1. **Wyznaczanie skorygowanych lokalizacji przystanków**

Baza danych Systemów Dynamicznej Informacji Pasażerskiej zawiera historyczne dane dotyczące pozycji autobusów w postaci punktów w formacie WGS84[1]. Każdy punkt powiązany jest z konkretnym pojazdem oraz konkretnym dziennym kursem. Każdy punkt zawiera również informację o dacie oraz czasie, kiedy został zarejestrowany. Zazwyczaj *czas próbkowania* wynosił kilkanaście sekund. Oznacza to, że można je odnieść do faktycznego przejazdu autobusu z przeszłości oraz otrzymać informacje o tym, na jakim przystanku powinien się zatrzymywać. Dane te zawarte są w tabeli *csiptrace*. W konkretnej bazie, która była używana w trakcie projektu tabela zawierała ponad 7 milionów rekordów. Dzięki lokalizacjom punktów oraz czasom, w którym pojazd się znajdował w tych punktach, można wyznaczyć średnią prędkość pomiędzy punktami a w konsekwencji, z dość dużą pewnością, faktyczne punkty zatrzymania.

1. **Algorytm wyznaczenia punktów zatrzymania**

Posiadając dane dotyczące jednego przejazdu, średnią prędkość można obliczyć poprzez wyznaczenie odległości między dwoma punktami oraz czasu, który upłynął podczas przejazdu między nimi zgodnie ze wzorem:

gdzie:

* – prędkość w punkcie *i*
* – odległość między punktami *i* oraz *i-1*
* - czas w punkcie *i*
* – czas w punkcie *i-1*

Odległość między dwoma punktami możemy wyznaczyć korzystając ze wzoru:

Gdzie:

* *R* – średni promień Ziemi (*R =* 6371 *km*)
* – szerokość geograficzna punktu *i*
* – długość geograficzna punktu *i*

Jak zostało wspomniane wcześniej przejazd dotyczy konkretnego kursu. Tabela *csiptrace* zawiera kolumnę *daycourseid* dzięki czemu wyszukując przejazd o konkretnym *daycourseid* jesteśmy w stanie odtworzyć trasę przejazdu autobusu a korzystając z dwóch powyższych wzorów również średnią prędkość na poszczególnych odcinkach. Wiąże się to oczywiście z pewną niedokładnością, ponieważ przy takim podejściu zakładamy, że autobus na poszczególnych odcinkach poruszał się po linii prostej, w rzeczywistości zaś mógł to być pewien łuk lub łamana. Jednak z perspektywy jednego z celu projektu, czyli znalezienia punktów zatrzymania pojazdu, nie ma to większego znaczenia. Większe znaczenie ma na pewno *czas próbkowania*. Mogła się bowiem przydarzyć sytuacja, w której autobus podczas czasu pomiędzy dwoma zarejestrowanymi punktami mógł zdążyć się zatrzymać i ruszyć pokonując jednocześnie krótki odcinek drogi. Wtedy nasza wyliczona średnia prędkość nie będzie równa zeru. Jednak z perspektywy bardziej szczegółowego celu projektu, czyli wykrycie lokalizacji przystanku to również może mieć niewielkie znaczenie, ponieważ zazwyczaj autobus stoi na przystanku dłużej niż 10 sekund. Kolejną przesłanką mówiącą o tym, że ma to niewielkie znaczenie jest to, że posiadamy wystarczającą ilość przejazdów, aby te skrajne przypadki nie były bardzo istotne. Przyjęliśmy założenie, że pojazd zatrzymał się, jeżeli średnia prędkość na pewnym odcinku była mniejsza niż 0,1km/h. Jako punkt zatrzymania ustaliliśmy punkt znajdujący się w połowie odcinka pomiędzy dwoma punktami z bazy.

1. **Wyznaczanie faktycznych lokalizacji przystanków**

Baza danych zawiera lokalizacje przystanków ustalone w niewiadomy nam sposób. Zakładając, że faktyczne lokalizacje przystanków znajdują się w dość bliskiej odległości lokalizacji przystanków (parametr algorytmu) z bazy można je wyznaczyć poprzez porównanie z wszystkimi wyznaczonymi punktami zatrzymania. Punkty zatrzymania należałoby zaś pogrupować w obszary zatrzymania oraz zliczyć ilość zatrzymań w danym obszarze. Za kryterium wyboru faktycznej lokalizacji przystanku można wtedy przyjąć pewien sztuczny wskaźnik wyznaczony na podstawie odległości przystanku z bazy od pewnego obszaru zatrzymania oraz ilości zatrzymań pojazdów w tym miejscu. Podejście takie wiąże się jednak z kilkoma mankamentami. Przede wszystkim w pobliżu (bliżej niż wcześniej ustalone 200m) może być przystanek po drugiej stronie i może to wpływać na błędy. Podejście to jest również bardzo kosztowne. Należałoby bowiem, dla każdego przystanku (jest ich niecałe 1000) wyznaczyć odległość od wszystkich punktów zatrzymania wyznaczonych z ponad 7 milionów punktów. Z tego względu zdecydowano się na inne podejście. Dla każdego przystanku istnieje możliwość wyznaczenia wariantów, czyli tras kursów. Dla każdego wariantu można zaś wyznaczyć konkretny kurs. Możliwe jest zatem wyznaczenie wszystkich punktów zatrzymania dla kursów które musiały przebiegać przez dany przystanek. W tym podejściu nie będzie możliwości brania pod uwagę zatrzymań na przystanku po drugiej stronie ulicy. Jest ono również mniej kosztowne, ponieważ zawężona zostaje liczba potrzebnych odległości do policzenia. Rozważana również była możliwość jeszcze większego zawężenia niezbędnych odległości do policzenia, ponieważ możliwe jest wyznaczenie z rozkładu w okolicy którego przystanku powinien znajdować się autobus w danym czasie. Jednak ze względu na potencjalne odchylenia czasowe (spowodowane opóźnieniami, awariami, błędami w rozkładzie jazdy) opcja ta została odrzucona.

Używany algorytm do wyznaczania faktycznych lokalizacji przystanków:

1. Pobranie z bazy wszystkich lokalizacji przystanków.
2. Dla każdego przystanku:
   1. Pobranie wszystkich możliwych ID wariantów(tras) przejazdów dla danego przystanku.
   2. Pobranie wszystkich odbytych kursów dla wyznaczonych wariantów.
   3. Dla każdego kursu wyznaczenie średniej prędkości w poszczególnych punktach oraz odrzucenie punktów gdzie średnia prędkość wynosiła mniej niż *speed\_treshold* (parametr algorytmu). De facto jest to wyznaczenie punktów zatrzymania. Ponieważ operacje pobierania kursów z bazy oraz obliczenia średnich są dość kosztowne wszystkie kursy dla danych wariantów zapisywane są w pamięci, gdyż mogą zostać wykorzystane podczas analizy kolejnych przystanku. Przechowywane dane zawierają ID wariantu oraz listę z kursami nalężącymi do wariantu. Kurs zawiera lokalizacje punktów zatrzymania oraz prędkość.
   4. Utworzenie listy mającej zawierać obszary zatrzymania. Element listy zawiera pierwszy zapisany w niej punkt zatrzymania, ilość zatrzymań w obszarze wokół tego punktów oraz listę z lokalizacjami wszystkich punktów zatrzymania wokół tego punktu.
   5. Jeżeli punkt zatrzymania jest w odległości większej niż *search\_distance* (parametr algorytmu) od lokalizacji przystanku z bazy odrzucić ten punkt (przejście do analizy kolejnego punktu)
   6. Jeżeli punkt zatrzymania znajduje się w odległości mniejszej niż *group\_distance* (parametr algorytmu) od któregoś z punktów zapisanych w liście z punktu **d.** to dla elementu listy zawierającej ten punkt zwiększyć ilość zatrzymań w obszarze oraz dodać ten punktu do listy z lokalizacjami wszystkich punktów zatrzymania wokół tego punktu. W przeciwnym wypadku (jeżeli punkt znajduje się w odległości większej niż równej niż 20m od wszystkich punktów zapisanych w liście z punktu **d.**) należy stworzyć nowy element i dodać go do listy z punktu **d.**
   7. Wprowadzony zostaje wskaźnik, na podstawie którego wyznaczamy obszar który najprawdopodobniej określa położenie przystanku:

Gdzie:

* N – ilość zatrzymań w danym obszarze
* *dist* – odległość od lokalizacji przystanku z bazy (został podniesiony do kwadratu, aby zmniejszyć jego wpływ na wartość wskaźnika *w,* ponieważ istotniejsza jest liczba zatrzymań)
  1. Znalezienie obszaru, dla którego wskaźnik przyjmuje największą wartość
  2. Wyznaczenie średniego punktu na podstawie punktów zatrzymania w tym obszarze. Punkt ten jest wyznaczoną lokalizacją przystanku.

Celem zwiększenia szybkości pobierania danych do bazy zostały dodane odpowiednie indeksy.