Safety21
INNOVATING SAFETY FOR ALL

US DOT National
University Transportation Center for Safety

# Mitigating Crash Risks in Work Zones: Causal Inference and Crash Modification Factors

## Phase I: Matching work zone data to the state networks: case study in Pennsylvania from 2018 to 2024

PI: Sean Qian (ORCID: 0000-0001-8716-8989)

Lead Research Assistants: Quintessa Guengerich (ORCID: 0009-0006-6236-854X); Tao Tao (ORCID: 0000-0001-9865-5985)

FINAL RESEARCH REPORT – July 30, 2024

Contract # 69A3551747111



All data generated from this project can be accessed from
https://github.com/qguengerich/map_matching_work_zones

## 1 Introduction

The impact of work zones on crash incidents remains largely underexplored. Despite the recognized importance of this issue, existing studies have only marginally addressed the correlation between work zones and crash risks. One crucial preliminary task in understanding this relationship is accurately map-matching the work zones to the road network. This foundational step is essential for any subsequent analysis of how work zones might influence crash risk.

In this report, we replicate the initial stages of this critical analysis. Our primary objective is to match work zones to their specific locations on the Pennsylvania road network through a process known as map matching. This involves aligning work zone data with the precise segments of the road network they affect. By establishing these spatial connections, we lay the groundwork for a more comprehensive examination of the potential impacts work zones have on crash incidents.

Understanding the spatial distribution and characteristics of work zones is vital for traffic safety analysis and for developing effective countermeasures. As such, this report serves as an essential step towards a deeper investigation into the safety implications of work zones, aiming to enhance road safety for all users by identifying high-risk factors and informing better management practices for work zone planning and implementation.

## 2 Data

We applied two main data sources in this research project, including the Road Condition Reporting System (RCRS) and the Pennsylvania state road network datasets, both of which are requested from the Pennsylvania Department of Transportation. The RCRS dataset contains data with particularly relevant columns of the starting and ending coordinates of a work zone, the state route number on which the work zone was operating, the direction of traffic on the side of the road of the work zone, the road closure reason, and the opening and closing date and time of the work zone.

The state network dataset comprises the map of Pennsylvania State Roads, and contain the following columns about each road segment, which facilitate map matching of work zones: state route numbers, direction of traffic, and coordinates of each vertex of the road segment.

This project map-matches work zones from the years 2018 to 2024 to the Pennsylvania road network. Only those work zones with their durations shorter than seven days are considered depending on the research design of the next-step causal analysis. Table 1 lists the number of work zones to be matched in each year.

**Table 1. Number of work zones to be matched for each year from 2018 to 2024**

| Year | Number of work zones to be matched |
|---|---|
| **2018** | 31,643 |
| **2019** | 34,885 |
| **2020** | 21,559 |
| **2021** | 20,771 |
| **2022** | 21,185 |
| **2023** | 24,413 |
| **2024** | 9,736 |
| **Total** | 164,192 |

Note: The number of work zones in 2024 is up to July 2024.

## 3 Method

We designed several modules to match work zones to the Pennsylvania road network given the work zone starting and ending coordinates, the state route number, and the direction of the respective roadway. This involves several steps, including steps taken to decrease the computation time via vectorization and filtering of the search space to limited possibilities. In general, the overall algorithm operates under the following assumptions:

- The correct state route number is recorded for each work zone. This means that each work zone should be mapped to their respective state route number, regardless of whether a different route is closer to the coordinates recorded.
- The direction recorded for each work zone is correct. This means that each work zone should be mapped to their respective state route number, going the direction recorded, regardless of whether the other direction is closer to the recorded coordinates of the work zone.

These assumptions significantly shrink the search space of possible road segments for each work zone. To further limit computation time, a buffer size of five meters is imposed on each work zone so that only those road segments that are within a reasonable distance of the work zone are included in the search. We present an example of one of these filtration functions is detailed in the pseudocodes in Figure 1. In Algorithm 1, the searchable road segment data frame is filtered to those that share the same state route number and direction as the work zone. In Algorithm 2, the data frame is further filtered to those road segments that are within a set distance of the work zone.

**Algorithm 1** Filter Search Space to Possible Road Segments

```
function FilterRoadSegments(workzone, roads)
    wz_st_rt_no ← workzone at row State Route Number
    route_filtered_roads ← roads where ST_RT_NO is wz_st_rt_no
    wz_dir ← workzone at row Direction
    dir_filtered_roads ← route_filtered_roads where Direction is wz_dir
    return dir_filtered_roads
end function
```

**Algorithm 2** Point In Box

```
function PointInBox(point, buffer, roads)
    in_lon_range ← roads within buffer longitude of point
    in_lat_range ← roads within buffer latitude of point
    return in_lon_range ∧ in_lat_range
end function
```

**Figure 1. Pseudocodes of filtration functions**

Using the algorithm in Figure 2, we calculated the distance from each segment to the respective point and selected the segment with the smallest distance.

**Algorithm 3** Calculate Distance to Point on Line

```
function CalcDistPL(x, y, line, index)
    point ← Point(x, y)
    distance ← Distance(point, line)
    closest_point_on_line ← Interpolate(point, line)
    return {index, closest_point_on_line, distance}
end function
```

**Figure 2. Pseudocodes for calculating distance from a starting or ending point to a roadway segment**

After we matched the starting and ending points of a work zone to their respective road segments, as depicted in Figure 3, we selected the connected road segments between them.

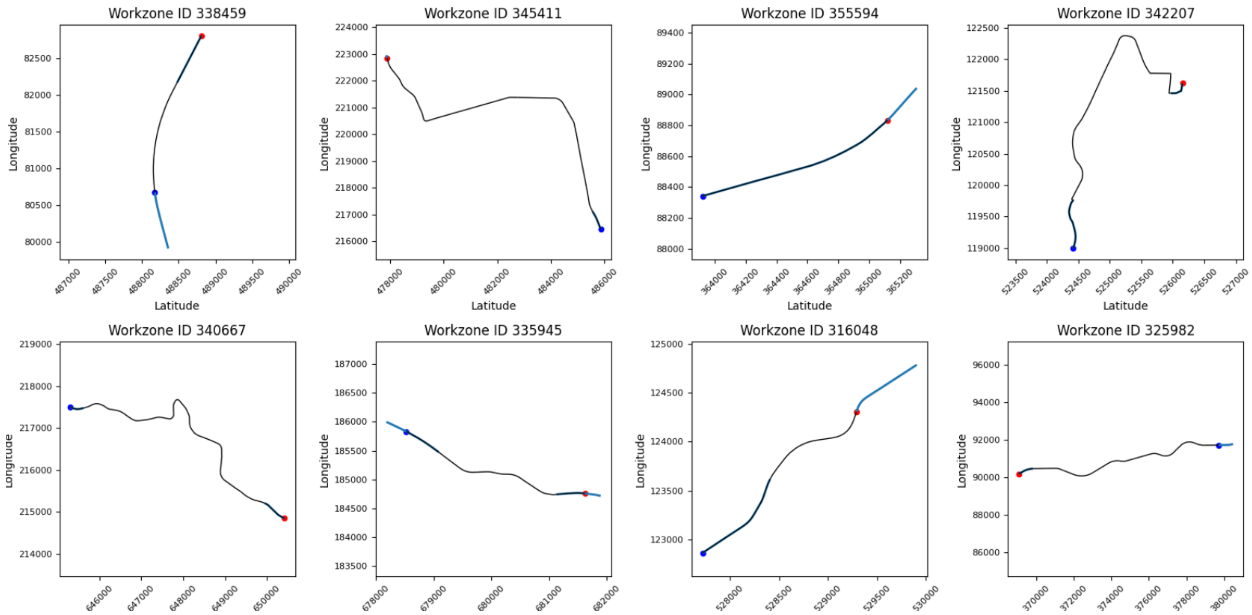


**Figure 3. Examples in which the starting and ending points of a work zone are matched to their respective road segments**

Algorithm 4 (Figure 4) traverses from end to end of each road segment until both the starting and ending points are included. The resulting geometry includes all road segments that unite the start of the work zone to the end of the work zone.

In most cases, work zone starting and ending points fall on the line of a road segment, and when that entire road segment is selected, it is not representative of the true length of the work zone. The last step of the algorithm (Algorithm 5 in Figure 4) trims the selected geometry down to the exact size of the work zone as indicated by the starting and ending coordinates by selecting only the geometry that exists between the exact points.

**Algorithm 4** Get Road Segments

```
function GetRoadSegments(roads, start_segment, end_segment)
    success ← true
    {segments, found_start, found_end, result_road} ←
TraverseSegments(roads, start_segment, end_segment, true)
    if not found_start or not found_end then
        {segments, found_start, found_end, result_road} ←
TraverseSegments(roads, start_segment, end_segment, false)
    end if
    if not found_start or not found_end then
        success ← false
    end if
    return {segments, success}
end function
```

**Algorithm 5** Traverse Segments

```
function TraverseSegments(roads, start, end, start_with_idx1)
    p1idx ← start road index
    p2idx ← end road index
    segments ← Empty list
    i ← if start_with_idx1 then p1idx else p2idx
    while not found_start or not found_end do
        if i is start and end index then
            found_start, found_end ← true
            Trim road geometry between start and end points
        else if i is start index then
            found_start ← true
            Trim road geometry from start point
        else if i is end index then
            found_end ← true
            Trim road geometry from end point
        end if
        segments ← concat segments with segment
        if road has more connections then
            i ← next connection
        else
            break
        end if
    end while
    return {segments, found_start, found_end, result_road}
end function
```

**Figure 4. Pseudocodes for selecting the geometry between the starting and ending points of the work zone**

An example of map-matched work zone is shown in Figure 5. We identified the starting and ending points and the road segments between these two points on the state road network. The final pseudocode is shown in Figure 6.

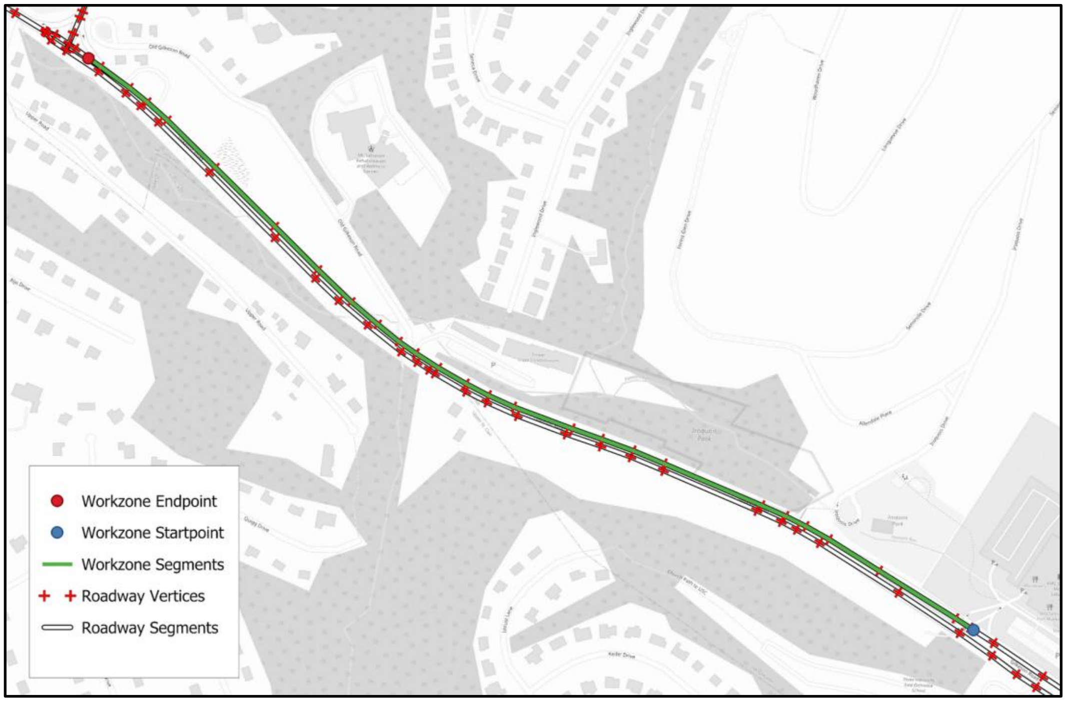


**Figure 5. One example of map-matched work zone**

**Algorithm 6** Map-matching Workzones to Roadways

```
function MapMatchingWorkzonesToRoadways(workzones, roads)
    starts_df ← Empty Dataframe
    ends_df ← Empty Dataframe
    workzone_segments ← Empty Dataframe
    for each workzone in workzones DataFrame do
        filtered_roads ← FilterRoadSegments(workzone, roads)
        possible_starts ← PointInBox(workzone start, buffer, filtered_roads)
        possible_ends ← PointInBox(workzone end, buffer, filtered_roads)
        for each index, segment in possible_starts do
            wz_x ← x row in workzone
            wz_y ← y row in workzone
            line ← geometry row in segment
            results ← CalcDistPL(wz_x, wz_y, line, index)
            starts_df ← concatenate starts_df with results
        end for
        for each index, segment in possible_ends do
            wz_x ← x row in workzone
            wz_y ← y row in workzone
            line ← geometry row in segment
            results ← CalcDistPL(wz_x, wz_y, line, index)
            ends_df ← concatenate ends_df with results
        end for
        closest_start_seg ← row in starts_df where dist column has min value
        closest_end_seg ← row in ends_df where dist column has min value
        {segments, success} ← GetRoadSegments(filtered_roads, closest_start_segment, closest_end_seg)
        if success then
            workzone_segments ← concat workzone_segments with segments
        end if
    end for
    return starts_df, ends_df, workzone_segments, roads
end function
```

**Figure 6. Completed pseudocode for map matching work zones**

In the Figure 7, we present an array of eight different examples of work zones of different lengths and complexities that have been matched to the Pennsylvania road network.

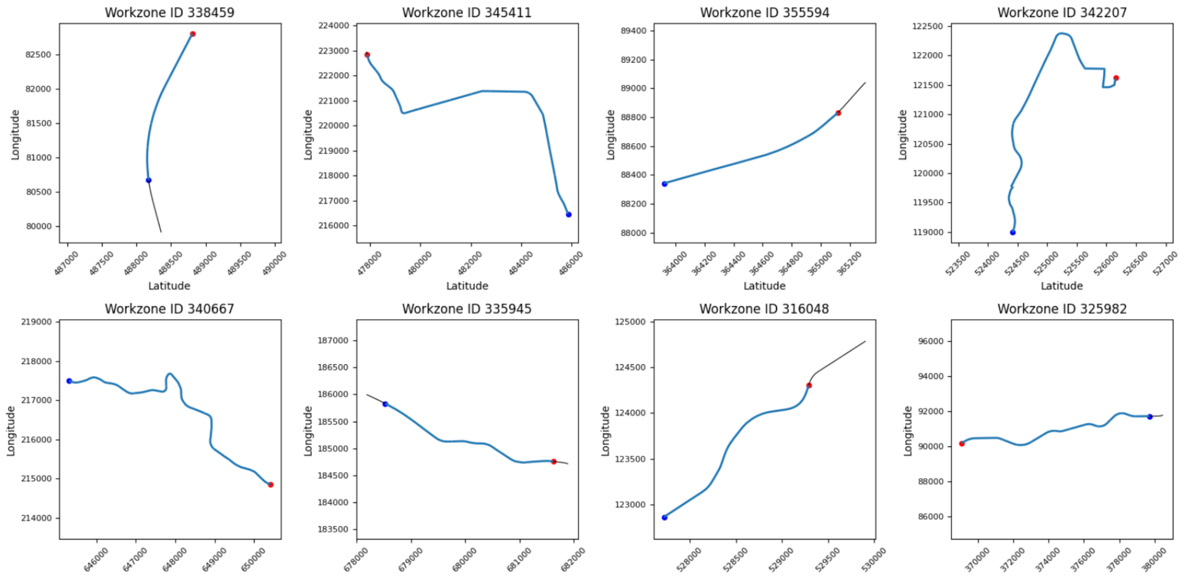


**Figure 7. Eight selected work zones of varied complexity and length, matched to their respective roadways.**

After the act of map matching, we observed that some work zones shared the same space at the same time. This may occur when the same work zone is recorded by different agencies. These records were identified and aggregated to one.

The map matching was carried out with Python. The related codes could be found in the shared GitHub link (https://github.com/qguengerich/map_matching_work_zones).

## 4 Results

In this section, we present the results of the map-matched work zones. Table 2 summarizes the results of the map matching and the subsequent aggregation of overlapping work zones.

**Table 2. Meta data about the number of work zones**

| Year | Number of Work Zones | % Successfully Map Matched | Number of Work Zones After Map Matching and Aggregation |
|---|---|---|---|
| **2018** | 31,487 | 94.0% | 31,643 |
| **2019** | 36,205 | 96.0% | 34,885 |
| **2020** | 22,324 | 96.4% | 21,559 |
| **2021** | 21,294 | 97.4% | 20,771 |
| **2022** | 21,712 | 97.4% | 21,185 |
| **2023** | 24,863 | 98.0% | 24,413 |
| **2024** | 9,976 | 97.5% | 9,736 |

As seen in Table 2, up to 6.0% of the work zones are not successfully map matched. There are two major reasons for this. First, a small number of work zones have invalid state route numbers, directions, and beginning and ending points, which make it impossible to search related road segments. Second, the state road network is not perfectly connected and has some road segment discontinued. Some road segments also have incorrect information of their directions. Algorithm 5 fails to handle such a case, because the algorithm relies on the end-to-end connectivity following correct directions of road segments to traverse and search for the work zone end points. This case accounts for most of the unsuccessful map matching work zones. We cannot address the cases of incorrect information. However, future versions of this map matching algorithm may handle the discontinuity of the state network through a spatial analysis operation to find all interconnected geometries and correct them.

Descriptive statistics and distributions of related characteristics of the map-matched work zones are shown below. Work zone lengths and work zone duration are plotted in Figure 8 and Figure 9 and summarized in Table 3. Most work zones are less than 24 hours and 12,500 meters.

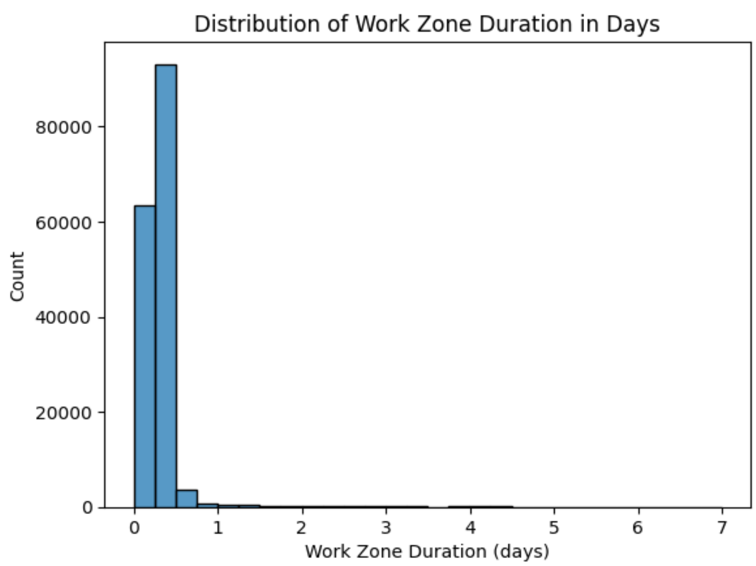


**Figure 8. Histogram of durations of the map-matched work zones**

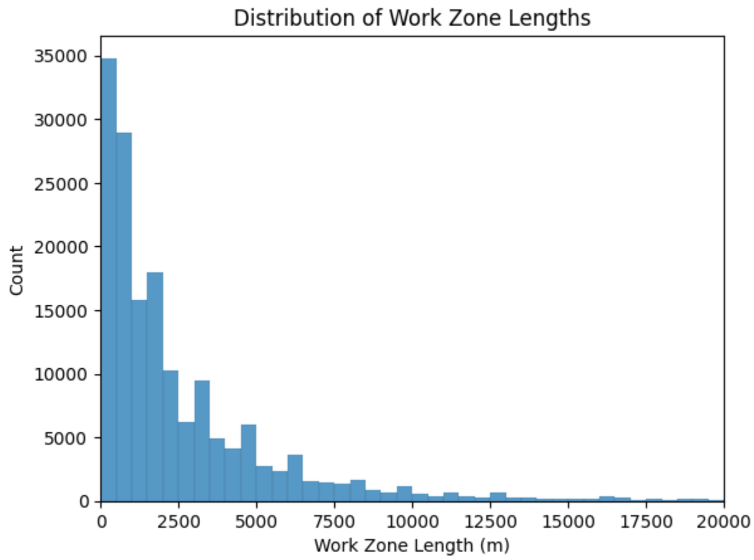


**Figure 9. Histogram of lengths of the map-matched work zones**

**Table 3. Durations and lengths of map-matched work zones**

| Year | Duration (day) | | Length (meter) | |
|---|---|---|---|---|
| | Average | Standard deviation | Average | Standard deviation |
| **2018** | 0.3264 | 0.4206 | 4086.00 | 7688.08 |
| **2019** | 0.3182 | 0.4119 | 3944.38 | 8628.90 |
| **2020** | 0.3487 | 0.4301 | 2384.69 | 3433.56 |
| **2021** | 0.3239 | 0.3771 | 2429.86 | 3325.54 |
| **2022** | 0.3244 | 0.3941 | 2416.86 | 3259.88 |
| **2023** | 0.3137 | 0.3222 | 2619.61 | 3480.54 |
| **2024** | 0.3119 | 0.3516 | 2433.10 | 3433.68 |
| **All years** | 0.3243 | 0.3941 | 3091.61 | 5883.18 |

A map of the Pennsylvania road network is shown below (Figure 10), with all map matched work zones shown in orange. Along certain road segments, work zones were constructed repeatedly throughout the 2018 – 2024 period. Please see the Appendix for maps of work zones during each year (Figure 11-Figure 17).

**Figure 10. All map matched work zones from 2018-2024.**

## 5 Conclusion and future research

In this research project, we applied a self-designed algorithm to map match the work zones to the Pennsylvania state road network based on their beginning and ending locations, route number, and traffic directions from 2018 to 2024. The algorithm successfully map matched over 94% of the work zones for each year. The average length of the map matched work zones is 3091 meters. The average duration of the map matched work zones is 0.32 days. Up to 6% of work zones cannot be successfully map matched for two main reasons. First, the work zones contain incorrect information about location, route number, and direction. Second, the state network has incorrect information about direction and disconnected road segments. Future improvement of the algorithm could focus on reconnecting the road segments through spatial analysis.

The results of this research project lay the groundwork for analyzing the impact of work zones on crash risk (e.g., (Zhang et al., 2022b, 2022a)) in two main ways. First, map matching assists in accurately locating work zones on corresponding road segments. This allows for the correlation of important road segment characteristics—such as the number of lanes, road classification, traffic volume, and speed—with work zones. By controlling for these factors, the analysis can provide a more precise estimate of the impact of work zones on crash risk. Additionally, the analysis can assess the heterogeneous impact of work zones based on different road characteristics, offering comprehensive insights into the varying effects.

Second, map matching helps eliminate irrelevant work zones, which often contain incorrect information. By removing these inaccuracies, the model's estimation accuracy is significantly improved.

## References


Zhang, Z., Akinci, B., & Qian, S. (2022a). Inferring heterogeneous treatment effects of work zones on crashes. *Accid Anal Prev*, *177*, 106811. https://doi.org/10.1016/j.aap.2022.106811

Zhang, Z., Akinci, B., & Qian, S. (2022b). Inferring the causal effect of work zones on crashes: Methodology and a case study. *Analytic Methods in Accident Research*, *33*, 100203. https://doi.org/10.1016/j.amar.2021.100203


## Appendix

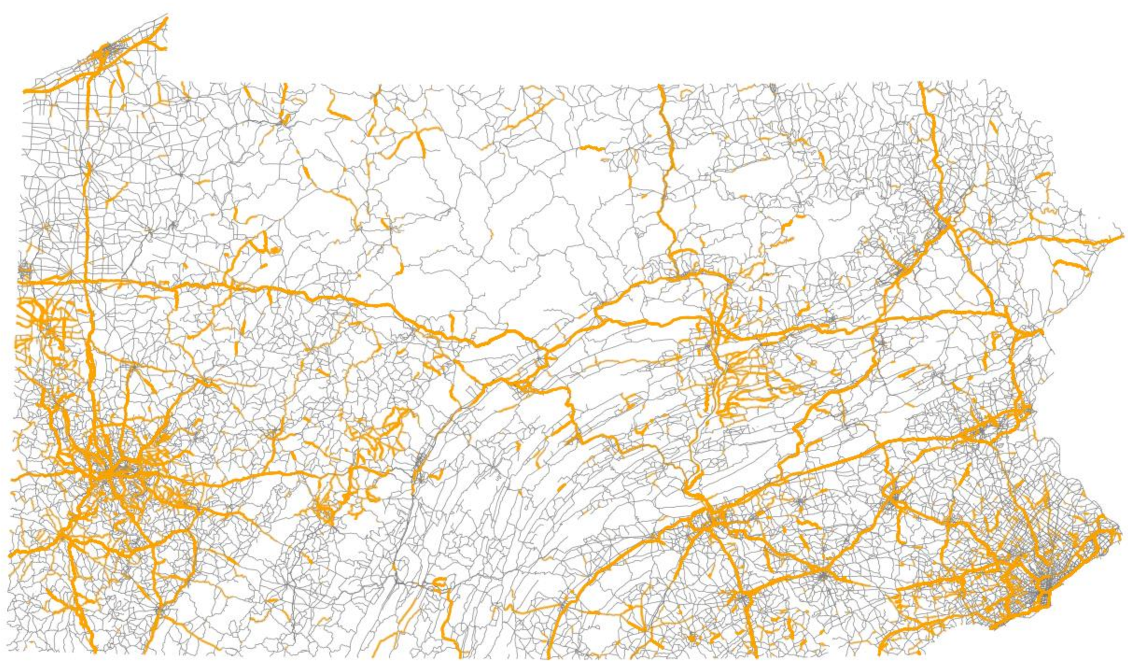

**Figure 11. All map matched work zones from 2018**

**Figure 12. All map matched work zones from 2019**

**Figure 13. All map matched work zones from 2020**

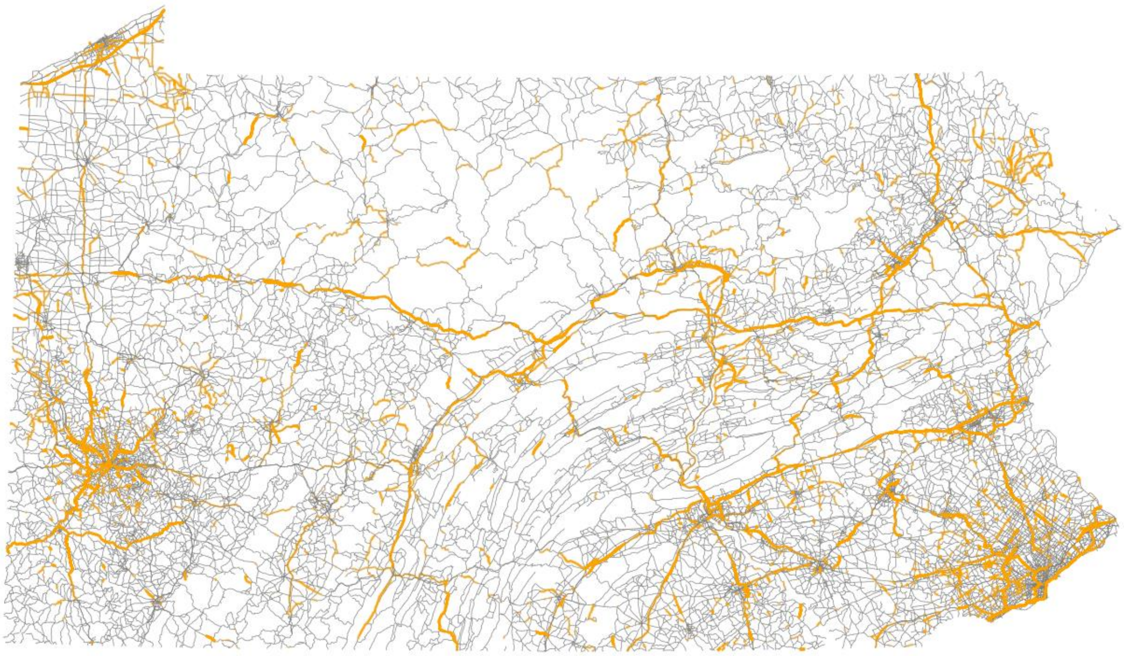

**Figure 14. All map matched work zones from 2021**

**Figure 15. All map matched work zones from 2022**

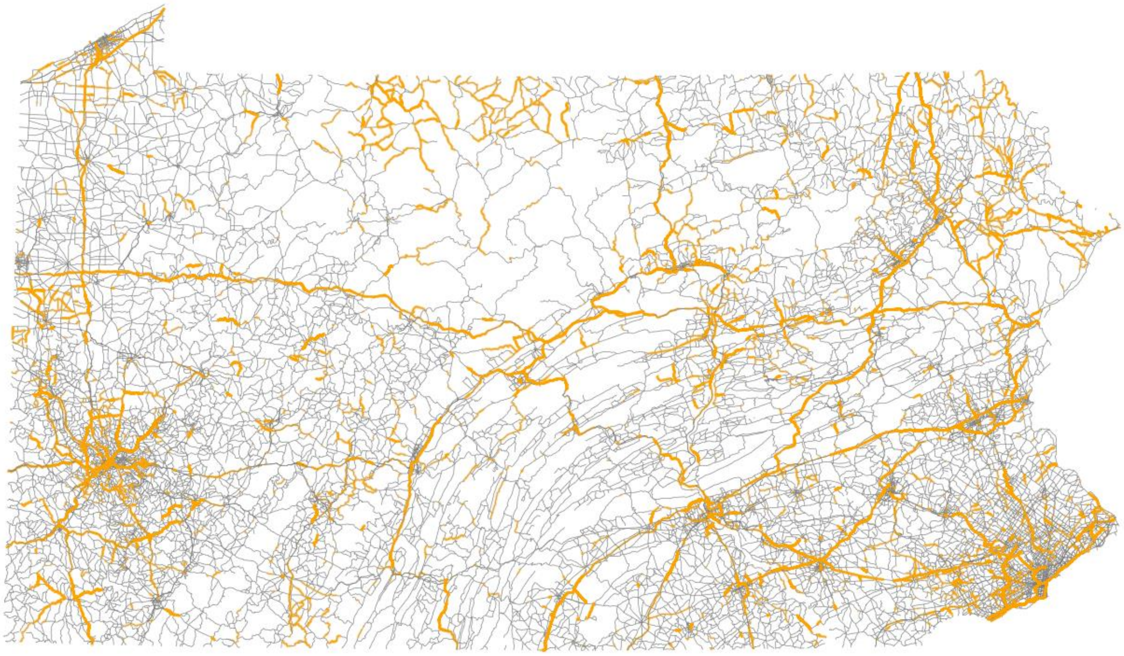

**Figure 16. All map matched work zones from 2023**

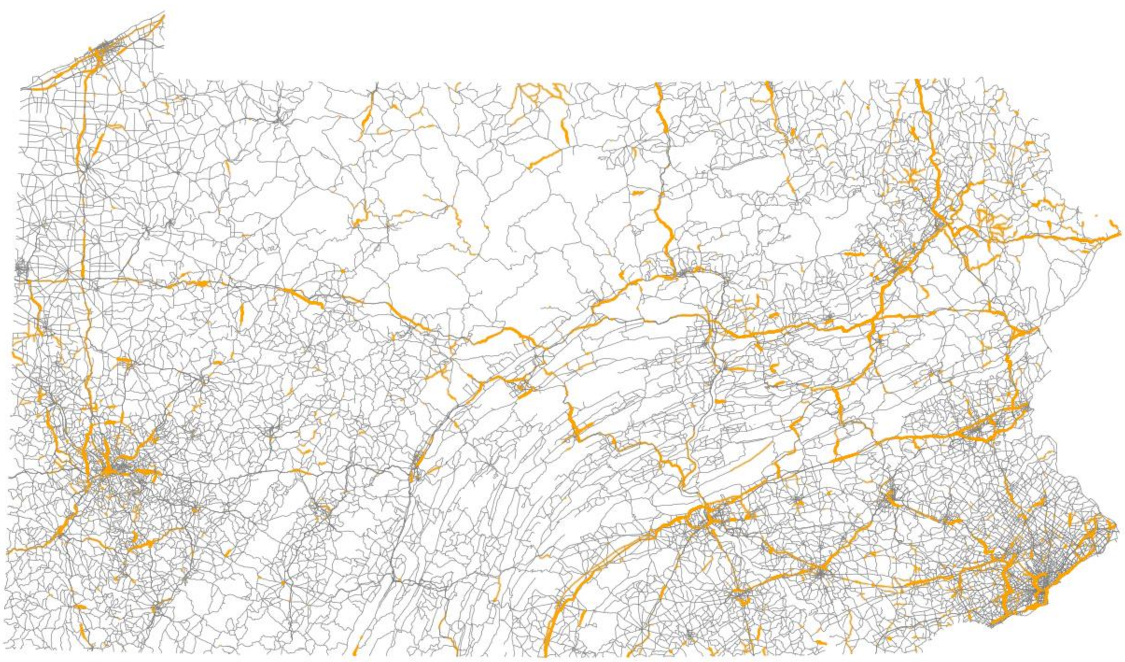

**Figure 17. All map matched work zones from 2024**

| 1. Report No. | 2. Government Accession No. | 3. Recipient's Catalog No. |
|---|---|---|
| . | | |

**4. Title and Subtitle**
Mitigating Crash Risks in Work Zones: Causal Inference and Crash Modification Factors. Phase I: Matching work zone data to the state networks: case study in Pennsylvania from 2018 to 2024

**5. Report Date**
Aug 1, 2024

**6. Performing Organization Code**
.

**7. Author(s)**
Quintessa Guengerich (ORCID: 0009-0006- 6236-854X)
Tao Tao (ORCID: 0000-0001-9865-5985)
Sean Qian (ORCID: 0000-0001-8716-8989)

**8. Performing Organization Report No.**
Enter any/all unique alphanumeric report numbers assigned by the performing organization, if applicable.

**9. Performing Organization Name and Address**
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213

**10. Work Unit No.**

**11. Contract or Grant No.**
Federal Grant No. 69A3552344811

**12. Sponsoring Agency Name and Address**
Safety21 University Transportation Center
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

**13. Type of Report and Period Covered**
Final Report (July 1, 2023-June 30, 2024)

**14. Sponsoring Agency Code**
USDOT

**15. Supplementary Notes**

**16. Abstract**
The impact of work zones on crash incidents remains largely underexplored. Despite the recognized importance of this issue, existing studies have only marginally addressed the correlation between work zones and crash risks. One crucial preliminary task in understanding this relationship is accurately map-matching the work zones to the road network. This foundational step is essential for any subsequent analysis of how work zones might influence crash risk. Our primary objective is to match work zones to their specific locations on the Pennsylvania road network through a process known as map matching. This involves aligning work zone data with the precise segments of the road network they affect. By establishing these spatial connections, we lay the groundwork for a more comprehensive examination of the potential impacts work zones have on crash incidents. Understanding the spatial distribution and characteristics of work zones is vital for traffic safety analysis and for developing effective countermeasures. As such, this report serves as an essential step towards a deeper investigation into the safety implications of work zones, aiming to enhance road safety for all users by identifying high-risk factors and informing better management practices for work zone planning and implementation.

**17. Key Words**
Work zone, GIS, networks, map-matching, crashes

**18. Distribution Statement**
No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161. Enter any other agency mandated distribution statements. Remove NTIS statement if it does not apply.

| 19. Security Classif. (of this report) | 20. Security Classif. (of this page) | 21. No. of Pages | 22. Price |
|---|---|---|---|
| Unclassifed | Unclassified | 16 | . |

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized