smoke-Free Policy in a Large University. Nicotine & Tobacco Research, 16(7), 976–983. https://doi.org/10.1093/ntr/ntu012

Marks, L., Mitchell, J., Rowbotham, S., Laird, Y., & Smith, B. J. (2024). Citizen science as a vehicle to address multiple policy objectives: Case studies from Australia. Health Promotion Journal of Australia, 35(4), 1326–1336. https://doi.org/10.1002/hpja.866

O’Sullivan, D., & Unwin, D. J. (2010). Point Pattern Analysis. In Geographic Information Analysis (pp. 121–156). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470549094.ch5

Sy, D. K. (2024). Tobacco industry accountability for marine pollution: Country and global estimates. Tobacco Control, 33(e2), e1–e4. https://doi.org/10.1136/tc-2022-057795