[Τίτλος εγγράφου]

[Υπότιτλος εγγράφου]

Στράτος Κ

[Έτος]

# **Ευχαριστίες**

# **Περίληψη**

# **Κατάλογος Σχημάτων**

# **Εισαγωγή**

# **Κεφάλαιο 1: Θεωρητικό Υπόβαθρο της Εφαρμογής**

Το πρόβλημα της πρόβλεψης της κυκλοφοριακής ροής σε ένα οδικό δίκτυο αποτελεί ένα ζήτημα που μπορεί να επιλυθεί με πολλούς τρόπους και μεθοδολογίες. Μάλιστα, υπάρχουν πολλά πρίσματα, υπό τα οποία μπορεί να εξετάσει κάποιος την κυκλοφοριακή ροή στους δρόμους. Για παράδειγμα, η τελευταία εξετάζεται σε μία ολόκληρη πόλη, σε μία μόνο οδό, σε μία εθνική οδό ή σε ένα χωριό;

Στο παρόν κεφάλαιο εξετάζεται η μεθοδολογία που ακολουθήσαμε για την επίτευξη αυτού του στόχου. Συγκεκριμένα, ορίζεται ρητά το πρόβλημα προς επίλυση και η οπτική γωνία υπό την οποία το εξετάζουμε. Δηλώνονται επίσης και οι ερμηνείες σημαντικών εννοιών, η κατανόηση των οποίων καθίσταται αναγκαία για την παρακολούθηση του κειμένου.

## **Παρουσίαση του Προβλήματος**

Η παρούσα μελέτη επικεντρώνεται στο πρόβλημα της πρόβλεψης της κυκλοφοριακής ροής σε ένα οδικό δίκτυο μιας συγκεκριμένης περιοχής. Η έννοια του οδικού δικτύου ταυτίζεται με το σύστημα οδών, δρόμων και διασταυρώσεων της περιοχής. Με τον όρο κυκλοφοριακή ροή εννοούμε το πλήθος των κινούμενων αντικειμένων που διέρχονται από ένα καθορισμένο μονοπάτι του οδικού δικτύου εντός ενός συγκεκριμένου χρονικού διαστήματος. Για την καλύτερη παρακολούθηση του προβλήματος, ορίζουμε  τις ακόλουθες έννοιες:

* ο όρος «**ακμή**» αναφέρεται σε ένα τμήμα του οδικού δικτύου που βρίσκεται ανάμεσα σε δύο διασταυρώσεις.
* ο όρος «**μονοπάτι**»| αναφέρεται σε μια ακολουθία από συνεχόμενες ακμές. Η συνεχόμενη φύση του μονοπατιού σημαίνει ότι η ακμή που ακολουθεί την προηγούμενη ακμή αποτελεί επίσης την αρχή του επόμενου τμήματος του μονοπατιού, διασφαλίζοντας τη συνεχόμενη σύνδεση των ακμών. Ένα μονοπάτι περιλαμβάνει τουλάχιστον δύο ακμές.
* ο όρος «**κινούμενο αντικείμενο**» υποδηλώνει ένα κινητό στοιχείο που διασχίζει το οδικό δίκτυο.

Στο εξής, η εμβέλεια των παραπάνω ορισμών θα έχει ισχύ σε όλο το κείμενο.

Για την επίλυση του προβλήματος της πρόβλεψης της κυκλοφοριακής ροής σε ένα οδικό δίκτυο, ακολουθούμε μία προσέγγιση που, εξ όσων γνωρίζουμε, είναι μοναδική. Με άλλα λόγια, δεν έχουμε καταφέρει να εντοπίσουμε έρευνες άλλων επιστημόνων που να ακολουθούν παρόμοια μεθοδολογία για την επίλυση του προβλήματος της πρόβλεψης της κυκλοφοριακής ροής.

Πιο συγκεκριμένα, προσπαθούμε να ορίσουμε έναν αριθμό από μοναδικά μονοπάτια μέσα στο οδικό δίκτυο, για τα οποία μετράμε την κυκλοφοριακή ροή σε κάθε μονοπάτι και προσπαθούμε να προβλέψουμε την μελλοντική κυκλοφοριακή ροή σε κάθε ένα από αυτά. Για τη μέτρηση της ροής της κυκλοφορίας σε κάθε μονοπάτι, χρησιμοποιούμε τη μεθοδολογία των Αυστηρών Ερωτημάτων Μονοπατιού (ΑΕΜ). Η ιδέα πίσω από αυτή την μεθοδολογία αναλύεται στο επόμενο υποκεφάλαιο. Έπειτα, ανάγουμε το πρόβλημα σε χρονοσειρές και προσπαθούμε να εφαρμόσουμε διάφορες τεχνικές και αλγορίθμους μηχανικής μάθησης, για να διεξάγουμε την πρόβλεψη.

## **Αυστηρά Ερωτήματα Μονοπατιού (ΑΕΜ)**

Ο όρος «Αυστηρά Ερωτήματα Μονοπατιού» ή στα αγγλικά ως «Strict Path Queries» (SPQs) αναφέρεται σε μια διαδικασία αναζήτησης που εκτελείται σε δεδομένα τροχιών κινούμενων αντικειμένων με στόχο την ανάκτηση όλων των τροχιών που διέρχονται αυστηρά από ένα προκαθορισμένο μονοπάτι (ακολουθώντας δηλαδή πιστά τις ακμές που αποτελείται το μονοπάτι μία προς μία και χωρίς να παρεκκλίνουν καθόλου από το μονοπάτι αυτό) μέσα σε ένα προεπιλεγμένο χρονικό διάστημα. [1], [2]

Στα πλαίσια της δικής τους έρευνας, οι συγγραφείς, για να εξετάσουν την μεθοδολογία τους χρησιμοποιούν δεδομένα κίνησης κινούμενων αντικειμένων, τα οποία καταγράφονται μέσω ενός συστήματος GPS, παρέχοντας πληροφορίες σχετικά με τη θέση τους στον τρισδιάστατο χώρο (x, y, t). Το x αντιστοιχεί στο γεωγραφικό μήκος (longitude), το y αντιστοιχίζεται στο γεωγραφικό πλάτος (latitude) και το t αναπαριστά τον χρόνο (time). Η καταγραφή αυτών των δεδομένων κίνησης για κάθε κινούμενο όχημα ακολουθεί σταθερή περιοδικότητα.  Κάθε αναφορά θέσης από το σύστημα GPS αναπαρίστανται από μια πλειάδα παραμέτρων της μορφής loc=(moid, ts, pos), όπου:

* το στοιχείο «moid» (από το moving object id) αναπαριστά το αναγνωριστικό του κινούμενου αντικειμένου,
* το στοιχείο «ts» αντιστοιχεί σε συγκεκριμένη χρονική στιγμή, και

* το στοιχείο «pos» (από το position) δηλώνει τη θέση του κινούμενου αντικειμένου κατά τη χρονική στιγμή «ts» με χρήση χωρικών συντεταγμένων (latitude και longitude).

Μέσω μιας διαδικασίας αντιστοίχισης σημείων GPS σε ψηφιακούς χάρτες, τα αρχικά σημεία GPS που παράγονται για κάθε κινούμενο όχημα, αντιστοιχίζονται σε μια ακολουθία ακμών εντός του οδικού δικτύου. Με την εφαρμογή αυτής της διαδικασίας, διαμορφώνεται μια τροχιά εντός του οδικού δικτύου για κάθε κινούμενο όχημα.

Κάθε τροχιά που συσχετίζεται με ένα συγκεκριμένο κινούμενο όχημα αποτελείται από πολλές εγγραφές της μορφής locmm=(tid, eid, tsenter, tsleave). Στην εν λόγω αναπαράσταση το «tid» δηλώνει το αναγνωριστικό της τροχιάς, το «eid» αναφέρεται στο αναγνωριστικό της ακμής στο οδικό δίκτυο, ενώ τα «tsenter» και «tsleave» αναφέρονται στους χρόνους εισόδου και εξόδου του κινούμενου αντικειμένου από την ακμή με αναγνωριστικό «eid» αντίστοιχα. Επομένως, τα αρχικά δεδομένα που καταγράφονται από το σύστημα GPS υποβάλλονται σε μια διαδικασία μετατροπής σε εγγραφές locmm. Μια τροχιά t προκύπτει ως ένα σύνολο τέτοιων εγγραφών, δηλαδή η τροχιά t ορίζεται ως εξής: t = [locmm1, locmm2, …, locmmn].

Όσον αφορά την επίδοση της μεθόδου, οι συγγραφείς υποστηρίζουν ότι τα αποτελέσματα που παρέχονται από τον αλγόριθμο Strict Path Queries (SPQs) μπορούν να θεωρηθούν ικανοποιητικά. Ο συγκεκριμένος αλγόριθμος διακρίνεται για την υψηλή του ακρίβεια και ταχύτητα εκτέλεσής του σε σύγκριση με άλλες προσεγγίσεις που έγιναν στα πλαίσια της ίδιας έρευνας.

Είναι σημαντικό να αναφερθεί ότι τα ΑΕΜ είναι απαραίτητα για πολλούς λόγους. Αφενός, μας παρέχουν την πληροφορία για το πόσες τροχιές διέσχισαν ένα συγκεκριμένο μονοπάτι από την αρχή του έως και το τέλος του, χωρίς να παρεκκλίνουν καθόλου από αυτό. Θέτοντας το διαφορετικά, μας βοηθούν να μετρήσουμε την ποσότητα της ροής των κινούμενων αντικειμένων εντός ενός ολοκλήρου μονοπατιού με μεγάλη ακρίβεια. Αφετέρου, μπορούμε να ορίσουμε για πιο χρονικό διάστημα επιλέγουμε να γίνει η αναζήτηση της κυκλοφοριακής ροής. Επομένως, με την χρήση αυτής της μεθοδολογίας μπορούν να εξαχθούν σημαντικά συμπεράσματα σχετικά με τη συμπεριφορά της κυκλοφορίας και της μετακίνησης κινούμενων οχημάτων, αναγνωρίζοντας ποια μονοπάτια διασχίζονται συχνότερα κατά τη διάρκεια διαφόρων χρονικών περιόδων, όπως οι ώρες αιχμής. Επιπλέον, μπορούν να ανακαλυφθούν μοτίβα συμφόρησης και να εντοπιστούν ανωμαλίες στην κυκλοφορία, επιτρέποντας την αντιμετώπιση πιθανών προβλημάτων.

## **Εφαρμογή των ΑΕΜ στην Μελέτη μας**

Όπως δηλώθηκε παραπάνω, η έρευνα που διεξάγουμε αναφέρεται στην πρόβλεψη της κυκλοφοριακής ροής κινούμενων αντικειμένων μέσα σε ολόκληρα μονοπάτια του οδικού δικτύου μίας περιοχής. Ο τρόπος με τον οποίο γίνεται αυτό περιγράφεται περιληπτικά στο παρόν κεφάλαιο, ενώ αναλυτικότερη εξήγηση θα δοθεί σε επόμενο κεφάλαιο.

 Αρχικά, τα δεδομένα που έχουμε στην διάθεσή μας είναι ένα σύνολο από αναφορές θέσεων GPS διαφόρων κινούμενων αντικειμένων. Κάθε αναφορά θέσης αποτελείται από μία πλειάδα τεσσάρων στοιχείων (id,lat,lon,time) όπου το στοιχείο «id» αναφέρεται στο αναγνωριστικό του κινούμενου αντικειμένου, το στοιχείο «lat» αναφέρεται στο γεωγραφικό πλάτος, το στοιχείο «lon» παραπέμπει στο γεωγραφικό μήκος και το στοιχείο «time» στην χρονική στιγμή που έγινε η καταγραφή της θέσης του κινούμενου αντικειμένου.

Τα δεδομένα GPS που μας δίνονται αρχικά δεν έχουν αντιστοιχηθεί στο οδικό δίκτυο της περιοχής που διερευνούμε. Για να γίνει αυτό, χρειάζεται να γίνει μία αντιστοίχιση των GPS δεδομένων σε ακμές του οδικού δικτύου.  Αφού γίνει η διαδικασία της αντιστοίχισης, ορίζουμε ένα αριθμό από τυχαία και μοναδικά μονοπάτια τυχαίου μήκους που θα δημιουργηθούν. Ο τρόπος με τον οποίο παράγεται ένα μονοπάτι ακολουθεί αυστηρά τον ορισμό που δόθηκε παραπάνω, ενώ οι ακμές από τις οποίες απαρτίζεται κάθε μονοπάτι προκύπτουν άμεσα από τα δεδομένα.

Στην συνέχεια, δημιουργούμε σταθερά χρονικά διαστήματα, για κάθε ένα από τα οποία μετράμε την κυκλοφοριακή ροή σε όλα τα μονοπάτια που έχουν δημιουργηθεί. Η μέτρηση της κυκλοφοριακής ροής των κινούμενων αντικειμένων πραγματοποιείται με την χρήση των ΑΕΜ. Με αυτόν τον τρόπο, τα αρχικά GPS δεδομένα μετατρέπονται πλέον σε δεδομένα χρονοσειρών, δηλαδή σε μια σειρά από μετρήσεις που καταγράφονται με χρονική σειρά και ανά σταθερά χρονικά διαστήματα μεταξύ τους. Οι χρονοσειρές αυτές αποτελούν τα ιστορικά δεδομένα κυκλοφοριακής ροής για κάθε μονοπάτι. Τέλος, εφαρμόζοντας αλγορίθμους μηχανικής και βαθιάς μάθησης, μπορούμε να προβλέψουμε την μελλοντική ροή της κυκλοφορίας σε όλα τα μονοπάτια που μας απασχολούν χρησιμοποιώντας τα ιστορικά δεδομένα.

Αξίζει να σημειωθεί ότι όσα περισσότερα μοναδικά μονοπάτια δημιουργηθούν, τόσο αυξάνεται και η πιθανότητα να καλυφθεί ολόκληρο το προς μελέτη οδικό δίκτυο. Επομένως, η παρούσα έρευνα, αν και εστιάζει στην πρόβλεψη της κυκλοφοριακής ροής εντός ολόκληρων μονοπατιών, μπορεί να χρησιμοποιηθεί και για την πρόβλεψη της κυκλοφοριακής ροής σε ένα σύνολο μονοπατιών που απαρτίζουν ολόκληρο το οδικό δίκτυο.

**Κεφάλαιο 2: Σχετικές Εργασίες**

Σε αυτό το κεφάλαιο παρατίθενται σχετικές εργασίες που έχουν υλοποιηθεί από άλλους ερευνητές, οι οποίοι με την σειρά τους λύνουν το πρόβλημα της πρόβλεψης της ροής κυκλοφορίας σε έναν οδικό δίκτυο. Οι περισσότερες από αυτές τις υλοποιήσεις υιοθετούν αναδρομικά νευρωνικά δίκτυα με μνήμη (LSTMs), απλά νευρωνικά δίκτυα (MLPs) και στατιστικά μοντέλα, όπως είναι τα μοντέλα ARIMA και SARIMA. Το κεφάλαιο αυτό οργανώνεται ως εξής: αναφέρεται, για κάθε έρευνα, το πρόβλημα που λύνεται και η μέθοδος με την οποία επιλύεται.

**Έγγραφο 1**

**Τίτλος:** Predicting Short-term Traffic Flow by Long Short-Term Memory Recurrent Neural Network

**Εισαγωγή**

Η εργασία αυτή εξετάζει τη σημασία της βραχυπρόθεσμης πρόβλεψης της ροής της κυκλοφορίας στις ευφυείς μεταφορές (ITS) και την εφαρμογή της στη διαχείριση της κυκλοφοριακής συμφόρησης, τη μείωση της ρύπανσης και την ενίσχυση της οδικής ασφάλειας. Επιπρόσθετα, επισημαίνει τις προκλήσεις της ακριβούς πρόβλεψης της εξαιρετικά μη γραμμικής και στοχαστικής φύσης της ροής κυκλοφορίας, λόγω διαφόρων παραγόντων, όπως οι καιρικές συνθήκες και η μορφολογία του εδάφους. Στόχος είναι η πρόβλεψη της κυκλοφοριακής ροής σε ένα συγκεκριμένο χρονικό διάστημα με βάση ιστορικά δεδομένα ανά διαστήματα των 15 λεπτών.

Στο πλαίσιο της παρούσας εργασίας, η κυκλοφοριακή ροή αναφέρεται στην κίνηση των οχημάτων σε έναν αυτοκινητόδρομο ανά διαστήματα των 15 λεπτών και αντιπροσωπεύει τον όγκο των οχημάτων που διέρχονται από ένα συγκεκριμένο σταθμό παρατήρησης σε μια δεδομένη χρονική στιγμή. Η εργασία έχει ως στόχο να προβλέψει αυτή την βραχυπρόθεσμη κυκλοφοριακή ροή με ακρίβεια, ώστε να παρέχει έγκαιρες και πολύτιμες πληροφορίες για διάφορους ενδιαφερόμενους, συμπεριλαμβανομένων των μεμονωμένων ταξιδιωτών, των επιχειρήσεων και των κυβερνητικών υπηρεσιών.

**Μεθοδολογία**

Στην έρευνα αυτή χρησιμοποιείται το μοντέλο Long Short Term Memory Recurrent Neural Network (LSTM RNN) για την βραχυπρόθεσμη πρόβλεψη της ροή κυκλοφορίας. Το LSTM RNN είναι ένας τύπος νευρωνικού δικτύου που είναι κατάλληλος για εργασίες πρόβλεψης χρονοσειρών. Για την εκτέλεση της πρόβλεψης, το μοντέλο LSTM RNN λαμβάνει ως είσοδο δεδομένα ιστορικής ροής κυκλοφορίας, τα οποία περιλαμβάνουν πληροφορίες σχετικά με προηγούμενους όγκους κίνησης εντός του επιλεγμένου χρονικού διαστήματος (π.χ. τα τελευταία 30 λεπτά ή δύο χρονικά βήματα στο παρελθόν).

Το μοντέλο μαθαίνει τα μοτίβα και τις εξαρτήσεις που υπάρχουν στα ιστορικά δεδομένα για να προβλέψει τη ροή κυκλοφορίας για το επόμενο διάστημα των 15 λεπτών. Η αρχιτεκτονική LSTM RNN περιλαμβάνει μπλοκ μνήμης που επιτρέπει στο δίκτυο να συλλαμβάνει και να αποθηκεύει πληροφορίες για μεγαλύτερες χρονικές περιόδους, αντιμετωπίζοντας την πρόκληση των χρονικών εξαρτήσεων στη ροή της κυκλοφορίας. Σε αντίθεση με τα παραδοσιακά μοντέλα νευρωνικών δικτύων που βασίζονται σε προκαθορισμένα και στατικά μήκη εισόδων, το LSTM RNN προσδιορίζει δυναμικά τις βέλτιστες χρονικές καθυστερήσεις και τις ενσωματώνει στη διαδικασία πρόβλεψης.

Το μοντέλο LSTM RNN εκπαιδεύεται χρησιμοποιώντας το σύνολο δεδομένων Caltrans Performance Measurement System (PeMS), παρέχοντας στον ερευνητή έναν μεγάλο αριθμό ιστορικών δεδομένων κυκλοφοριακής ροής. Η απόδοση του μοντέλου αξιολογείται και σε σύγκριση με άλλα μοντέλα, όπως το Random Walk (RW), τον αλγόριθμο Support Vector Machine (SVM) και τα Feed Forward Neural Networks (FFNN) για την αξιολόγηση της ακρίβειάς του και δυνατότητες γενίκευσης.

Τα αποτελέσματα δείχνουν την υπεροχή του μοντέλου LSTM RNN, όσον αφορά την ακρίβεια πρόβλεψης και τις δυνατότητες γενίκευσης σε άλλα σύνολα δεδομένων (π.χ. δεδομένα ελέγχου). Το LSTM RNN υπερτερεί έναντι άλλων μοντέλων στην αποτύπωση των μη γραμμικών και στοχαστικών χαρακτηριστικών της ροής κυκλοφορίας.

**Έγγραφο 2**

**Τίτλος:** Supervised deep learning based for traffic flow prediction

**Εισαγωγή**

Σε αυτή την έρευνα, η ροή κυκλοφορίας ορίζεται ως η κίνηση, η ποσότητα και τα χαρακτηριστικά των οχημάτων ή /και πεζών στο οδικό δίκτυο μίας συγκεκριμένης περιοχής και χρονικής περιόδου. Πιο συγκεκριμένα, η κυκλοφοριακή ροή περιλαμβάνει πλη όπως το όχημα ταχύτητα, πυκνότητα και όγκο και παρέχει πληροφορίες για το επίπεδο συμφόρησης, τα μοτίβα κυκλοφορίας και συνολικά αποτελεσματικότητα των συστημάτων μεταφορών. Η ακριβής πρόβλεψη της ροής της κυκλοφορίας είναι ζωτικής σημασίας για την αποτελεσματική κυκλοφορία διαχείριση, σχεδιασμός και λήψη αποφάσεων προκειμένου να αντιμετωπιστούν ζητήματα όπως η συμφόρηση, τα ατυχήματα, ο αέρας ρύπανση και κατανάλωση ενέργειας στις μητροπολιτικές περιοχές.

Μεθοδολογία

Η μεθοδολογία που χρησιμοποιείται σε αυτή τη μελέτη για την πρόβλεψη της ροής της κυκλοφορίας είναι μια εποπτευόμενη βαθιά μάθηση Προσέγγιση βασισμένης πρόβλεψης ροής κυκλοφορίας (SDLTFP) χρησιμοποιώντας ένα πλήρως συνδεδεμένο βαθύ νευρωνικό δίκτυο (FC-DNN). Οι συγγραφείς αναγνωρίζουν τη σημασία της έγκαιρης πρόβλεψης και αντιμετωπίζουν πιθανές προκλήσεις όπως π.χ υπερβολική και χρονοβόρα προπόνηση.

Για να μετριαστεί η υπερβολική προσαρμογή και να βελτιωθεί η αποτελεσματικότητα της προπόνησης, η ομαλοποίηση παρτίδων (BN) και η εγκατάλειψη οι τεχνικές ενσωματώνονται στο νευρωνικό δίκτυο. Το BN βοηθά στην ομαλοποίηση των εισόδων σε κάθε επίπεδο, μειώνοντας μετατόπιση εσωτερικής συμμεταβλητής και κάνοντας την προπονητική διαδικασία ταχύτερη και πιο σταθερή. Η εγκατάλειψη είναι συνηθισμένη απενεργοποιήστε τυχαία ένα κλάσμα νευρώνων κατά τη διάρκεια της προπόνησης, αποτρέποντας την εξάρτηση από συγκεκριμένα μοτίβα και ενισχύοντας την ικανότητα γενίκευσης του δικτύου.

Το Stochastic Gradient Descent (SGD) με ορμή χρησιμοποιείται για ενημερώσεις βάρους στην προπόνηση επεξεργάζομαι, διαδικασία. Το SGD υπολογίζει τη διαβάθμιση της συνάρτησης απώλειας για κάθε δείγμα προπόνησης και εκτελεί το βάρος ενημερώνει αναλόγως. Η συμπερίληψη της ορμής βοηθά στην επιτάχυνση της σύγκλισης και εξομαλύνει το τροχιά κατά τη βελτιστοποίηση.

Τα πειράματα διεξάγονται χρησιμοποιώντας ανοιχτά δεδομένα ως ιστορικά δεδομένα κίνησης, τα οποία χρησιμοποιούνται για εκπαίδευση την προτεινόμενη μέθοδο και μοντέλο SDLTFP. Το μέσο απόλυτο ποσοστό σφάλματος (MAPE) χρησιμοποιείται ως ένα μέτρηση αξιολόγησης για την αξιολόγηση της ακρίβειας των προβλέψεων ροής κυκλοφορίας. Το MAPE υπολογίζεται και για τα δύο δείγματα δεδομένα (εντός 5%) και δεδομένα εκτός δείγματος (που κυμαίνονται μεταξύ 15% και 20%), που καταδεικνύουν την ικανότητα πρόβλεψης της ροής της κυκλοφορίας με ακρίβεια.

Χρησιμοποιώντας την προτεινόμενη μεθοδολογία, οι συγγραφείς στοχεύουν να παρέχουν μια ακριβή ροή κυκλοφορίας πρόβλεψη για τη διευκόλυνση του Συστήματος Διαχείρισης Κυκλοφορίας (TMS) των Ευφυών Συστημάτων Μεταφορών (ITS). Η ενσωμάτωση των τεχνικών BN και Dropout όχι μόνο ενισχύει την προπονητική διαδικασία αλλά βοηθά και μειώστε την υπερπροσαρμογή, με αποτέλεσμα αξιόπιστες προβλέψεις ροής κυκλοφορίας.

## **Έγγραφο 3**

**Τίτλος:** Traffic Flow prediction with big data: A deep learning approach

**Εισαγωγή**

Στο παρόν έγγραφο, η ροή κυκλοφορίας αναφέρεται στην κίνηση των οχημάτων σε δίκτυα μεταφορών, όπως οι δρόμοι ή οι αυτοκινητόδρομοι. Συγκεκριμένα αναφέρεται στην ποσότητα των οχημάτων που διέρχονται από μια συγκεκριμένη τοποθεσία σε διαφορετικές χρονικά διαστήματα. Η παρατηρούμενη ποσότητα ροής κυκλοφορίας, η οποία συμβολίζεται ως Xt\_i, αντιπροσωπεύει τον όγκο κυκλοφορίας ή άλλες μετρικές που σχετίζονται με την κυκλοφορία που μετρήθηκαν στην i-οστή θέση παρατήρησης κατά τη διάρκεια του t-οστού χρονικού διαστήματος.

Η πρόβλεψη της κυκλοφοριακής ροής αποσκοπεί στην παροχή ακριβών και έγκαιρων πληροφοριών σχετικά με τις αναμενόμενες συνθήκες κυκλοφορίας, η οποία είναι ζωτικής σημασίας για τη διαχείριση των μεταφορών, τα ευφυή συστήματα μεταφορών και διάφορες εφαρμογές στον έλεγχο και τη βελτιστοποίηση της κυκλοφορίας.

Μεθοδολογία

Η μεθοδολογία που χρησιμοποιείται στην εργασία για την πρόβλεψη της κυκλοφοριακής ροής περιλαμβάνει την εφαρμογή ενός στοιβαγμένου αυτοκωδικοποιητή (SAE), ο οποίος είναι μια αρχιτεκτονική βαθιάς μάθησης. Το μοντέλο SAE εκπαιδεύεται με τη χρήση κυκλοφοριακών δεδομένων ροής που συλλέγονται από τη βάση δεδομένων του συστήματος μέτρησης επιδόσεων (PeMS) της Caltrans. Εδώ είναι ένα ανάλυση της μεθοδολογίας:

1. Συλλογή δεδομένων: Τα δεδομένα κυκλοφοριακής ροής συλλέγονται από τη βάση δεδομένων PeMS της Caltrans. Αυτή η βάση δεδομένων περιέχει συγκεντρωτικές πληροφορίες από ένα μεγάλο αριθμό ανιχνευτών που βρίσκονται σε διάφορες μεταφορικές δίκτυα.
2. Προεπεξεργασία: Τα δεδομένα που συλλέγονται υποβάλλονται σε προεπεξεργασία για την απομάκρυνση τυχόν ακραίων τιμών ή ασυνεπειών. Είναι στη συνέχεια χωρίζονται σε σύνολα εκπαίδευσης και δοκιμής για την αξιολόγηση του μοντέλου.
3. Αρχιτεκτονική μοντέλου: Το μοντέλο SAE εισάγεται ως η πρωταρχική μεθοδολογία για τη ροή κυκλοφορίας πρόβλεψης. Το SAE είναι ένα μοντέλο βαθιάς μάθησης που αποτελείται από πολλαπλά στρώματα αυτοκωδικοποιητών. Ένα αυτοκωδικοποιητής είναι ένας τύπος νευρωνικού δικτύου που στοχεύει στην ανακατασκευή της εισόδου του στο στρώμα εξόδου, ουσιαστικά μαθαίνοντας να αναπαριστά τα δεδομένα σε συμπιεσμένη μορφή. Το μοντέλο SAE χρησιμοποιεί τους αυτοκωδικοποιητές ως δομικά στοιχεία για τη δημιουργία ενός βαθιού δικτύου.
4. Διαδικασία εκπαίδευσης: Το μοντέλο SAE εκπαιδεύεται με άπληστο τρόπο κατά στρώματα, όπου κάθε στρώμα είναι προ-εκπαιδεύεται ξεχωριστά πριν από τη λεπτομερή ρύθμιση ολόκληρου του μοντέλου. Αυτή η διαδικασία επιτρέπει στο μοντέλο να μάθει ιεραρχικές αναπαραστάσεις των χαρακτηριστικών της ροής κυκλοφορίας.
5. Αξιολόγηση: Η απόδοση του μοντέλου SAE αξιολογείται με τη χρήση διαφόρων μετρικών, όπως η πρόβλεψη ακρίβεια πρόβλεψης και μέσο απόλυτο σφάλμα. Η απόδοση του μοντέλου συγκρίνεται με άλλες μεθόδους πρόβλεψης, συμπεριλαμβανομένων των μοντέλων BP NN, τυχαίου περιπάτου, SVM και RBF NN.
6. Αποτελέσματα και ανάλυση: Στην εργασία παρουσιάζονται τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν με τη χρήση του SAE μοντέλο. Επισημαίνεται η ανώτερη απόδοση του μοντέλου SAE στην ακριβή πρόβλεψη της ροής κυκλοφορίας σε συνθήκες βαριάς και μέσης κυκλοφορίας. Ωστόσο, σημειώνει ότι το μοντέλο δυσκολεύεται να αποδώσει καλά σε χαμηλές συνθήκες ροής κυκλοφορίας, παρόμοια με τις υπάρχουσες μεθόδους πρόβλεψης.

Συνολικά, η μεθοδολογία επικεντρώνεται στην εκπαίδευση και την αξιολόγηση του μοντέλου SAE με τη χρήση πραγματικών δεδομένων ροής κυκλοφορίας. Καταδεικνύει τις δυνατότητες των τεχνικών βαθιάς μάθησης, και συγκεκριμένα του μοντέλου SAE, για την ακριβή πρόβλεψη της ροής της κυκλοφορίας. Η εργασία παρέχει πληροφορίες σχετικά με τα δυνατά σημεία και τους περιορισμούς του προτεινόμενου μεθοδολογίας και τη συγκρίνει με τις υπάρχουσες μεθόδους πρόβλεψης.

**Έγγραφο 4:** DeepCrowd -A Deep Model for Large-Scale Citywide Crowd Density and Flow Prediction

Εισαγωγή

Το πρόβλημα που αντιμετωπίζεται στην εργασία είναι η πρόβλεψη της ροής του πλήθους ή της κυκλοφορίας σε μια πόλη σε επίπεδο πόλης, χρησιμοποιώντας δεδομένα μεγάλης κινητικότητας που παράγονται από πηγές όπως κινητά τηλέφωνα, συστήματα πλοήγησης αυτοκινήτων, και αισθητήρες κυκλοφορίας. Οι συγγραφείς υπογραμμίζουν τη σημασία τέτοιων προβλέψεων στη διαχείριση έκτακτης ανάγκης, ρύθμιση της κυκλοφορίας και τον αστικό σχεδιασμό. Προτείνουν μια λύση που περιλαμβάνει τη διαίρεση μιας μεγάλης αστικής περιοχής σε λεπτόκοκκα πλέγματα και την αναπαράσταση της πυκνότητας και της ροής του πλήθους σε ολόκληρη την πόλη ως τετραδιάστατη τανυστή.

Οι συγγραφείς ορίζουν τη ροή κυκλοφορίας ως την κίνηση των ατόμων εντός των πλεγμάτων-πλεγμάτων. Οι εισάγουν την έννοια της πυκνότητας του πλήθους, η οποία αναφέρεται στον αριθμό των ατόμων που βρίσκονται σε ένα συγκεκριμένο πλέγμα-πλέγμα σε μια συγκεκριμένη χρονική στιγμή. Εξετάζεται επίσης η ροή εισόδου-εξόδου του πλήθους, η οποία αντιπροσωπεύει την κίνηση των ατόμων μεταξύ διαδοχικών χρονικών σημείων εντός ενός συγκεκριμένου πλέγματος ματιών.

Μεθοδολογία

Για την αντιμετώπιση αυτού του προβλήματος, οι συγγραφείς προτείνουν ένα νέο μοντέλο βαθιάς μάθησης που ονομάζεται DeepCrowd, το οποίο αξιοποιεί αρχιτεκτονικές πυραμίδας και μηχανισμούς προσοχής υψηλής διάστασης που βασίζονται σε Convolutional LSTM.

Η μελέτη χρησιμοποιεί ένα μεγάλης κλίμακας συγκεντρωτικό σύνολο δεδομένων ανθρώπινης κινητικότητας που δημιουργήθηκε από ένα πραγματικό κόσμο εφαρμογή smartphone. Αυτό το σύνολο δεδομένων προσφέρει πλεονεκτήματα σε σχέση με τα υπάρχοντα, συμπεριλαμβανομένου ενός μεγαλύτερου αριθμού πλέγματος, λεπτότερη κοκκομετρία στο μέγεθος του πλέγματος και μεγαλύτερο αριθμό δειγμάτων χρηστών. Με την αναπαράσταση των πυκνότητα του πλήθους σε όλη την πόλη ως τετραδιάστατος τανυστής (χρονικό βήμα, ύψος, πλάτος, κανάλι), παρόμοια με τα βίντεο δεδομένα, οι συγγραφείς είναι σε θέση να συλλάβουν και να αναλύσουν τις χωροχρονικές πτυχές του πλήθους και της κυκλοφορίας μοτίβα.

Η μεθοδολογία του DeepCrowd περιλαμβάνει τη χρήση αρχιτεκτονικών πυραμίδας και υψηλής διάστασης μηχανισμούς προσοχής. Οι πυραμίδες ConvLSTM επιτρέπουν την εξαγωγή χαρακτηριστικών πολλαπλών κλιμάκων συνδυάζοντας χαρακτηριστικά από διαφορετικά επίπεδα της πυραμίδας. Αυτό συμβάλλει στη σύλληψη τόσο της παγκόσμιας όσο και της τοπικής χωρικής πληροφορίας για την πρόβλεψη της πυκνότητας και της ροής του πλήθους. Επιπλέον, το μπλοκ προσοχής 4D υψηλής διάστασης ενισχύει την ικανότητα του μοντέλου να εστιάζει σε πληροφοριακές περιοχές στις χωρικές και χρονικές διαστάσεις. Ο μηχανισμός earlyfusion συνδυάζει διαφορετικές μορφές δεδομένων για ακριβέστερες προβλέψεις.

Για την αξιολόγηση των επιδόσεων του DeepCrowd, οι συγγραφείς διεξάγουν ενδελεχή πειράματα και συγκρίνουν τα αποτελέσματά τους με πολλαπλές σύγχρονες μεθόδους. Τέσσερις μετρικές, MSE, RMSE, MAE και MAPE, χρησιμοποιούνται για την αξιολόγηση. Τα αποτελέσματα των πειραμάτων αποδεικνύουν την αποτελεσματικότητα και την αποδοτικότητα της προτεινόμενου μοντέλου για την πρόβλεψη πλήθους σε ολόκληρη την πόλη σε δύο μεγάλες πόλεις, το Τόκιο και την Οσάκα.Συνοπτικά, η εργασία παρουσιάζει ένα πλήρες πλαίσιο για την πρόβλεψη της πυκνότητας του πλήθους σε ολόκληρη την πόλη και ροής, από την προεπεξεργασία δεδομένων έως τα μοντέλα βαθιάς μάθησης. Παρουσιάζει ένα νέο μοντέλο βαθιάς μάθησης, DeepCrowd, το οποίο ενσωματώνει αρχιτεκτονικές πυραμίδας και μηχανισμούς προσοχής υψηλής διάστασης. Το συγγραφείς δημιουργούν και δημοσιεύουν επίσης ένα νέο σύνολο δεδομένων υψηλής ποιότητας για την πρόβλεψη της πυκνότητας και της ροής του πλήθους. Το αποτελέσματα αξιολόγησης δείχνουν την υπεροχή του DeepCrowd σε σύγκριση με τις υπάρχουσες μεθόδους.

**GitHub Link:** https://github.com/deepkashiwa20/DeepCrowd

**Έγγραφο 5:** Hybrid LSTM Neural Network for Short-Term Traffic Flow Prediction

Εισαγωγή

Στο παρόν έγγραφο, οι συγγραφείς ορίζουν τη ροή της κυκλοφορίας ως την κίνηση των οχημάτων μέσω ενός δικτύου κυκλοφορίας κατά τη διάρκεια μιας συγκεκριμένης χρονικής περιόδου. Η ροή μετράται συνήθως με βάση τον αριθμό των οχημάτων που διέρχονται από ένα συγκεκριμένη θέση ή τμήμα οδού.

Μεθοδολογία

Η μεθοδολογία που ακολουθείται στην εργασία περιλαμβάνει δύο βασικά στοιχεία: την πειραματική περιβάλλον και το σύνολο δεδομένων, και τους δείκτες αξιολόγησης που χρησιμοποιούνται.

Πρώτον, περιγράφεται το πειραματικό περιβάλλον, το οποίο περιλαμβάνει το υλικό, το λογισμικό και το γλώσσα προγραμματισμού που χρησιμοποιείται. Το υλικό που χρησιμοποιήθηκε για τα πειράματα είναι ένας Intel Core i5-3210M 2,5 GHz με διπύρηνο επεξεργαστή και μνήμη 4 GB. Το λειτουργικό σύστημα είναι τα Windows 7 64 bit. Το σύστημα προγραμματισμού γλώσσα που επιλέχθηκε για την υλοποίηση είναι η Python και το πλαίσιο βαθιάς μάθησης που χρησιμοποιήθηκε είναι το Keras. Αυτό το περιγραφή παρέχει το απαραίτητο πλαίσιο για την κατανόηση της πειραματικής διάταξης και εξασφαλίζει αναπαραγωγιμότητα.

Στη συνέχεια, περιγράφεται το σύνολο δεδομένων που χρησιμοποιήθηκε στα πειράματα. Το σύνολο δεδομένων αποτελείται από τοπικό οδικό δίκτυο δεδομένα κυκλοφορίας από την 1η έως την 8η Σεπτεμβρίου στην περιοχή Yunyan της πόλης Guiyang, επαρχία Guizhou. Το σύνολο δεδομένων περιέχει πάνω από 230.000 εγγραφές οχημάτων. Κάθε εγγραφή στο σύνολο δεδομένων περιλαμβάνει τρία πεδία: Τρέχον κόμβος (TrafficLightID), κόμβος προέλευσης (FromID) και ροή κυκλοφορίας (traffic\_flow). Το πεδίο Current Node αναπαριστά το αναγνωριστικό του σημερινού φωτεινού σηματοδότη του οχήματος, ενώ το πεδίο Source Node αναπαριστά το το αναγνωριστικό του φωτεινού σηματοδότη του οχήματος την τελευταία στιγμή. Το πεδίο Traffic Flow είναι ένας πίνακας διάστασης 1680, που αναπαριστά τις ροές κυκλοφορίας ενός οδικού τμήματος σε διάστημα 30 δευτερολέπτων. Για παράδειγμα, traffic\_flow[0] αντιπροσωπεύει τη ροή κυκλοφορίας από τη διασταύρωση FromID προς τη διασταύρωση TrafficLightID τη χρονική στιγμή t=0. Παρέχεται επίσης η δομή του τοπικού οδικού δικτύου που εμπλέκεται στο σύνολο δεδομένων, η οποία απεικονίζεται στο σχήμα 3, όπου οι κοίλοι κύκλοι αντιπροσωπεύουν διασταυρώσεις, οι πλευρές τους επισημαίνονται με μοναδικούς αριθμούς αναγνώρισης, και οι συμπαγείς γραμμές αντιπροσωπεύουν τα οδικά τμήματα. Αυτή η λεπτομερής περιγραφή του συνόλου δεδομένων επιτρέπει στους αναγνώστες να να κατανοήσουν τα χαρακτηριστικά των κυκλοφοριακών δεδομένων και τη δομή του δικτύου που συμμετέχουν στα πειράματα.

Επιπλέον, οι δείκτες αξιολόγησης που χρησιμοποιήθηκαν για την αξιολόγηση της απόδοσης του υβριδικού μοντέλου LSTM συζητούνται. Ο δείκτης που επιλέχθηκε είναι το μέσο τετραγωνικό σφάλμα (RMSE). Το RMSE επιλέγεται επειδή μειώνει τον αντίκτυπο των χονδροειδών σφαλμάτων στο τελικό αποτέλεσμα και διατηρεί την ίδια μονάδα με την προβλεπόμενη κυκλοφορία δεδομένα ροής κυκλοφορίας. Η μαθηματική έκφραση του RMSE δεν παρέχεται ρητά στο απόσπασμα, αλλά είναι αναφέρεται ότι ακολουθείται μια αναφορά [30]. Αυτό υποδηλώνει ότι το έγγραφο αναφέρεται σε μια συγκεκριμένη πηγή για την μαθηματική διατύπωση του RMSE. Η χρήση του RMSE ως μετρικής αξιολόγησης επιτρέπει την ποσοτική αξιολόγηση της ακρίβειας του υβριδικού μοντέλου LSTM στην πρόβλεψη των ροών κυκλοφορίας.

Τα ευρήματα καταδεικνύουν ότι το μέγιστο σχετικό σφάλμα μεταξύ της πραγματικής και της προβλεπόμενης ροές οχημάτων για κάθε οδικό τμήμα είναι 1,03% και για κάθε διασταύρωση είναι 1,18%. Συνεπώς, το υβριδικό LSTM μοντέλο επιδεικνύει υψηλότερο επίπεδο ακρίβειας και ανταποκρίνεται στις απαιτήσεις πραγματικού χρόνου της βραχυπρόθεσμης κυκλοφορίας πρόβλεψης της ροής κυκλοφορίας. Επιπλέον, αποδεικνύεται ότι είναι κατάλληλο για διαφορετικές συνθήκες κυκλοφορίας εντός μιας πραγματικής κυκλοφοριακής δίκτυο.

## **Έγγραφο 6**

**Τίτλος:** Graph Hierarchical Convolutional Recurrent Neural Network (GHCRNN) for Vehicle Condition Prediction

**Εισαγωγή**

Στο παρόν έγγραφο, η κυκλοφοριακή ροή αναφέρεται στην κίνηση των κινούμενων αντικειμένων σε αστικές περιοχές. Η τελευταία συμβολίζεται ως ο αριθμός των οχημάτων που διέρχονται από μια συγκεκριμένη τοποθεσία ή περιοχή σε μια συγκεκριμένη χρονική περίοδο. Η συγκεκριμένη μελέτη επικεντρώνεται στην πρόβλεψη της κυκλοφοριακής ροής και της ταχύτητας των αστικών οχημάτων, η οποία είναι ζωτικής σημασίας για τις «έξυπνες πόλεις» ως προς την λήψη αποφάσεων σε κυκλοφοριακά θέματα.

**Μεθοδολογία**

Η μεθοδολογία που χρησιμοποιείται στην παρούσα εργασία περιλαμβάνει την εφαρμογή μίας παραλλαγής αναδρομικού νευρωνικού δικτύου (RNN), γνωστή και ως Gated Recurrent Unit (GRU), για τη μοντελοποίηση των χρονικών χαρακτηριστικών και την πρόβλεψη της ροής της κυκλοφορίας σε αστικές περιοχές. Το μοντέλο GRU είναι ικανό να συλλαμβάνει μακροπρόθεσμες και βραχυπρόθεσμες εξαρτήσεις στη διάσταση του χρόνου, καθιστώντας το κατάλληλο για τη σύλληψη των χρονικών μοτίβων της κυκλοφοριακής ροής.

Το πρώτο βήμα της μεθοδολογίας περιλαμβάνει την προεπεξεργασία των δεδομένων, η οποία περιλαμβάνει τον καθαρισμό των δεδομένων, την κανονικοποίηση ως προς μία συγκεκριμένη κλίμακα και την επιλογή των χαρακτηριστικών που θα χρησιμοποιηθούν για την πρόβλεψη. Στη συνέχεια, τα δεδομένα χωρίζονται σε σύνολα εκπαίδευσης και ελέγχου. Το μοντέλο GRU εκπαιδεύεται χρησιμοποιώντας το σύνολο εκπαίδευσης των δεδομένων με στόχο εκμάθηση διαφόρων μοτίβων και σχέσεων εντός των δεδομένων. Με άλλα λόγια, το μοντέλο προσπαθεί να ελαχιστοποιήσει τη συνάρτηση σφάλματος, η οποία μετρά τη διαφορά μεταξύ των προβλεπόμενων τιμών ροής κυκλοφορίας και των πραγματικών τιμών.

Αφού εκπαιδευτεί το μοντέλο GRU, το τελευταίο χρησιμοποιείται για την πραγματοποίηση των προβλέψεων στο σύνολο ελέγχου. Η απόδοση του μοντέλου αξιολογείται χρησιμοποιώντας διάφορες μετρικές, όπως το μέσο απόλυτο σφάλμα (MAE) και μέσο τετραγωνικό σφάλμα (MSE), για τη μέτρηση της ακρίβειας των προβλέψεων.

Επιπλέον, οι ερευνητές συγκρίνουν την απόδοση του προτεινόμενου μοντέλου με άλλα υπάρχοντα μοντέλα, συμπεριλαμβανομένου του Historical Average (HA), του στατιστικού μοντέλου ARIMA, του αλγορίθμου Support Vector Regression (SVR) και του LSTM. Η αξιολόγηση αυτών των μοντέλων γίνεται με βάση τις ίδιες μετρικές και το ίδιο σύνολο δεδομένων. Με βάση τα αποτελέσματα, το μοντέλο GRU, ιδιαίτερα το προτεινόμενο μοντέλο GHCRNN (Gated Highway Convolutional Recurrent Neural Network), παρουσίαζουν ανώτερη απόδοση σε σύγκριση με τα προαναφερθέντα μοντέλα, αποδεικνύοντας την αποτελεσματικότητά του στην πρόβλεψη της κυκλοφοριακής ροής σε αστικές περιοχές.

Συνολικά, η μεθοδολογία συνδυάζει τη δύναμη των αναδρομικών νευρωνικών δικτύων, συγκεκριμένα του μοντέλου GRU όταν αυτό τροφοδοτείται με μη γραμμικά δεδομένα.

**Έγγραφο 7**

**Τίτλος:** Short-Term Traffic Flow Prediction Based on XGBoost

**Εισαγωγή**

Στην παρούσα εργασία, ως ροή κυκλοφορίας ορίζεται η κίνηση των οχημάτων εντός μιας συγκεκριμένης λωρίδας κυκλοφορίας σε μια δεδομένη χρονική στιγμή. Η τελευταία μετριέται με βάση παραμέτρους όπως ο αριθμός των διερχόμενων οχημάτων, ο μέσος όρος ταχύτητας και η πληρότητα (το ποσοστό του χρόνου κατά τον οποίο η λωρίδα είναι κατειλημμένη από οχήματα). Αυτοί οι παράμετροι παρέχουν πληροφορίες σχετικά με τον όγκο και τα χαρακτηριστικά της κίνησης των οχημάτων και βοηθούν στην ανάλυση και την πρόβλεψη των μοτίβων κυκλοφορίας.

**Μεθοδολογία**

Η μεθοδολογία της παρούσας εργασίας συνδυάζει την τεχνική της αποσύνθεσης κυματιδίων (wavelet decomposition) με τον αλγόριθμο XGBoost, θέτοντας ως στόχο τη βραχυπρόθεσμη πρόβλεψη της κυκλοφοριακής ροής. Η αποσύνθεση κυματιδίων χρησιμοποιείται για την εξαγωγή επιπρόσθετης πληροφορίας από το χαρακτηριστικό της ροής κυκλοφορίας (το προς πρόβλεψη μέγεθος), ενώ η ανακατασκευή του χαρακτηριστικού συνδυάζει τις πληροφορίες χαμηλής και υψηλής συχνότητας που παρήχθησαν για τη δημιουργία του εν λόγω χαρακτηριστικού που θα χρησιμοποιηθεί στα δεδομένα εκπαίδευσης του αλγορίθμου.

Στη συνέχεια εφαρμόζεται ο αλγόριθμος XGBoost για να την πρόβλεψη της ροής κυκλοφορίας. Η αξιολόγηση της μεθοδολογίας πραγματοποιείται μέσω σύγκρισης με άλλες μεθόδους, δηλαδή των αλγορίθμων SVM και XGBoost χωρίς την μέθοδο της αποσύνθεσης κυματιδίων. Επιπλέον, οι μετρικές αξιολόγησης που χρησιμοποιούνται είναι το μέσο τετραγωνικό σφάλμα (RMSE) και το μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE).

Τα αποτελέσματα δείχνουν ότι η προτεινόμενη μέθοδος επιτυγχάνει χαμηλότερο RMSE και MAPE σε σύγκριση με τα υπόλοιπα μοντέλα. Συνολικά, η παρούσα μεθοδολογία ενισχύει την ακρίβεια και την αποτελεσματικότητα της βραχυπρόθεσμης πρόβλεψης της ροής της κυκλοφορίας.

**Κεφάλαιο 3: Δεδομένα**

Η κυκλοφοριακή ροή είναι ένα μέγεθος που δεν χαρακτηρίζεται από γραμμικότητα. Αυτό σημαίνει ότι ο αριθμός των οχημάτων που διέρχονται από ένα συγκεκριμένο σημείο ενός οδικού δικτύου κατά μια δεδομένη χρονική στιγμή, δεν αυξάνεται ή μειώνεται σε σταθερό ρυθμό, αλλά μπορεί να διακυμαίνεται εξαιτίας διαφόρων παραγόντων, όπως ο χρόνος, οι καιρικές συνθήκες και τα τροχαία ατυχήματα. Επομένως, η χρήση δεδομένων είναι κρίσιμη για την κατανόηση και την πρόβλεψη της κυκλοφοριακής ροής. Για τους σκοπούς της έρευνάς μας, έχουμε εκμεταλλευτεί δύο σύνολα δεδομένων:

* τα δεδομένα τροχιών (trajectory data): αναφέρονται στις πληροφορίες που καταγράφουν τις θέσεις των κινούμενων αντικειμένων σε όλη τη διάρκεια του χρονικού διαστήματος που μας ενδιαφέρει.
* τα δεδομένα καιρού (weather data): αφορούν μετρήσεις και πληροφορίες σχετικά με τις καιρικές συνθήκες σε μια συγκεκριμένη περιοχή. Η καταγραφή και η ανάλυση των δεδομένων καιρού μπορεί να προσφέρει πληροφορίες σχετικά με τις συνθήκες που επηρεάζουν την κίνηση των οχημάτων, καθώς και την ασφάλεια και την απόδοσή τους στους δρόμους.

Και τα δύο προαναφερθέντα είδη δεδομένων συνδυάζονται συχνά με σκοπό την καλύτερη κατανόηση της κυκλοφοριακής ροής και της συμπεριφορά των οχημάτων υπό διάφορες καιρικές συνθήκες.

## **Δεδομένα Τροχιών**

Για να εξασφαλίσουμε την υψηλή ποιότητα της έρευνας, έχουμε προσπαθήσει να αναζητήσουμε ένα σύνολο δεδομένων θέσεων GPS σημείων εντός μίας περιοχής, το οποίο ικανοποιεί ορισμένες αυστηρές προδιαγραφές. Συγκεκριμένα, κατά την επιλογή του ιδανικού συνόλου δεδομένων, πρέπει να συμπεριληφθούν υπόψιν τα ακόλουθα:

* πρώτο, περιέχει δεδομένα θέσεων κινούμενων αντικειμένων GPS που καλύπτουν ένα σχετικά ευρύ γεωγραφικό χώρο (για παράδειγμα τον χώρο μίας ολόκληρης πόλης).
* δεύτερο, στο σύνολο δεδομένων είναι σημαντικό να υπάρχουν καταγραφές θέσεων GPS πολλών διαφορετικών κινούμενων αντικειμένων, ώστε η μελέτη μας να είναι όσο το δυνατόν ακριβέστερη και πληρέστερη.
* τρίτο, τα συνεχόμενα δεδομένα θέσεων GPS για κάθε κινούμενο όχημα πρέπει να καταγράφονται εντός ενός μικρού χρονικού διαστήματος το ένα με το άλλο.
* τέταρτο, ο χρόνος καταγραφής για κάθε σημείο GPS πρέπει να δίνεται ως πληροφορία στο σύνολο δεδομένων.

Ένα σύνολο δεδομένων που καλύπτει όλες αυτές τις προϋποθέσεις είναι γνωστό ως Cab Mobility Traces. Αυτό το σύνολο δεδομένων αποτελεί προϊόν της συνεργασίας μεταξύ του Exploratorium (το μουσείο επιστήμης, τέχνης και ανθρώπινης αντίληψης) και του Scott Snibbe.

Το συγκεκριμένο σύνολο δεδομένων προσφέρει μια λεπτομερή εικόνα των μοτίβων κίνησης των κίτρινων ταξί στην πόλη του San Francisco. Τα δεδομένα έχουν συγκεντρωθεί μέσω ενός καινοτόμου συστήματος παρακολούθησης GPS που ενσωματώθηκε σε κάθε κίτρινο ταξί της πόλης του San Francisco. Αυτό το σύστημα μεταδίδει δεδομένα πραγματικού χρόνου, περιλαμβάνοντας τον αριθμό ταυτοποίησης του ταξί, τις γεωγραφικές του συντεταγμένες (γεωγραφικό πλάτος και γεωγραφικό μήκος) εκείνη τη χρονική στιγμή, τον χρόνο που γίνεται η καταγραφή και την κατάσταση του taxi, δηλαδή εάν εκτελεί δρομολόγιο ή όχι εκείνη τη χρονική στιγμή. Όλη αυτή η πληροφορία συγκεντρώνεται σε ένα κεντρικό διακομιστή.

Στο σύνολο δεδομένων Cab Mobility Traces καταγράφεται η κίνηση των ταξί κατά τον μήνα Μάιο του έτους 2008. Η πορεία κάθε ταξί ενσωματώνεται σε ένα ξεχωριστό αρχείο που φέρει ως όνομα το αναγνωριστικό του ταξί.

Το παρών σύνολο δεδομένων αντιπροσωπεύει μια επιστημονικά σημαντική πηγή, παρέχοντας σε ερευνητές και ακαδημαϊκούς μια μοναδική ευκαιρία να εξερευνήσουν τα μοτίβα κίνησης στην αστική περιοχή του San Francisco και την αλληλεπίδραση των συστημάτων μεταφοράς με το αστικό περιβάλλον. Το σύνολο δεδομένων είναι δημόσια προσβάσιμο μέσω της ιστοσελίδας του έργου Cabspotting [εδώ](http://cabspotting.org), όπου οι χρήστες μπορούν να ανακτήσουν το σύνολο δεδομένων για τις δικές τους αναλύσεις.

## **Δεδομένα Καιρού**

Η συμπερίληψη δεδομένων καιρού είναι απαραίτητη κατά την πρόβλεψη της ροής κυκλοφορίας σε ένα οδικό δίκτυο, καθώς οι καιρικές συνθήκες έχουν έντονη επίδραση στη συμπεριφορά των οχημάτων και, κατά συνέπεια, στην κυκλοφοριακή κατάσταση των οδών. Ο καιρός μπορεί να επηρεάσει την ταχύτητα, την ορατότητα, την πρόσφυση των ελαστικών, τη συμπεριφορά των οδηγών και την κυκλοφοριακή ροή.

Επομένως, για να έχουμε ακόμα καλύτερα αποτελέσματα, έχουμε συμπεριλάβει δεδομένα καιρού της πόλης του San Francisco κατά τον μήνα Μάϊο του έτους 2008. Τα δεδομένα αυτά είναι διαθέσιμα [εδώ](https://www.visualcrossing.com/weather/weather-data-services). Τα δεδομένα καιρού που χρησιμοποιούμε έχουν καταγραφεί σε ωριαία βάση. Τα πιο σημαντικά στοιχεία που συμπεριλαμβάνονται στα δεδομένα αυτά είναι τα ακόλουθα:

* **θερμοκρασία:** η θερμοκρασία μπορεί να επηρεάσει την ταχύτητα των οχημάτων, καθώς και την ορατότητα.
* **υγρασία:** η υγρασία μπορεί να επηρεάσει την πρόσφυση των ελαστικών στον δρόμο και την ασφάλεια της οδήγησης.
* **ταχύτητα του ανέμου:** η ταχύτητα του ανέμου επηρεάζει την συμπεριφορά των οχημάτων και την σταθερότητα της κίνησης.
* **ατμοσφαιρική πίεση:** η ατμοσφαιρική πίεση μπορεί να προσφέρει πληροφορίες σχετικά με τις αλλαγές στις καιρικές συνθήκες και την πιθανή επίδρασή τους στην κυκλοφορία.
* **ορατότητα:** η ορατότητα είναι κρίσιμη για την ασφάλεια της οδήγησης και μπορεί να επηρεάσει την ταχύτητα και τον τρόπο κίνησης.

Αξίζει να σημειωθεί ότι η ανάλυση αυτών των δεδομένων παρέχει πολύτιμες πληροφορίες για τον τρόπο με τον οποίο οι καιρικές συνθήκες επηρεάζουν την κυκλοφοριακή ροή. Η σύνδεση των δεδομένων καιρού με τα δεδομένα κίνησης μπορεί να αποκαλύψει ποιες συνθήκες οδήγησης έχουν τη μεγαλύτερη επίδραση στην κίνηση των οχημάτων, προσφέροντας ένα πληρέστερο και περιεκτικό πλαίσιο για την κατανόηση των παραγόντων που επηρεάζουν την κυκλοφορία.

**Κεφάλαιο 4: Τεχνολογίες**

Το συγκεκριμένο κεφάλαιο έχει δημιουργηθεί για να παραθέσει περισσότερες πληροφορίες σχετικά με τις τεχνολογίες που υιοθετήθηκαν για την διεκπεραίωση της παρούσας έρευνας. Πιο συγκεκριμένα, αναλύονται πληροφορίες για την γλώσσα προγραμματισμού Python που χρησιμοποιήθηκε. Αναφέρονται επίσης πληροφορίες για τα περιβάλλοντα εκτέλεσης του κώδικα, δηλαδή τα Jupyter Notebook και Google Colab. Περιγράφεται, επιπλέον, ο τρόπος λειτουργίας της βάσης δεδομένων PostgreSQL και της γλώσσας PL/pgSQL που χρησιμοποιήθηκαν στην έρευνα μόνο για συγκριτικούς σκοπούς. Τέλος,  αναλύεται ο τρόπος λειτουργίας των λογισμικών Docker και Valhalla, η σύνθεση των οποίων κατέστη εφικτή την διαδικασία αντιστοίχισης των τροχιών των κινούμενων αντικειμένων επάνω στο οδικό δίκτυο.

**Η γλώσσα Python**

Η Python είναι μια γλώσσα προγραμματισμού υψηλού επιπέδου που δημιουργήθηκε από τον Guido van Rossum και πρωτοκυκλοφόρησε το 1991. Έχει κερδίσει ευρεία αναγνώριση χάρη στην απλή και ευανάγνωστη σύνταξή της, καθώς και στην ευελιξία και ισχυρή κοινότητα που την υποστηρίζει. Μερικά από τα βασικά χαρακτηριστικά της Python περιλαμβάνουν:

* κατανοητή σύνταξη: η Python χρησιμοποιεί απλή και ευανάγνωστη σύνταξη, που διευκολύνει την ανάπτυξη και τη συντήρηση του κώδικα.
* διερμηνευμένη φύση: η διερμηνευμένη φύση της Python αναφέρεται στο γεγονός ότι ο κώδικας δεν χρειάζεται να μεταγλωττιστεί προκαταβολικά σε γλώσσα μηχανής πριν από την εκτέλεσή του. Αντί αυτού, ο κώδικας εκτελείται απευθείας από τον διερμηνέα της Python κατά τη διάρκεια της εκτέλεσης του προγράμματος.
* δομημένη οργάνωση: υποστηρίζεται η οργάνωση του κώδικα σε μονάδες, όπως οι συναρτήσεις και οι κλάσεις. Επομένως, η συγκεκριμένη γλώσσα συνδυάζει τα χαρακτηριστικά του διαδικαστικού και του αντικειμενοστραφούς προγραμματισμού.
* δυναμική δήλωση τύπων: δεν απαιτείται να δηλώνεται εξαρχής ο τύπος μιας μεταβλητής, καθώς ο τελευταίος αναγνωρίζεται αυτόματα κατά την εκτέλεση του προγράμματος.
* υποστήριξη από πολυάριθμες βιβλιοθήκες: η Python παρέχει μια πλούσια συλλογή από ενσωματωμένες βιβλιοθήκες για διάφορες εργασίες που υποστηρίζουν από ανάλυση δεδομένων μέχρι γραφικά και σχεδίαση ιστοσελίδων.

Η Python υποστηρίζει επίσης πολλούς τύπους δεδομένων, όπως ακέραιους αριθμούς, πραγματικούς αριθμούς, σύνθετους αριθμούς, συμβολοσειρές, λίστες, πλειάδες και λεξικά. Κάθε τύπος έχει τις δικές του εφαρμογές και χρήσεις, και μπορούν να συνδυαστούν για να δημιουργήσουν πολύπλοκες δομές και λειτουργίες στα προγράμματά.

Συνολικά, η Python είναι μια απλή και ισχυρή γλώσσα προγραμματισμού με πολλές δυνατότητες. Για αυτό τον λόγο, αποτελεί μια εξαιρετική επιλογή τόσο για αρχάριους όσο και για προχωρημένους προγραμματιστές. Δεν είναι λοιπόν παράλογο το γεγονός ότι η συγκεκριμένη γλώσσα χρησιμοποιείται εκτεταμένα σε πολλούς τομείς, όπως είναι η ανάπτυξη λογισμικού, οι επιστημονικοί υπολογισμοί, η ανάλυση δεδομένων, η κατασκευή ιστοσελίδων, η τεχνητή νοημοσύνη και πολλοί άλλοι.

**Python και Επιστήμη Δεδομένων**

Στην παρούσα έρευνα που πραγματοποιούμε, εκμεταλλευόμαστε τη γλώσσα προγραμματισμού Python για την ανάλυση και επίλυση ενός προβλήματος που ανήκει στο πεδίο της μηχανικής μάθησης, δηλαδή την πρόβλεψη της κυκλοφοριακής ροής σε ένα οδικό δίκτυο. Στο πλαίσιο της επιστήμης των δεδομένων, η Python αποτελεί ένα ευρέως χρησιμοποιούμενο εργαλείο, υποστηρίζοντας τις δραστηριότητες ανάλυσης, εξόρυξης και ερμηνείας δεδομένων, προκειμένου να αντληθούν πληροφορίες και γνώση από αυτά. Οι πιο συνήθεις τρόποι χρήσης της Python στο πεδίο της επιστήμης των δεδομένων περιλαμβάνουν:

* ανάλυση και εξόρυξη δεδομένων: η Python παρέχει βιβλιοθήκες, όπως η Pandas και η Numerical Python (NumPy), για την ανάλυση και τον χειρισμό των δεδομένων, ενώ η βιβλιοθήκη Scikit-learn προσφέρει μία ποικιλία από αλγορίθμους μηχανικής μάθησης προκειμένου να εξορύξουμε γνώση από αυτά.
* μηχανική μάθηση: η Python αποτελεί συνήθως την πρώτη επιλογή για την ανάπτυξη μοντέλων μηχανικής μάθησης με τη χρήση βιβλιοθηκών όπως η Scikit-learn, το TensorFlow και το Keras. Αυτές οι βιβλιοθήκες επιτρέπουν τη δημιουργία μοντέλων πρόβλεψης, ταξινόμησης, συσταδοποίησης και πολλών άλλων.
* οπτικοποίηση δεδομένων: η Python παρέχει βιβλιοθήκες όπως η Matplotlib και η Seaborn για τη δημιουργία γραφημάτων που βοηθούν στην οπτικοποίηση των δεδομένων, ενισχύοντας την κατανόηση τους.

Επομένως, η ευελιξία της Python, μαζί με τις πλούσιες βιβλιοθήκες που προσφέρει, καθιστούν αυτήν τη γλώσσα ένα ισχυρό εργαλείο για την ανάλυση και την εξόρυξη δεδομένων. Μάλιστα, αυτό αποτέλεσε και το βασικό κίνητρο επιλογής της συγκεκριμένης γλώσσας στην έρευνά μας.

**Jupyter Notebook και Google Colab**

Όσον αφορά το προγραμματιστικό περιβάλλον που χρησιμοποιήθηκε, η παρούσα έρευνα που έχουμε διεξάγει έχει αναπτυχθεί κυρίως στα περιβάλλοντα του Jupyter Notebook και του Google Colab (ο όρος Colab προέρχεται από τη λέξη Collaboratory). Τόσο το Jupyter Notebook, όσο και το Google Colab αποτελούν δημοφιλή περιβάλλοντα προγραμματισμού, τα οποία χρησιμοποιούνται ευρέως για την ανάπτυξη και εκτέλεση κώδικα Python, καθώς και για την οπτικοποίηση και ανάλυση δεδομένων. Επομένως, έχουν εξαιρετική απήχηση στον χώρο της επιστήμης των δεδομένων και της μηχανικής μάθησης.

**Jupyter Notebook**

Το Jupyter Notebook αντιπροσωπεύει ένα περιβάλλον προγραμματισμού που επιτρέπει τη δημιουργία κώδικα Python και τον κοινό διαμοιρασμό εγγράφων που περιλαμβάνουν εκτελέσιμο κώδικα, κείμενο, εικόνες, γραφήματα και άλλα, με άλλους χρήστες.

Αναλυτικότερα, το Jupyter Notebook, όταν εγκατασταθεί τοπικά, λειτουργεί σε έναν τοπικό διακομιστή. Κατά την εκτέλεσή του, το γραφικό περιβάλλον της εφαρμογής εμφανίζεται μέσω ενός προγράμματος περιήγησης σε μια προκαθορισμένη διεύθυνση IP, συνήθως στην http://localhost:8888. Επιπρόσθετα, το κλείσιμο του προγράμματος περιήγησης δεν συνεπάγεται και την αποσύνδεση από τον διακομιστή που φιλοξενεί το Jupyter Notebook.

Όσον αφορά την σχεδίαση του Jupyter Notebook, αυτό είναι οργανωμένο σε σημειωματάρια, όπως βέβαια δηλώνει και το όνομα του προγράμματος. Κάθε σημειωματάριο είναι οργανωμένο σε εκτελέσιμα κελιά. Ένα κελί μπορεί να περιλαμβάνει διάφορα είδη πληροφοριών, όπως κώδικα Python ή άλλους τύπους πληροφορίας, για παράδειγμα κείμενο. Οι χρήστες έχουν τη δυνατότητα να εκτελούν κάθε ένα αυτά τα κελιά, ξεχωριστά το ένα με το άλλο, και να παρατηρούν άμεσα τα αποτελέσματα. Το Jupyter Notebook αποτέλεσε το βασικό εργαλείο συγγραφής κώδικα κατά την διάρκεια της έρευνας, λόγω της ευχρηστότητάς του.

**Google Colab**

Το Google Colab αντιπροσωπεύει μια υπηρεσία που προσφέρεται από την Google και λειτουργεί παρόμοια με το Jupyter Notebook. Η μόνη διαφορά είναι ότι η εφαρμογή αυτή παρέχει ένα περιβάλλον προγραμματισμού που εκτελείται στο υπολογιστικό νέφος της Google προσφέροντας στον τελικό χρήστη δωρεάν ή προς πληρωμή πόρους, όπως επεξεργαστική ισχύ και μνήμη για την εκτέλεση κώδικα. Το Colab αποδεικνύεται ιδιαίτερα χρήσιμο για την εκπαίδευση μοντέλων μηχανικής μάθησης, αφού παρέχει πρόσβαση σε βιβλιοθήκες όπως το TensorFlow και το Keras. Τέλος, το Google Colab δίνει τη δυνατότητα της εύκολης κοινής χρήσης των σημειωματάριων με άλλους χρήστες.

Στην εν λόγω έρευνα, εκμεταλλευτήκαμε αρκετά τις κάρτες γραφικών που προσφέρει η υπηρεσία, ώστε να εκτελέσουμε τον τελικό κώδικά μας και να εκπαιδεύσουμε τα μοντέλα μηχανικής μάθησης που αναπτύχθηκαν σε αρκετά μικρότερο χρόνο.

**PostgreSQL και PLpgSQL**

Το σύστημα διαχείρισης βάσεων δεδομένων της PostgreSQL είναι μία τεχνολογία, η οποία χρησιμοποιήθηκε στην έρευνά μας μόνο για συγκριτικούς σκοπούς. Περισσότερες πληροφορίες για την χρήση της δίνονται στο επόμενο κεφάλαιο. Λόγω του γεγονότος ότι η χρήση της είναι περιορισμένη στην παρούσα εργασία, παρουσιάζουμε τα πιο σημαντικά στοιχεία που αφορούν αυτήν την τεχνολογία.

**Γενικά χαρακτηριστικά**

Η PostgreSQL, γνωστή και ως Postgres, αποτελεί ένα προηγμένο σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων (ΣΔΒΔ). Αυτό το ΣΔΒΔ είναι εύκολα επεκτάσιμο μέσω διάφορων προσθέτων, πολλά από τα οποία είναι ανοιχτού κώδικα και διαθέσιμα δωρεάν μέσω του διαδικτύου. Επομένως, πρόκειται για μία τεχνολογία με πολλές δυνατότητες.

Σε ένα σχεσιακό ΣΔΒΔ, τα δεδομένα αποθηκεύονται σε πίνακες με στήλες και γραμμές. Κάθε στήλη αντιπροσωπεύει έναν συγκεκριμένο τύπο δεδομένων, ενώ κάθε γραμμή αντιπροσωπεύει μια εγγραφή με συγκεκριμένες τιμές για κάθε στήλη. Το σχεσιακό μοντέλο επιτρέπει τη δημιουργία συσχετίσεων μεταξύ των πινάκων, δημιουργώντας ένα περίπλοκο δίκτυο συνδέσεων για την ανάκτηση σχετικών δεδομένων. Ένα σχεσιακό ΣΔΒΔ προσφέρει πολλά πλεονεκτήματα, όπως:

* δομημένη οργάνωση: τα δεδομένα οργανώνονται σε πίνακες ή σχέσεις, πράγμα που καθιστά εύκολη την οργάνωση και την ανάκτησή τους.
* ευέλικτη ανάκτηση: οι χρήστες μπορούν να κάνουν πολύπλοκα ερωτήματα με βάση διάφορα κριτήρια και συνθήκες.
* κοινόχρηστη πρόσβαση: πολλοί χρήστες μπορούν να έχουν πρόσβαση στα ίδια δεδομένα ή στην ίδια βάση δεδομένων ταυτόχρονα.
* αποκλεισμός ανωμαλιών: το ΣΔΒΔ παρέχει μηχανισμούς αποκλεισμού ανωμαλιών και ανάκαμψης μετά από κάποιο σφάλμα.

Ένα σημαντικό χαρακτηριστικό που υποστηρίζεται από την PostgreSQL είναι τα ευρετήρια (indexes). Τα ευρετήρια βελτιστοποιούν την απόδοση και την αναζήτηση δεδομένων στη βάση επιτρέποντας γρήγορη πρόσβαση σε συγκεκριμένες εγγραφές μίας σχέσης, μειώνοντας έτσι τον χρόνο αναζήτησης και ανάκτησης των δεδομένων. Η PostgreSQL υποστηρίζει διάφορους τύπους ευρετηρίων που εξυπηρετούν διάφορες ανάγκες. Τα πιο σημαντικά και ευρέως χρησιμοποιούμενα ευρετήρια στην postgres είναι τα ευρετήρια B+ δέντρων και τα ευρετήρια κατακερματισμού.

Τέλος, η PostgreSQL υποστηρίζει διεπαφές με πολλές γλώσσες προγραμματισμού, όπως είναι οι Java, Python, C, C++, PHP, C#, επιτρέποντας στις εφαρμογές να αλληλοεπιδρούν άμεσα με τη βάση δεδομένων.

**Η γλώσσα PL/pgSQL**

Η γλώσσα προγραμματισμού PL/pgSQL (από το Procedural Language/PostgreSQL) αποτελεί μια πολύτιμη επέκταση του διαχειριστικού συστήματος βάσεων δεδομένων της PostgreSQL. Πιο αναλυτικά, τα κυριότερα χαρακτηριστικά της PL/pgSQL συνοψίζονται παρακάτω:

* δημιουργία συναρτήσεων και διαδικασιών: με την PL/pgSQL, μπορούν να δημιουργηθούν συναρτήσεις που επιστρέφουν τιμές και διαδικασίες που εκτελούν ενέργειες χωρίς επιστροφή τιμών. Αυτό επιτρέπει την οργάνωση εντός της βάσης δεδομένων.
* ενσωμάτωση με την PostgreSQL: ένα από τα πλεονεκτήματα της PL/pgSQL είναι ότι είναι πλήρως ενσωματωμένη με το σύστημα διαχείρισης βάσεων δεδομένων της PostgreSQL. Αυτό σημαίνει ότι οι διαδικασίες και οι συναρτήσεις που γράφονται σε γλώσσα PL/pgSQL είναι άμεσα εκτελέσιμες από τον ίδιο τον εξυπηρετητή της βάσης. Με τη δυνατότητα εκτέλεσης πολύπλοκων λειτουργιών απευθείας στη βάση δεδομένων, εξοικονομούνται πόροι, βελτιώνοντας την απόδοση. Η PL/pgSQL αναδεικνύεται ως ένα απαραίτητο εργαλείο για τη δημιουργία πολυπλοκότερων και αποτελεσματικότερων διαχειριστικών λειτουργιών εντός του περιβάλλοντος της PostgreSQL.
* εύκολη Ανάπτυξη Κώδικα: η σύνταξη της PL/pgSQL είναι παρόμοια με αυτή της κλασικής SQL, επιτρέποντας στους προγραμματιστές και τους διαχειριστές βάσεων δεδομένων να εξοικειωθούν γρήγορα με τη νέα γλώσσα.
* δήλωση μεταβλητών και σταθερών: ένα από τα σημαντικότερα χαρακτηριστικά της PL/pgSQL είναι η δυνατότητα δήλωσης μεταβλητών για την αποθήκευση δεδομένων κατά τη διάρκεια της εκτέλεσης προγραμμάτων, καθώς και η δυνατότητα αποθήκευσης σταθερών τιμών, οι τιμές των οποίων παραμένουν σταθερές κατά τη διάρκεια της εκτέλεσης του κώδικα.
* έλεγχος ροής: η PL/pgSQL παρέχει δομές ελέγχου ροής, επιλογής και επανάληψης, οι οποίες επιτρέπουν τη διαχείριση της ροής εκτέλεσης του προγράμματος βάσει συνθηκών και κριτηρίων.
* παραμετροποίηση: η δυνατότητα χρήσης παραμέτρων και μεταβλητών στην PL/pgSQL παρέχει ευελιξία κατά την εκτέλεση των διαδικασιών, επιτρέποντας την προσαρμογή της συμπεριφοράς του προγράμματος βάσει δυναμικών παραμέτρων.
* αναγνώριση σφαλμάτων: σε περίπτωση εμφάνισης σφάλματος, η PL/pgSQL παρέχει σαφή μηνύματα σφάλματος προς τον προγραμματιστή, διευκολύνοντας τον εντοπισμό και τη διόρθωσή τους.

Συνολικά, η γλώσσα PL/pgSQL αντιπροσωπεύει μια ισχυρή εργαλειοθήκη που συμβάλλει στην επέκταση των δυνατοτήτων της PostgreSQL, καθιστώντας τη ένα περιβάλλον κατάλληλο όχι μόνο για την αποθήκευση δεδομένων, αλλά και για την εκτέλεση πολυπλοκότερων εργασιών και λογικών σεναρίων. Ουσιαστικά, η PL/pgSQL επιτρέπει στους χρήστες να εκμεταλλεύονται την ισχύ μίας γλώσσας προγραμματισμού, για να διαμορφώσουν τη βάση δεδομένων όπως αυτοί επιθυμούν.

**Το λογισμικό Docker**

Το λογισμικό Docker αποτελεί μια εξαιρετικά σημαντική τεχνολογία, η οποία συνδέεται στενά με τον τρόπο που αναπτύσσονται, μεταφέρονται και εκτελούνται εφαρμογές και λογισμικά. Το κύριο πλεονέκτημα του Docker είναι ότι επιτρέπει τη δημιουργία ενός «κουτιού εκτέλεσης» (containers), το οποίο μπορεί να εκτελεστεί σε οποιοδήποτε σύστημα και λειτουργικό, εξαλείφοντας έτσι τις ανησυχίες περί συμβατότητας. Αναλυτικότερα, κάθε κουτί εκτέλεσης περιλαμβάνει:

* **εκτελέσιμο αρχείο:** το εκτελέσιμο αρχείο της εφαρμογής.
* **βιβλιοθήκες:** οι βιβλιοθήκες και οι εξαρτήσεις που απαιτούνται για τη λειτουργία της εφαρμογής. Αυτές χρησιμοποιούνται κατά τη διάρκεια της εκτέλεσης.
* **ρυθμίσεις περιβάλλοντος:** είναι τα πρότυπα που καθορίζουν πώς η εφαρμογή πρέπει να εκτελείται, όπως είναι οι μεταβλητές περιβάλλοντος και οι παράμετροι εκτέλεσης.
* **εκτελέσιμες εντολές:** οι εκτελέσιμες εντολές αναφέρονται στις ενέργειες που πρέπει να ληφθούν για να ξεκινήσει η εφαρμογή εντός του container.
* **δεδομένα εφαρμογής:** πρόκειται για οποιαδήποτε δεδομένα ή αρχεία που χρειάζεται η εφαρμογή για να λειτουργήσει σωστά.

Η αξία του Docker έγκειται στην ευκολία ανάπτυξης, στην ευελιξία και στην αξιοπιστία εκτέλεσης των εφαρμογών. Όσον αφορά την επιρροή του Docker στον παγκόσμιο χώρο, η τεχνολογία αυτή έχει υιοθετηθεί από πολυάριθμους χρήστες και επιχειρήσεις, διαμορφώνοντας έτσι νέα πρότυπα για την αποτελεσματική και αξιόπιστη διαχείριση των εφαρμογών.

**Το λογισμικό Valhalla**

Το λογισμικό Valhalla αποτελεί ένα πρόγραμμα ανοιχτού κώδικα που αναπτύχθηκε αρχικά από την εταιρεία Mapzen. Η συγκεκριμένη εταιρία ιδρύθηκε το 2013 και ειδικευόταν στις υπηρεσίες χαρτογράφησης γεωχωρικών δεδομένων. Η συγκεκριμένη εφαρμογή έχει ως κύριο στόχο να εκτελεί υπηρεσίες χαρτογράφησης GPS δεδομένων, προσφέροντας ακριβείς δρομολογήσεις για αυτοκίνητα, ποδήλατα και άλλα μέσα μεταφοράς, καθώς επίσης και τη δυνατότητα αντιστοίχισης τροχιών GPS σε ψηφιακούς χάρτες.

Η διαδικασία ανάπτυξης του Valhalla εκκινείται το 2014 και ολοκληρώνεται το 2019, όταν η εταιρεία Mapzen αποσύρεται από τον επιχειρηματικό τομέα. Μετά το κλείσιμο της Mapzen, το Valhalla και τα σχετικά δεδομένα που το απαρτίζουν μεταφέρονται στον Οργανισμό Ελεύθερου Λογισμικού (Software Freedom Conservancy) για τη διαχείριση, ανάπτυξη και συντήρηση του λογισμικού Valhalla.

Στον πυρήνα του Valhalla είναι συγκεντρωμένοι αλγόριθμοι υπεύθυνοι για την επεξεργασία των γεωχωρικών δεδομένων, όπως η δρομολόγηση και η αντιστοίχιση τροχιών σε χάρτες, εξυπηρετώντας έτσι προγραμματιστές και χρήστες που ψάχνουν μία τέτοια υπηρεσία.  Επιπρόσθετα, οι αλγόριθμοι είναι σχεδιασμένοι με τέτοιο τρόπο,  ώστε να προσδίδουν ακρίβεια και αξιοπιστία στα αποτελέσματα που παράγουν.

Το λογισμικό Valhalla είναι συμβατό με διάφορα λειτουργικά συστήματα, όπως το Linux, το macOS και τα Windows. Επιπλέον, το Valhalla είναι προσαρμόσιμο για εκτέλεση τόσο σε τοπικούς υπολογιστές, όσο και σε απομακρυσμένους διακομιστές (servers). Τέλος, λόγω του γεγονότος ότι το εν λόγω λογισμικό είναι ανοιχτού κώδικα, σημαίνει ότι διανέμεται και δωρεάν στους ενδιαφερόμενους χρήστες.

**Valhalla Meili**

Το Valhalla Meili αποτελεί ένα υποσύνολο του ευρύτερου project Valhalla και αναλαμβάνει να εκτελέσει μία συγκεκριμένη λειτουργία χαρτογράφησης. Η κύρια διαφορά μεταξύ του Valhalla Meili και του Valhalla είναι ότι το πρώτο είναι υπεύθυνο για την αντιστοίχιση τροχιών GPS σε ένα ψηφιακό χάρτη, ενώ το δεύτερο παρέχει στο χρήστη ένα ευρύτερο φάσμα λειτουργιών χαρτογράφησης, όπως η αναζήτηση της συντομότερης διαδρομής μεταξύ δύο σημείων.

Συγκεκριμένα, το Valhalla Meili αναλαμβάνει την λήψη της πορείας μιας παρατηρηθείσας τροχιάς αποτελούμενη από διαδοχικά GPS σημεία (ζεύγη από γεωγραφικά μήκη και γεωγραφικά πλάτη) και δίνει ως αποτέλεσμα μια πιθανή αντιστοιχισμένη διαδρομή στον ψηφιακό χάρτη.

Στην μελέτη μας, χρησιμοποιούμε ως ψηφιακό χάρτη τα δεδομένα από το Open Street Map (OSM) της περιοχής του San Francisco, California. Το OSM είναι μια δωρεάν, δημόσια βάση γεωγραφικών δεδομένων που ενημερώνεται και συντηρείται συνεχώς από μια κοινότητα εθελοντών. Μέσα σε αυτή την βάση αποθηκεύονται πληροφορίες σχετικά με το οδικό δίκτυο μίας περιοχής (για παράδειγμα δρόμοι, διασταυρώσεις, εθνικές οδοί, πεζοδρόμια κ.α.) και την τοπολογία μίας περιοχής (παραδείγματος χάριν βουνά, λίμνες, ποτάμια, πεδιάδες, θάλασσες κ.α.). Το OpenStreetMap χρησιμοποιείται ευρέως για την εξαγωγή ηλεκτρονικών χαρτών σε μορφή αρχείων και την οπτικοποίηση γεωχωρικών δεδομένων, καθώς διατίθεται ελεύθερα υπό την άδεια της ανοιχτής βάσης δεδομένων.

Στην διερεύνηση που έχουμε διεξάγει, χρησιμοποιούμε την υπηρεσία Valhalla Meili σε συνδυασμό με το λογισμικό Docker. Ουσιαστικά, έχουμε εγκαταστήσει το Valhalla Meili σε ένα κουτί (container) στο Docker, μέσα στο οποίο το πρόγραμμα εκτελείται ανεξάρτητα από το λειτουργικό σύστημα και τους περιορισμούς του μηχανήματος που διαθέτουμε. Μετά την εγκατάσταση του Valhalla Meili, δημιουργείται ένας εξυπηρετητής που αναλαμβάνει να δέχεται αιτήματα αντιστοίχισης GPS σημείων και να επιστρέφει την αντιστοιχισμένη πληροφορία στον ψηφιακό χάρτη OSM που χρησιμοποιούμε.

Η απάντηση που λαμβάνουμε από τον εξυπηρετητή του Valhalla μετά από κάθε αίτημα αντιστοίχισης που στέλνουμε σε αυτόν, είναι εμπλουτισμένη με χρήσιμες πληροφορίες. Οι πληροφορίες αυτές καθίστανται προσβάσιμες εξαιτίας της λειτουργίας trace attributes που διαθέτει το Valhalla Meili. Αυτό έχει ως αποτέλεσμα κάθε τροχιά να συνοδεύεται από επιπρόσθετα δεδομένα, όπως είναι η ταχύτητα του κινούμενου αντικειμένου, η κατάστασή του (π.χ. είναι σε κίνηση ή σε στάση), το είδος του κινούμενου οχήματος (π.χ. αυτοκίνητο, ποδήλατο, πεζός) κ.α.

Εξαιρετικά χρήσιμο χαρακτηριστικό της λειτουργίας αυτής αποτελεί επίσης και τη δυνατότητα να γνωρίζουμε το μοναδικό αναγνωριστικό της ακμής (Edge ID), στην οποία αντιστοιχίζεται το GPS σημείο που εξετάζουμε. Αυτά τα μοναδικά Edge IDs που παρέχει το Valhalla Meili δεν είναι τυχαία, αλλά προέρχονται απευθείας από τη βάση δεδομένων του Open Street Map. Αυτό σημαίνει ότι μπορούμε εύκολα να αναζητήσουμε την αντίστοιχη ακμή, απλά κάνοντας αναζήτηση στη βάση δεδομένων του OSM. Μία ακόμα λεπτομέρεια που αξίζει να προσέξουμε είναι ότι κάθε Edge ID που υπάρχει στη βάση δεδομένων του OSM είναι μοναδικό.

Αντιστοιχίζοντας σημεία GPS με τα Edge IDs, ο αλγόριθμος μπορεί να ανακατασκευάσει ακριβώς τη διαδρομή που ακολούθησε το κινούμενο αντικείμενο, όπως ένα όχημα ή ένας χρήστης. Η διαδικασία αντιστοίχισης δεδομένων GPS που απαρτίζουν μία τροχιά στο οδικό δίκτυο περιλαμβάνει την εύρεση της ακολουθίας των Edge IDs με σωστή σειρά. Αυτή η ακολουθία των Edge IDs αντιστοιχίζεται όσο το δυνατόν καλύτερα στο οδικό δίκτυο δίνοντας ως αποτέλεσμα μια ακριβή ή αρκετά καλή αναπαράσταση της διαδρομής που διένυσε το κινούμενο αντικείμενο πάνω στον OSM χάρτη.

**Κεφάλαιο 5: Υλοποίηση της Εφαρμογής**

Στο κεφάλαιο αυτό παρέχεται όλη η απαραίτητη γνώση σχετικά με τον κώδικα που αναπτύχθηκε. Ταυτόχρονα, αναλύονται οι τεχνικές και οι βελτιστοποιήσεις που λήφθηκαν υπόψιν, ώστε να παραχθούν ακριβέστερα και ποιοτικότερα αποτελέσματα.

Ο κώδικας έχει χωριστεί σε τρία αρχεία, τα οποία φέρουν τους τίτλους Notebook1, Notebook2 και Notebook3. Στο πρώτο αρχείο γίνεται όλη η προεπεξεργασία των δεδομένων: η αντιστοίχιση των σημείων GPS στο οδικό δίκτυο του San Francisco και η αναγωγή του προβλήματος σε χρονοσειρές. Στο αρχείο Notebook2 δείχνουμε γιατί είναι σημαντική η χρήση της μεθόδου των Αυστηρών Ερωτημάτων Μονοπατιού (ΑΕΜ). Τέλος, στο τρίτο αρχείο, εισάγουμε τα δεδομένα καιρού, διεξάγουμε οπτικοποιήσεις των δεδομένων και συγκρίνουμε τέσσερα μοντέλα μηχανικής μάθησης, όπου κάθε ένα από αυτά αποσκοπεί στην επίλυση του βασικού προβλήματος, δηλαδή της πρόβλεψης της κυκλοφοριακής ροής σε ολόκληρα μονοπάτια εντός του οδικού δικτύου του San Francisco.

## **Προεπεξεργασία των δεδομένων**

Η αρχική φάση του κώδικα είναι αφιερωμένη στην συλλογή και την προετοιμασία του αρχικού συνόλου δεδομένων. Όσον αφορά την συλλογή των δεδομένων, λόγω του ότι αυτά βρίσκονται μοιρασμένα σε πολλά διαφορετικά αρχεία, προσπαθούμε να τα ενοποιήσουμε όλα μαζί σε ένα αρχείο. Το σύνολο δεδομένων που προκύπτει, περιλαμβάνει όλες τις καταγραφές κίνησης όλων των ταξί που κινούνται στην πόλη του San Francisco κατά τον μήνα Μάιο του έτους 2008. Το σύνολο δεδομένων είναι σχολαστικά δομημένο, ενσωματώνοντας χαρακτηριστικά όπως τα εξής:

* **Taxi ID:** το μοναδικό αναγνωριστικό του κάθε Ταξί.
* **Latitude:** το γεωγραφικό πλάτος της θέσης του ταξί όταν έγινε η καταγραφή.
* **Longitude:** το γεωγραφικό μήκος της θέσης του ταξί όταν έγινε η καταγραφή.
* **Occupied:** δηλώνει εάν το ταξί μετέφερε επιβάτη ή όχι όταν έγινε η καταγραφή της θέσης του (είναι η μόνη πληροφορία που δεν θα χρειαστεί στην έρευνά μας).
* **Date Time:** Η ημερομηνία και η ώρα που έγινε η καταγραφή. Είναι της μορφής χρονιά-μήνας-ημέρα ώρα:λεπτό:δευτερολεπτο.

Το σύνολο δεδομένων περιλαμβάνει περίπου έντεκα εκατομμύρια εγγραφές. Για να διευκολύνουμε την διαδικασία της έρευνάς μας, έχουμε περιορίσει τον αριθμό των εγγραφών που χρησιμοποιούμε, απομονώνοντας τα δεδομένα που καταγράφηκαν κατά τη διάρκεια μιας εβδομάδας - συγκεκριμένα, από τις 18 έως τις και τις 25 Μαΐου του έτους 2008. Μάλιστα, θεωρήσαμε φρόνιμο το γεγονός ότι η κυκλοφοριακή ροή είναι ένα μέγεθος που για να προβλεφθεί η συμπεριφορά του δεν χρειάζεται να ανατρέξουμε πολύ πίσω στον χρόνο.

Το συγκεκριμένο σύνολο δεδομένων είναι ιδανικό, καθώς περιέχει καταγραφές σημείων GPS ανά μικρά χρονικά διαστήματα. Ωστόσο, για κάθε διαφορετικό ταξί (δηλαδή για κάθε διαφορετικό Taxi ID) υπάρχει μία συνεχόμενη ροή από καταγραφές θέσεων GPS για το χρονικό διάστημα ολόκληρου του μήνα Μάιου. Επομένως, για κάθε ταξί διατίθεται μία μονοκόμματη τροχιά. Ένα πρόβλημα που δημιουργείται είναι η διαδικασία διαχωρισμού της τροχιάς σε υποτροχιές, δηλαδή σε μικρότερες τροχιές της ίδιας μεγαλύτερης τροχιάς. Επιπλέον, θέλουμε σε κάθε υποτροχιά να περιλαμβάνονται σημεία GPS που απέχουν ανά δύο ένα συγκεκριμένο χρονικό διάστημα. Στην έρευνα έχουμε ορίσει ως μέγιστο χρονικό διάστημα τα ενενήντα δευτερόλεπτα. Με άλλα λόγια, κάθε τροχιά ενός ταξί θα διαμοιραστεί σε υποτροχιές που θα περιέχουν η κάθε μία διαδοχικά σημεία που απέχουν χρονικά ανά δύο έως και ενενήντα δευτερόλεπτα. Αυτή η διαδικασία ονομάζεται “Διαχωρισμός Τροχιάς”. Η προσέγγιση αυτή έρχεται να εισάγει μία καινούρια πληροφορία στο σύνολο δεδομένων, την στήλη “**Traj ID**” που δηλώνει το αναγνωριστικό της υποτροχιάς. Πλέον, κάθε μοναδικό ζεύγος (Taxi ID, Traj ID) ταυτοποιεί μοναδικά μία τροχιά. Μετά το πέρας αυτής της ενέργειας, το πλήθος των τροχιών που υπάρχει στο σύνολο δεδομένων αυξάνεται.

## **Αντιστοίχιση τροχιών στο οδικό δίκτυο**

Ένα επιπλέον πρόβλημα που προκύπτει με τα δεδομένα, είναι ότι οι τροχιές δεν έχουν αντιστοιχηθεί σε κάποιον ψηφιακό χάρτη. Αυτό οδηγεί στο ακόλουθο πρόβλημα: τα δεδομένα GPS ίσως να μην είναι ακριβή, αφού μπορεί να υπήρξε θόρυβος κατά την συλλογή τους. Επομένως, μία διαδικασία αντιστοίχισης των σημείων GPS στο οδικό δίκτυο που χρησιμοποιούμε καθίσταται απαραίτητη.

Ο κώδικας εκκινεί τη διαδικασία αντιστοίχισης τροχιών στον χάρτη της περιοχής του San Francisco, εκμεταλλευόμενο το Valhalla Meili API.  Αυτό καθιστά δυνατή την ευθυγράμμιση των πορειών GPS του κάθε ταξί με το υποκείμενο οδικό δίκτυο. Κάθε τροχιά, που αναπαρίσταται από ζεύγη γεωγραφικού πλάτους και μήκους μέσα στο σύνολο δεδομένων, υποβάλλεται ως αίτημα στον εξυπηρετητή του Valhalla Meili, λαμβάνοντας ως έξοδο ένα νέο σύνολο δεδομένων που φέρει το όνομα **visited\_segments** και περιέχει όλη την απαραίτητη πληροφορία, δηλαδή τις αντιστοιχιζόμενες τροχιές επάνω στο οδικό δίκτυο. Συγκεκριμένα, οι πληροφορίες εξάγονται από το trace attributes του Valhalla Meili και περιλαμβάνουν τις ακόλουθες στήλες:

* **αναγνωριστικό ταξί (Taxi ID):** το μοναδικό αναγνωριστικό που αντιστοιχεί σε κάθε ταξί, επιτρέποντας τη διάκριση των τροχιών.

* **αναγνωριστικό τροχιάς (Traj ID):** το αναγνωριστικό της υποτροχιάς της κύριας τροχιάς που διένυσε το ταξί με αναγνωριστικό Taxi ID.
* **αναγνωριστικό διαδρομής OSM (OSM Way ID):** δηλώνει το αναγνωριστικό της αντιστοιχιζόμενης στο οδικό δίκτυο ακμής.
* **ώρα έναρξης (Start Time):** χρονοσφραγίδα που υποδηλώνει τη στιγμή που η τροιχιά εισέρχεται στην ακμή με αναγνωριστικό OSM Way ID.

* **ώρα λήξης (End Time):** χρονοσφραγίδα που υποδηλώνει τη στιγμή που η τροχιά εξέρχεται από την ακμή με αναγνωριστικό OSM Way ID.

Μία σημαντική διευκρίνιση που πρέπει να γίνει εδώ είναι ότι στο καινούριο σύνολο δεδομένων που δημιουργήθηκε έχουμε για κάθε ξεχωριστή τροχιά (δηλαδή ένα μοναδικό ζεύγος Taxi ID και Traj ID) τις διαδοχικές ακμές που αυτή διένυσε, όπως και την χρονική στιγμή που αυτή εισήλθε και εξήλθε από κάθε ακμή.

## **Αναγωγή του προβλήματος σε χρονοσειρές**

Χρησιμοποιώντας τα δεδομένα που μας έδωσε ως έξοδο ο αλγόριθμος Valhalla Meili και την μεθοδολογία των ΑΕΜ, καταλήγουμε σε ένα τελικό σύνολο δεδομένων που αποτελείται από χρονοσειρές. Μάλιστα, έχουμε υλοποιήσει την ιδέα των ΑΕΜ σε δύο προγραμματιστικά περιβάλλοντα, επιλέγοντας στο τέλος την ταχύτερη λύση. Στο παρόν υποκεφάλαιο εξετάζουμε όλα τα βήματα κατασκευής του τελικού συνόλου δεδομένων.

### **Υλοποίηση των Αυστηρών Ερωτημάτων Μονοπατιού**

Ένα από τα πιο σημαντικά κομμάτια αυτής της πτυχιακής εργασίας είναι και η συγγραφή της συνάρτησης που υλοποιεί τα ΑΕΜ. Υπενθυμίζεται ότι ένα τέτοιο ερώτημα βρίσκει όλες τις τροχιές που ακολουθούν επακριβώς ένα συγκεκριμένο μονοπάτι οποιουδήποτε μήκους και εντός ενός συγκεκριμένου χρονικού διαστήματος. Η έναρξη αυτού του χρονικού διαστήματος υποδηλώνει τον χρόνο που η τροχιά εισέρχεται στο συγκεκριμένο μονοπάτι, ενώ η λήξη του χρονικού διαστήματος υποδηλώνει τον ανεκτό χρόνο που η εκάστοτε τροχιά θα πρέπει να έχει εξέλθει από αυτό.

Πρέπει να σημειωθεί ότι όλος ο κώδικας έχει γραφεί σε python σε μορφή συνάρτησης, ενώ η συνολική διάρκεια εκτέλεσης μίας συνάρτησης διαρκεί περίπου 300 – 500 ms κατά προσέγγιση. Οι συγκεκριμένοι χρόνοι είναι σχετικοί και προκύπτουν από ένα σύνολο δεδομένων με μέγεθος 3.300.000 εγγραφών. Με άλλα λόγια, η συγκεκριμένη χρονική απόδοση είναι πάρα πολύ ικανοποιητική! Επιπλέον, η συνάρτηση προσπελάβνει κάθε φορά τον πίνακα visited\_segments που αναφέραμε προηγουμένως. Οι παράμετροι που λαμβάνει σαν είσοδο η συνάρτηση SPQ είναι οι ακόλουθες:

* **path:** η διαδρομή που θα πρέπει να ακολουθήσουν οι τροχιές ακριβώς (ακμή προς ακμή), χωρίς να παρεκκλίνουν από αυτή. Αυτό το μονοπάτι μπορεί να έχει οποιοδήποτε μήκος ακμών μεγαλύτερο ή ίσο των δύο.
* **time\_enter:** ο χρόνος, κατά τον οποίο η τροχιά θα πρέπει να έχει εισέλθει στην πρώτη ακμή της διαδρομής που δίνεται ως είσοδος.
* **time\_leave:** ο χρόνος κατά τον οποίο η τροχιά πρέπει να έχει εγκαταλείψει την τελευταία ακμή της διαδρομής που δίνεται ως είσοδος.

Ακολουθεί τώρα ο ψευδοκώδικας που περιγράφει την συνάρτηση που υλοποιεί τα ΑΕΜ. Το όνομα της συνάρτησης είναι SPQ:

ΑΛΓΟΡΙΘΜΟΣ SPQ

Είσοδος: ένα μονοπάτι, ο χρόνος time\_enter, ο χρόνος time\_leave.

1. Υπολόγισε το μήκος του μονοπατιού και αποθήκευσέ το σε μία μεταβλητή path\_length
2. Φτιάξε μία κενή λίστα trajectories
3. Βρες όλες τις εγγραφές του πίνακα visited\_segments που έχουν Start Time >= time\_enter και End Time <= time\_leave, αποθήκευσέ τες σε μία μεταβλητή με όνομα examined\_data.
4. Βρες όλα τα αναγνωριστικά των γραμμών που περιέχουν σαν OSM Way ID την πρώτη ακμή στο μονοπάτι π και αποθήκευσέ τα σε μία λίστα needed\_indexes.

**4.** Για κάθε στοιχείο index στη λίστα needed\_indexes επανέλαβε:

**4.1** Βρες το Taxi ID του στοιχείου index και αποθήκευσέ το σε μία μεταβλητή taxi\_id

**4.2** Βρες το Traj ID του στοιχείου index και αποθήκευσέ το σε μία μεταβλητή traj\_id

**4.3** Όρισε την τιμής μίας νέας μεταβλητής inter = 1

**4.4** Από I = 1 έως path\_length επανέλαβε:

**4.4.1** Έλεγξε εάν η γραμμή με αναγνωριστικό index+i περιέχει σαν Taxi ID == taxi\_id ΚΑΙ Traj ID == traj\_id ΚΑΙ OSM Way ID την επόμενη σε σειρά ακμή στο μονοπάτι.

**4.4.2** Εάν ισχύει η παραπάνω συνθήκη αύξησε τον μετρητή inter κατά 1

**4.5** Τέλος εσωτερικού βρόγχου

**4.6** Εάν path\_length == inter, πρόσθεσε το ζευγάρι (taxi\_id,traj\_id) στη λίστα trajectories

**4.7** Tέλος εξωτερικού βρόγχου

**5**. Διέγραψε τα διπλότυπα ζευγάρια (εάν υπάρχουν) από τη λίστα trajectories και επέστρεψε το μήκος της.

ΤΕΛΟΣ ΑΛΓΟΡΙΘΜΟΥ SPQ

Η ιδεά της χρήσης των ΑΕΜ έχει προέλθει από άλλη έρευνα, όπως δηλώσαμε και στο πρώτο κεφάλαιο. Η όλη ιδέα έχει υλοποιηθεί σε PostgreSQL και PL/pgSQL. Στην παρούσα μελέτη έχουμε καταφέρει να υλοποιήσουμε την ίδια μεθοδολογία των ΑΕΜ και σε γλώσσα PL/pgSQL εκμεταλλευόμενοι τις δυνατότητες της βάσης περί ταχύτητας δημιουργώντας ευρετήρια B+ δέντρων σε στήλες και σε συνδυασμό στηλών του πίνακα «visited\_segments».

Αρχικά, δημιουργούμε ένα ευρετήριο πάνω στην στήλη «OSM\_Way\_ID». Αυτό το ευρετήριο βοηθά στην ταχεία ανάκτηση δεδομένων όταν αναφερόμαστε σε συγκεκριμένες ακμές του δρόμου. Στην συνέχεια, δημιουργούμε ένα δεύτερο ευρετήριο πάνω στις στήλες «Time\_Enter» και «Time\_Leave». Αυτό το ευρετήριο επιτρέπει την αποτελεσματική αναζήτηση και το φιλτράρισμα εγγραφών με βάση τα χρονικά διαστήματα εισόδου και εξόδου. Τέλος, δημιουργούμε ένα τρίτο ευρετήριο πάνω στις στήλες «OSM\_Way\_ID», «Traj\_ID» και «Taxi\_ID». Αυτό το ευρετήριο βοηθά στην γρήγορη αναζήτηση και συγκριτική ανάλυση εγγραφών με βάση τα κριτήρια που περιλαμβάνουν τα αναγνωριστικά των ακμών και των τροχιών. Τέλος, υλοποιούμε την συνάρτηση SPQ σε γλώσσα PL/pgSQL με τον ίδιο τρόπο, όπως ακριβώς υλοποιήθηκε και στην γλώσσα Python, χρησιμοποιώντας τις ίδιες παραμέτρους σε κάθε κλήση της.

Αφού έχουμε στην διάθεσή μας δύο όμοιες συναρτήσεις υλοποιημένες σε διαφορετικές γλώσσες, μπορούμε να τις συγκρίνουμε ως προς την ταχύτητα. Για να είμαστε όσο τον δυνατόν ακριβέστεροι, στο πείραμα αυτό καλούμε κάθε συνάρτηση με τις ίδιες παραμέτρους. Πρώτον, ορίζεται ένα σύνολο από δεκατέσσερις δέσμες. Κάθε δέσμη περιλαμβάνει είκοσι διαφορετικά μονοπάτια ίσου μήκους σε συγκεκριμένα χρονικά διαστήματα. Η πρώτη δέσμη περιλαμβάνει είκοσι μονοπάτια μήκους δύο, η δεύτερη δέσμη περιλαμβάνει είκοσι μονοπάτια μήκους τρία κ.ο.κ. Ουσιαστικά, κάθε δέσμη είναι είκοσι κλήσεις της συνάρτησης των ΑΕΜ. Δεύτερο, σε κάθε προγραμματιστικό περιβάλλον (Python ή Postgres) εκτελούμε μαζικά μία δέσμη και μετράμε τον συνολικό χρόνο εκτέλεσης των είκοσι κλήσεων της συνάρτησης SPQ σε αυτό. Το αποτέλεσμα αυτού του πειράματος συνοψίζεται στο ακόλουθο γράφημα:

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμμή, γράφημα

Περιγραφή που δημιουργήθηκε αυτόματα

Στον οριζόντιο άξονα περιλαμβάνεται το μήκος των μονοπατιών που περιέχει κάθε δέσμη. Ο κατακόρυφος άξονας περιέχει τον χρόνο εκτέλεσης των είκοσι ερωτημάτων της δέσμης. Με μπλε χρώμα σημειώνεται το περιβάλλον της PostgreSQL και με πορτοκαλί, το περιβάλλον της Python.

Συμπεραίνουμε ότι οι χρόνοι εκτέλεσης σε Python είναι ταχύτεροι, συγκριτικά με αυτούς σε PostgreSQL. Για αυτό τον λόγο έχουμε υιοθετήσει την υλοποίηση της συνάρτησης SPQ σε Python. Επομένως, δεν θεωρούμε σημαντική την εξήγηση του αλγορίθμου που δημιουργήθηκε σε γλώσσα PL/pgSQL. Ωστόσο, ο κώδικας παρατίθεται στο αρχείο «**SQP func in PLpgsql.txt**»

### **Το τελικό σύνολο δεδομένων**

Στην ανάλυσή μας ασχολούμαστε με την κατασκευή ενός συνόλου δεδομένων που εστιάζει στην καταγραφή της ροής της κυκλοφορίας στο οδικό δίκτυο ανά μονοπάτι και ανά χρονικό διάστημα. Για αυτή την διαδικασία χρησιμοποιούμε όλα τα αντιστοιχισμένα στο οδικό δίκτυο δεδομένα που υπάρχουν στον πίνακα «visited\_segments». Στην διάθεσή μας έχουμε δεδομένα μιας ολόκληρης εβδομάδας. Επομένως, χωρίζουμε αυτό το μεγάλο χρονικό διάστημα  σε μικρότερα διαστήματα διάρκειας μισής ώρας έκαστο. Επομένως, σε κάθε ημέρα αντιστοιχούν (24\*60)/30 = 48  χρονικά διαστήματα μισής ώρας.

Προχωρώντας, πρέπει να οριστούν τα μονοπάτια, πάνω στα οποία θα γίνει η ανάλυση της κυκλοφοριακής ροής. Για αυτό τον λόγο, δημιουργούμε χίλια μοναδικά μονοπάτια διαφορετικού μήκους το κάθε ένα. Ο λόγος για τον οποίο επιλέχθηκαν τόσα μονοπάτια είναι για να επιταχύνουμε την ταχύτητα εκτέλεσης του κώδικα. Κάθε μονοπάτι αντιπροσωπεύει μια ακολουθία συνεχόμενων οδικών ακμών που έχει διανύσει ένα ταξί. Οι ακμές αυτές δεν είναι τυχαίες, αλλά πηγάζουν άμεσα από τα δεδομένα μας, διασφαλίζοντας ότι τα μονοπάτια που δημιουργούνται ακολουθούν την ιδιότητα της συνέχειας, όπως ορίστηκε στο πρώτο κεφάλαιο. Το μήκος αυτών των μονοπατιών μπορεί να κυμαίνεται από δύο έως και δεκαπέντε ακμές.

Πλέον, η κατασκευή του συνόλου δεδομένων χρονοσειρών είναι εύκολη υπόθεση, αφού τα δομικά συστατικά που θα αποτελείται είναι στην διάθεσή μας. Οι χρονοσειρές αποθηκεύονται σε έναν πίνακα με όνομα «time\_series\_SPQ». Κάθε εγγραφή σε αυτόν τον πίνακα περιλαμβάνει το μονοπάτι (δηλαδή την λίστα από τις ακμές που αποτελείται), την τροχιά (Taxi ID, Traj ID) που βρίσκεται αυτό το μονοπάτι, το μήκος του μονοπατιού ως προς τις ακμές του και το πλήθος των ταξί που διέσχισαν το συγκεκριμένο μονοπάτι ανά μισή ώρα. Συγκεκριμένα, οι στήλες που αποτελείται ο νέος πίνακας είναι οι εξής: «Path», «Length», «Taxi ID», «Traj ID» και τα χρονικά διαστήματα. Κάθε χρονικό διάστημα είναι και μία διαφορετική στήλη στον πίνακα.

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, αριθμός, γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα

Το τελικό σύνολο δεδομένων

Ωστόσο, για να διευκολύνουμε την ανάλυσή μας, έχουμε τροποποιήσει το παραπάνω σύνολο δεδομένων ως προς τις στήλες, χωρίς να υπάρχει απώλεια πληροφορίας. Οι προηγούμενες στήλες που αναφέρθηκαν παραμένουν ως έχει, εκτός από τα χρονικά διαστήματα που έχουν τοποθετηθεί σε μία καινούρια στήλη με όνομα «Time Column». Ταυτόχρονα, η κυκλοφοριακή ροή (ο αριθμός των αυτοκινήτων που διέσχισαν ένα μονοπάτι) έχει προστεθεί ως ξεχωριστή στήλη στο σύνολο των δεδομένων (στήλη «Traffic Flow»). Ουσιαστικά, μετασχηματίσαμε το τελικό σύνολο δεδομένων σε μία μορφή που είναι περισσότερο ευανάγνωστη, όπως φαίνεται στην επόμενη φωτογραφία:

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, αριθμός

Περιγραφή που δημιουργήθηκε αυτόματα

Απόσπασμα από το μετασχηματισμένο σύνολο δεδομένων. Η κυκλοφοριακή ροή και η χρονική πληροφορία πλέον αποτελούν ξεχωριστές στήλες, κάνοντας τα δεδομένα πιο κατανοητά.

Τώρα, αυτός ο νέος πίνακας θα αποτελέσει τον βασικό πρωταγωνιστή της μελέτης μας.

## **Πώς επηρεάζουν τα ΑΕΜ την πληροφορία της κυκλοφοριακής ροής;**

Η επιλογή των ΑΕΜ στο πρόβλημα της πρόβλεψης της κυκλοφοριακής ροής εντός ολόκληρων μονοπατιών μέσα σε ένα οδικό δίκτυο, δηλαδή την πρόβλεψη του αριθμού των ταξί που θα διανύσουν ένα μονοπάτι την επόμενη χρονική στιγμή, ευθύνεται στην ακόλουθη γνώση: τα ΑΕΜ μας εγγυόνται ότι τα ταξί δεν θα παρεκκλίνουν της πορείας τους κατά την διάσχιση του μονοπατιού. Επίσης, η διάσχιση αυτή είναι σίγουρο ότι θα γίνεται με την σωστή σειρά, δηλαδή οι ακμές που αποτελείται το μονοπάτι θα διανύονται μία προς μία από την πρώτη έως και την τελευταία.

Για να δούμε την διαφορά στην κυκλοφοριακή ροή που συλλέγουμε όταν απουσιάζει η χρήση της μεθοδολογίας των ΑΕΜ, έχουμε δημιουργήσει ένα καινούριο σύνολο δεδομένων χρονοσειρών, με τον ίδιο τρόπο που περιεγράφηκε προηγουμένως. Η κύρια διαφορά τώρα είναι ότι σε κάθε μονοπάτι δεν μετράμε την κυκλοφοριακή ροή με την βοήθεια της συνάρτησης SPQ. Σε αυτή την περίπτωση, η κυκλοφοριακή ροή σε κάθε μονοπάτι ορίζεται ως το πλήθος των τροχιών που έχουν διανύσει τουλάχιστον μία φορά όλες τις ακμές που απαρτίζουν το μονοπάτι εντός ενός συγκεκριμένου χρονικού διαστήματος.

Το παρακάτω γράφημα δείχνει την διαφορά ανάμεσα στα δύο σύνολα δεδομένων. Με μπλε χρώμα απεικονίζεται η πληροφορία που προκύπτει εφαρμόζοντας την μέθοδο των ΑΕΜ, ενώ με πορτοκαλί χρώμα παρουσιάζεται η καταγραφή όταν απουσιάζει αυτή η μέθοδος. Επίσης, για κάθε χρονική στιγμή έχουμε συμπεριλάβει το άθροισμα των ταξί που διένυσαν όλα τα μονοπάτια του συνόλου δεδομένων.

Εικόνα που περιέχει κείμενο, γραμματοσειρά, στιγμιότυπο οθόνης, γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα

Στον οριζόντιο άξονα απεικονίζεται ο χρόνος, ενώ ο κατακόρυφος άξονας μετράει το συνολικό άθροισμα της κυκλοφοριακής ροής σε όλα τα μονοπάτια.

Αξιοσημείωτο είναι το γεγονός ότι η κυκλοφοριακή ροή παρουσιάζει παρόμοια συμπεριφορά (τάση, περιοδικότητα) και στα δύο σύνολα δεδομένων. Ωστόσο, η παρουσία περιορισμών που διακατέχει την μέθοδο των ΑΕΜ οδηγεί σε εγκυρότερα δεδομένα. Προτιμάται, λοιπόν, η χρήση των ΑΕΜ, αφού αναφερόμαστε σε κινητικότητα μέσα σε ολόκληρο μονοπάτι.

## **Προσθήκη επιπλέον πληροφορίας στο τελικό dataset**

Ενσωμάτωση δεδομένων καιρού

Ένας παράγοντας που επηρεάζει συχνά την κυκλοφορία στους δρόμους είναι και ο καιρός. Για αυτό τον λόγο, σε συνδυασμό με τα δεδομένα κίνησης, έχουμε συμπεριλάβει και δεδομένα καιρού. Tα δεδομένα αυτά έχουν περιγραφεί αναλυτικά στο κεφάλαιο 3.

Επομένως, έχουμε ενοποιήσει τα δύο σύνολα δεδομένων σε ένα, κάνοντας κατάλληλη επεξεργασία. Και τα δύο σύνολα δεδομένων ανταποκρίνονται στο ίδιο χρονικό πλαίσιο.  Επιπλέον, τα δεδομένα κίνησης καταγράφονται ανά μισή ώρα, ενώ τα δεδομένα καιρού καταγράφονται ανά μία ώρα. Άρα, χρειάζεται να συνδεθούν με σωστό τρόπο οι δύο πίνακες: σε δύο εγγραφές δεδομένων κίνησης αντιστοιχίζεται μία εγγραφή δεδομένων καιρού. Με αυτόν τον τρόπο, η χρονική πληροφορία δεν χάνεται, αλλά παραμένει ακλόνητη. Το μόνο μειονέκτημα είναι ότι οι εγγραφές των δεδομένων καιρού αντιγράφονται συνολικά δύο φορές έκαστη.

Ενσωμάτωση χαρακτηριστικών που σχετίζονται με το χρόνο

Πολλά χαρακτηριστικά που σχετίζονται με τον χρόνο εξάγονται από τις πληροφορίες χρονοσφραγίδων στο σύνολο δεδομένων. Τα χαρακτηριστικά αυτά περιλαμβάνουν την ώρα, την ημέρα της εβδομάδας, την ημέρα του μήνα και τα λεπτά. Εφαρμόζεται κυκλική κωδικοποίηση σε ορισμένα χαρακτηριστικά (ώρα, ημέρα της εβδομάδας, ημέρα και λεπτό), η οποία αποτυπώνει την κυκλική φύση τους με την πάροδο του χρόνου. Επιπρόσθετα, το χαρακτηριστικό «3hour\_interval» εισάγεται, για να υποδεικνύει σε ποιο 3-ωρο χρονικό διάστημα της ημέρας εντοπίζεται αυτή η καταγραφή. Το χαρακτηριστικό αυτό λαμβάνει  τιμές από το ένα έως και το οκτώ (αφού 24 ώρες ανά ημέρα / 3 ώρες = 8 διαστήματα ανά ημέρα). Αυτή η πληροφορία μπορεί ενδεχομένως να καταγράψει τις διακυμάνσεις στη ροή της κυκλοφορίας κατά τη διάρκεια διαφορετικών τμημάτων της ημέρας.

Συμπέρασμα

Συνδυάζοντας δεδομένα καιρού και χαρακτηριστικά που σχετίζονται με τον χρόνο με τα υπάρχοντα δεδομένα ροής κυκλοφορίας, η προεπεξεργασία αποσκοπεί στη δημιουργία ενός ολοκληρωμένου συνόλου δεδομένων που ενσωματώνει τόσο εξωτερικούς περιβαλλοντικούς παράγοντες (καιρός) όσο και εγγενή χρονικά πρότυπα (χαρακτηριστικά που σχετίζονται με τον χρόνο). Αυτό το εμπλουτισμένο σύνολο δεδομένων μπορεί δυνητικά να ενισχύσει τις προγνωστικές δυνατότητες ενός μοντέλου πρόβλεψης ροής κυκλοφορίας, επιτρέποντάς του να εξετάσει ένα ευρύτερο φάσμα επιρροών στις μεταβολές της ροής κυκλοφορίας. Ο γενικός στόχος αυτών των προσθηκών είναι η βελτίωση της ακρίβειας και της αποτελεσματικότητας του επακόλουθου μοντέλου πρόβλεψης ροής κυκλοφορίας.

## **Οπτικοποίηση των δεδομένων**

Μία από τις βασικές αρχές στην επιστήμη των δεδομένων αποτελεί η οπτικοποίηση των δεδομένων. Με αυτόν τον τρόπο, ο ερευνητής μπορεί εύκολα να κατανοήσει σημαντικές πτυχές στα δεδομένα που δεν μπορούν να παρατηρηθούν αλλιώς. Σε αυτό το υποκεφάλαιο προσπαθούμε να ανακαλύψουμε την συμπεριφορά της κυκλοφοριακής ροής στο χρονικό διάστημα της μίας εβδομάδας που εξετάζουμε χρησιμοποιώντας διαγράμματα.

**Πρώτο Διάγραμμα:** Συνολική Ροή Κυκλοφορίας σε Κάθε Ημέρα

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γράφημα, γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα

Σε αυτό το διάγραμμα, φαίνεται η συνολική ροή της κυκλοφορίας, δηλαδή το άθροισμα της κυκλοφοριακής ροής σε όλα τα μονοπάτια του συνόλου δεδομένων, ανά χρονικό διάστημα. Στον οριζόντιο άξονα έχει τοποθετηθεί ο χρόνος, ο οποίο κυμαίνεται από τις 18 έως και 24 του Μάιου (μία ολόκληρη εβδομάδα). Το χρώμα της γραμμής αντιπροσωπεύει την ημέρα της εβδομάδας.

Μέσα από αυτό το γράφημα μπορούμε να παρατηρήσουμε πως αλλάζει η ροή της κυκλοφορίας κατά τη διάρκεια της εβδομάδας. Συγκεκριμένα, σε κάθε ημέρα, κατά τις πρωινές και βραδινές ώρες  η κυκλοφοριακή ροή είναι αυξημένη, ενώ τις μεσημεριανές και απογευματινές ώρες παρατηρείται μικρότερη κινητικότητα μέσα στο οδικό δίκτυο. Επομένως, υπάρχει μία σταθερή περιοδικότητα στα δεδομένα.

Στην συνέχεια, γίνεται μία αναλυτικότερη απεικόνιση της κίνησης με βάση την ημέρα και το 3-ωρο χρονικό διάστημα. Για παράδειγμα, το επόμενο γράφημα χωρίζει το άθροισμα της κυκλοφοριακής ροής όλων των μονοπατιών κατά την ημέρα 2008-05-18 σε 3-ωρα χρονικά διαστήματα. Κάθε τρίωρο χρονικό διάστημα φαίνεται με διαφορετικό χρώμα. Το υπόμνημα βοηθάει στην αποσαφήνιση τίνος τρίωρου αντιστοιχείται κάθε χρώμα. Η πληροφορία σε κάθε άξονα είναι η ίδια όπως και στο προηγούμενο γράφημα.

Εικόνα που περιέχει γράφημα, γραμμή, διάγραμμα, κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Η παραπάνω απεικόνιση μας επιτρέπει να αναφέρουμε ότι κατά το τέταρτο τρίωρο της ημέρας παρατηρήθηκε η χαμηλότερη αθροιστική κυκλοφορία. Από την άλλη, κατά το τρίτο τρίωρο παρατηρήθηκε η υψηλότερη αθροιστική κυκλοφορία. Επιπλέον, μπορεί να εκφραστεί με σιγουριά ότι στα πρώτα τρία τρίωρα, δηλαδή τις πρώτες εννέα ώρες της ημέρας υπάρχει αυξημένη κινητικότητα, ενώ στο τέταρτο τρίωρο (δηλαδή για τις επόμενες τρεις ώρες) παρατηρείται χαμηλή κινητικότητα κ.ο.κ.

Με παρόμοια γραφήματα, εξετάζεται η κυκλοφοριακή ροή ανά τρίωρο για κάθε μία από τις υπόλοιπες ημέρες της εβδομάδας. Τα διαγράμματα αυτά είναι διαθέσιμα στο φάκελο **Images/ Info about time series dataset**. Γενικά, μέσω αυτών των διαγραμμάτων, γίνονται κατανοητά τα παρακάτω:

* συνολική κυκλοφορία κάθε ημέρας: καθίστανται ευδιάκριτες οι τάσεις και τα μοτίβα της κυκλοφορίας κατά τη διάρκεια της εβδομάδας. Παρατηρείται εάν υπάρχει κάποια συγκεκριμένη μέρα με υψηλότερη ή χαμηλότερη κυκλοφορία.
* κορυφές και κοιλάδες: μπορούμε να εντοπίζουμε τις ώρες κατά τις οποίες η κυκλοφορία είναι στο αποκορύφωμά της κατά τη διάρκεια μιας συγκεκριμένης ημέρας. Ταυτόχρονα, εντοπίζεται εύκολα πότε η κυκλοφοριακή ροή είναι χαμηλή.
* συγκρίσεις ημερών: δίνεται δυνατότητα να συγκρίνουμε την κυκλοφορία μεταξύ διαφορετικών ημερών της εβδομάδας και να παρατηρούμε αν υπάρχουν διαφορές στα μοτίβα κυκλοφορίας μεταξύ των ημερών.

## **Χρήση Μοντέλων Μηχανικής και Βαθιάς Μάθησης**

### **Διαχωρισμός σε σύνολα εκπαίδευσης και ελέγχου**

Η κατάλληλη προετοιμασία των δεδομένων οδηγεί σε πιο ποιοτικά αποτελέσματα. Σε αυτό το βήμα, τα δεδομένα διαιρούνται σε δύο σύνολα: το σύνολο εκπαίδευσης (train set) και το σύνολο ελέγχου (test set). Το πρώτο σύνολο αποτελείται από δεδομένα που θα χρησιμοποιηθούν για την εκπαίδευση κάθε αλγορίθμου μηχανικής μάθησης, με σκοπό οι τελευταίοι να μάθουν διάφορα μοτίβα και σχέσεις σε αυτά. Mε άλλα λόγια, κατά την φάση της εκπαίδευσης, το μοντέλο προσαρμόζει τις παραμέτρους του, για να μπορεί να προβλέπει σωστά τα αποτελέσματα σε νέα μη γνωστά δεδομένα. Αντιθέτως, τα δεδομένα ελέγχου αποτελούν δεδομένα, στα οποία το μοντέλο δεν έχει εκπαιδευτεί και εξετάζουν την ικανότητα γενίκευσής του. Για αυτό τον λόγο, η απόδοση του μοντέλου αξιολογείται με βάση το πόσο καλά προβλέπει τα μοτίβα που υπάρχουν στα δεδομένα ελέγχου. Τελευταίο αλλά εξίσου σημαντικό είναι και το σύνολο επικύρωσης (validation set) που χρησιμοποιείται εκτεταμένα στην έρευνά μας. Τα δεδομένα αυτά χρησιμοποιούνται για τη βελτιστοποίηση των υπερπαραμέτρων του μοντέλου κατά τη διάρκεια της εκπαίδευσης, εξασφαλίζοντας πως το μοντέλο γενικεύει καλά σε νέα δεδομένα, χωρίς να επηρεάζεται από την επίδοσή του στα δεδομένα εκπαίδευσης. Συχνά, το validation set ταυτίζεται με το test set στην μελέτη αυτή.

Το χρονικό διάστημα όλων των παρατηρήσεων που υπάρχουν στην διάθεσή μας κυμαίνεται μεταξύ των ημερομηνιών 2008-05-18 και 2008-05-24. Στην μελέτη αποφασίστηκε το σύνολο εκπαίδευσης (train set) να περιέχει όλα τα δεδομένα για κάθε μονοπάτι μέχρι και την 2008-05-23. Τα υπόλοιπα δεδομένα (τα πιο πρόσφατα δηλαδή) βρίσκονται στο σύνολο ελέγχου (test set).

### **Διαχωρισμός των χαρακτηριστικών σε σύνολα Feature και Label**

Αφού γίνει ο διαχωρισμός σε σύνολα εκπαίδευσης και ελέγχου, προσπαθούμε να επιλέξουμε εκείνα τα χαρακτηριστικά που θα βοηθήσουν τους αλγορίθμους να προβλέψουν τις εκάστοτε τιμές της κυκλοφοριακής ροής, γνωστά και ως features. Ο πίνακας συσχέτισης (correlation matrix) είναι ένα εργαλείο που βοηθά στην επιλογή αυτών των χαρακτηριστικών κατά την ανάπτυξη ενός μοντέλου μηχανικής μάθησης. Ο ρόλος του είναι να παρουσιάσει τις συσχετίσεις μεταξύ των διαφόρων χαρακτηριστικών στα δεδομένα και να βοηθήσει τον προγραμματιστή στην κατασκευή ενός καλύτερου μοντέλου. Συγκεκριμένα, ο πίνακας συσχέτισης συμβάλλει στην:

* βελτιστοποίηση του μοντέλου: εάν το μοντέλο υποφέρει από υπερεκπαίδευση (overfitting), δηλαδή δεν μπορεί να γενικευτεί σε άλλα σύνολα δεδομένων, μια προσέγγιση που υιοθετείται είναι να μειωθεί ο αριθμός των χαρακτηριστικών που χρησιμοποιούνται κατά την εκπαίδευση. Ο πίνακας συσχέτισης μπορεί να βοηθήσει στην επιλογή των χαρακτηριστικών που πρέπει να διατηρηθούν, για να βελτιωθεί η γενίκευση του μοντέλου.
* ανίχνευση κοινών χαρακτηριστικών: κοινά χαρακτηριστικά που περιγράφουν τη ίδια πληροφορία μπορύν να εισαγάγουν θόρυβο στο μοντέλο μπερδεύοντάς το. Ο πίνακας συσχέτισης μπορεί να αναδείξει χαρακτηριστικά που έχουν υψηλή σχέση με άλλα χαρακτηριστικά, υποδηλώνοντας ότι μπορεί να είναι περιττά και ίσως να χρειαστεί να μην χρησιμοποιηθούν στην εκπαίδευση.

Κατά την εκπαίδευση των μοντέλων μηχανικής και βαθιάς μάθησης έχει χρησιμοποιηθεί ένας τέτοιος πίνακας συσχέτισης, όπως φαίνεται στην ακόλουθη εικόνα:

Εικόνα που περιέχει στιγμιότυπο οθόνης, πολυχρωμία, μοτίβο, γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα

Σε αυτό το γράφημα παρουσιάζεται ένας πίνακας διαστάσεων N\*N, όπου Ν είναι το πλήθος των χαρακτηριστικών στο σύνολο δεδομένων. Οι χρωματικές αντιστοιχίες σε κάθε κελί χρησιμοποιούνται για να απεικονίσουν τις τιμές συσχέτισης μεταξύ των χαρακτηριστικών που αναλογούν σε αυτό το κελί. Οι νόμιμες τιμές που λαμβάνει μία μίτρα συσχέτισης κυμαίνονται μεταξύ του μείον ένα και του ενός (από το -1 έως και το 1). Τα σκούρα χρώματα (μωβ, μπλε) αντιστοιχούν σε αρνητική τιμή συσχέτισης (κοντά στο -1) ή ακόμα και ανύπαρκτη συσχέτιση (κοντά στο 0). Αν η τιμή της συσχέτισης είναι κοντά στο μείον ένα, τότε υπάρχει αρνητική γραμμική συσχέτιση μεταξύ των δύο μεταβλητών. Αυτό σημαίνει ότι όταν μία μεταβλητή αυξάνεται, η άλλη μειώνεται σύμφωνα με μια γραμμική σχέση. Από την άλλη, όταν η τιμή της συσχέτισης είναι κοντά στο μηδέν, δεν υπάρχει γραμμική συσχέτιση ανάμεσα στις μεταβλητές. Αυτό σημαίνει ότι οι μεταβλητές δεν συσχετίζονται με τρόπο που να μπορεί να περιγραφεί με μια γραμμική σχέση. Ωστόσο, αυτό δεν σημαίνει απαραιτήτως ότι δεν υπάρχει καμία άλλη γραμμική συσχέτιση μεταξύ τους. Τέλος, τα ανοιχτά χρώματα (κίτρινο, πράσινο) αντιστοιχούν σε υψηλή τιμή συσχέτισης (κοντά στο 1), δηλώνοντας ότι οι δύο μεταβλητές παρουσιάζουν θετική συσχέτιση. Κάθε αλλαγή στην τιμή της μίας μεταβλητής, επηρεάζει την άλλη κατά ανάλογο τρόπο. Με άλλα λόγια, υπάρχει μία γραμμική σχέση που διέπει τις δύο μεταβλητές.

### **Εφαρμογή της τεχνικής του κυλιόμενου παραθύρου στον αλγόριθμο XGBoost**

Το κυλιόμενο παράθυρο (ή sliding window) είναι μια τεχνική που χρησιμοποιείται στην ανάλυση χρονοσειρών εξασφαλίζοντας προβλέψεις με βάση προηγούμενες παρατηρήσεις (ιστορικά δεδομένα). Στην ουσία, χρησιμοποιείται ένα κινούμενο παράθυρο που κυλάει κατά μήκος της χρονοσειράς, επιτρέποντας τη δημιουργία πολλαπλών προβλέψεων. Σε ένα πρόβλημα ανάλυσης ή πρόβλεψης χρονοσειρών, η τεχνική του συρόμενου παραθύρου χρησιμοποιείται για διάφορους λόγους:

* οι παρατηρήσεις από το παρελθόν μπορούν να χρησιμοποιηθούν, για να προβλέψουν το μέλλον. Με το sliding window, δημιουργούνται διαδοχικά παράθυρα που περιλαμβάνουν τις προηγούμενες παρατηρήσεις, και το μοντέλο εκπαιδεύεται σε αυτά τα παράθυρα για να παράγει τις μελλοντικές τιμές του μεγέθους που θέλουμε να προβλέψουμε.
* με την πρόοδο του χρόνου, το sliding window επιτρέπει τη συνεχή ενημέρωση του μοντέλου με νεότερες παρατηρήσεις, ενισχύοντας την ικανότητα πρόβλεψης με βάση τις τελευταίες πληροφορίες.

Η μεθοδολογία αυτή εγγυάται ότι ο αλγόριθμος εκπαιδεύεται σε ολόκληρη την χρονοσειρά (ή τις χρονοσειρές) που συμπεριλαμβάνεται (συμπεριλαμβάνονται) στα δεδομένα εκπαίδευσης. Η ίδια ακριβώς τεχνική χρησιμοποιείται και στα δεδομένα ελέγχου. Χρησιμοποιώντας διαφορετικά μήκη παραθύρου, δηλαδή το πόσα ιστορικά δεδομένα χρησιμοποιούνται κάθε φορά στο παράθυρο για την πρόβλεψη, μπορούμε να αξιολογήσουμε πως η εφαρμογή τους επηρεάζει την απόδοση του μοντέλου. Αυτό βοηθά να ευρεθεί το βέλτιστο μέγεθος παραθύρου για τη συγκεκριμένη χρονοσειρά.

Ο αλγόριθμος XGBoost

Ο αλγόριθμος XGBoost (Extreme Gradient Boosting) είναι ένας πανίσχυρος αλγόριθμος μηχανικής μάθησης που δημιουργήθηκε από τον Tianqi Chen. Βασίζεται σε δέντρα αποφάσεων και χρησιμοποιείται ευρέως για προβλήματα παλινδρόμησης (Regression) και ταξινόμησης (Classification).

Ο μηχανισμός αυτός έχει ως θεμέλιο λίθο την τεχνική Gradient Boosting. Η τελευταία πρόκειται για μια ευρέως χρησιμοποιούμενη μέθοδο μηχανικής μάθησης και απαντάται συχνά σε αλγορίθμους που υιοθετούν δέντρα, για να προβαίνουν σε προβλέψεις. Η τεχνική Gradient Boosting εμπίπτει στην κατηγορία της μάθησης συνόλου (ensemble learning), η οποία συνδυάζει ασθενέστερα μοντέλα δέντρων (week learners) για τη δημιουργία ενός ισχυρού (strong learner). Κάθε δέντρο κατασκευάζεται το ένα μετά το άλλο. Η ιδιαιτερότητα της μεθόδου είναι ότι το επόμενο δέντρο απόφασης που δημιουργείται προσπαθεί να μειώσει το σφάλμα του προηγούμενου. Η τελική πρόβλεψη που θα προκύψει αποτελεί το άθροισμα όλων των προβλέψεων όλων των δέντρων.

Ο αλγόριθμος XGBoost αποτελεί μια βελτιωμένη έκδοση αυτής της μεθόδου και έχει χρησιμοποιηθεί από πολλούς ερευνητές λόγω των εντυπωσιακών επιδόσεών του. Επιπρόσθετα, ο αλγόριθμος συνδυάζεται με ένα πλήθος υπερπαραμέτρων που πρέπει να οριστούν από τον ερευνητή κατά την αρχικοποίησή του. Οι υπερπαράμετροι αυτοί χρησιμοποιούνται για την βελτιστοποίηση των προβλέψεων και την αποφυγή της υπερεκπαίδευσης (overfitting). Στην έρευνα που έχουμε κάνει, χρησιμοποιούνται πέντε υπερπαράμετροι, οι **gamma**, **alpha**, **max\_depth**, **n\_estimators** και **learning\_rate**.

1. υπερπαράμετρος gamma: βοηθάει το μοντέλο να αποφεύγει τo overfitting. Καθώς δημιουργούνται week learners, τα δεδομένα που αναπαρίστανται σε αυτά αποθηκεύονται σε κόμβους και κλαδιά. Το δέντρο μπορεί τελικά να φτάσει σε ένα μεγάλο βάθος (το βάθος ή τα επίπεδα που μπορεί να έχει το δέντρο επιλέγονται από τον προγραμματιστή). Όσο πιο βαθύ είναι το δέντρο, τόσο περισσότερο έχει εντρυφήσει το συγκεκριμένο δέντρο στο σύνολο εκπαίδευσης, οδηγώντας το σε υπερεκπαίδευση. Η υπερπαράμετρος αυτή χρησιμοποιείται για να «κλαδεύει» τα δέντρα απόφασης ώστε να μειωθεί το βάθος τους και να αποφευχθεί το overfitting.
2. υπερπαράμετρος alpha: προσθέτει έναν επιπλέον όρο στην συνάρτηση σφάλματος που χρησιμοποιείται κατά την εκπαίδευση. Ανάλογα με την τιμή του alpha, το μοντέλο γίνεται πιο αυστηρό ή ανεκτό σε σφάλματα, επηρεάζοντας ανάλογα και τις τιμές των παραμέτρων κατά την φάση της εκπαίδευσης. Τελικά, αυτός ο επιπλέον όρος ελέγχει την πολυπλοκότητα των δέντρων που δημιουργούνται και αποτρέπει το overfitting.
3. υπερπαράμετρος n\_estimators: καθορίζει τον τελικό αριθμό των δέντρων (week learners) που θα δημιουργηθούν. Όσο περισσότερα δέντρα προστίθενται, τόσο πιο πολύπλοκο γίνεται το μοντέλο ενέχοντας τον κίνδυνο υπερεκπαίδευσης. Η εύρεση της κατάλληλης τιμής της υπερπαραμέτρου αυτής είναι σημαντική για την ισορροπία μεταξύ απόδοσης του μοντέλου και χρόνου εκπαίδευσης.
4. υπερπαράμετρος max\_depth: ορίζει το μέγιστο βάθος που μπορεί να έχει ένα δέντρο. Ένα βαθύ δέντρο μπορεί να προσδιορίσει σύνθετες σχέσεις στα δεδομένα, αλλά συνήθως οδηγεί σε overfitting. Η σωστή τιμή για το max\_depth βοηθά στο να δημιουργηθούν δέντρα που γενικεύουν καλά τα δεδομένα.
5. υπερπαράμετρος learning\_rate: ελέγχει το βήμα, με το οποίο το μοντέλο προσαρμόζεται στα δεδομένα εκπαίδευσης.

Εύρεση βέλτιστων υπερπαραμέτρων

Ένα βασικό ερώτημα που προκύπτει είναι πώς θα ευρεθούν οι κατάλληλες τιμές για αυτές τις υπερπαραμέτρους. Την απάντηση σε αυτό το πρόβλημα δίνει η μέθοδος Grid Search Cross Validation (GridSearchCV). Πρόκειται για μια τεχνική που χρησιμοποιείται για την αυτόματη εύρεση των βέλτιστων τιμών των υπερπαραμέτρων ενός μοντέλου. Οι υπερπαράμετροι είναι παράμετροι που δεν μαθαίνονται από το μοντέλο κατά τη διάρκεια της εκπαίδευσης, αλλά πρέπει να προκαθοριστούν πριν από αυτή. Η μέθοδος GridSearchCV λειτουργεί εφαρμόζοντας τα παρακάτω βήματα:

1. ορίζεται μια λίστα πιθανών τιμών για κάθε υπερπαράμετρο που χρειάζεται να βελτιστοποιηθεί.
2. η μέθοδος GridSearchCV δημιουργεί όλους τους δυνατούς συνδυασμούς των τιμών των υπερπαραμέτρων που έχουν καθοριστεί. Για παράδειγμα, εάν έχουν οριστεί δύο υπερπαράμετροι προς βελτιστοποίηση, καθεμία με τρεις δυνατές τιμές, θα δημιουργηθούν συνολικά 3 \* 3 = 9 διαφορετικοί συνδυασμοί υπερπαραμέτρων.
3. για κάθε συνδυασμό υπερπαραμέτρων, το μοντέλο εκπαιδεύεται υιοθετώντας το εκάστοτε συνδυασμό και αξιολογείται με τον τρόπο που ορίζει η τεχνική Cross Validation.
4. μετά την αξιολόγηση όλων των συνδυασμών, η GridSearchCV επιλέγει τον συνδυασμό υπερπαραμέτρων που παρήγαγε τα καλύτερα αποτελέσματα.

Συμπεραίνοντας, είναι κατανοητό ότι η παρούσα μεθοδολογία αποτελεί μια πολύ χρήσιμη τεχνική για την αυτοματοποίηση της διαδικασίας επιλογής υπερπαραμέτρων και τη βελτιστοποίηση της απόδοσης ενός μοντέλου.

Βελτιστοποίηση του αλγορίθμου XGBoost

Για το σύνολο δεδομένων που υπάρχει στην κατοχή μας, έχουμε εκπαιδεύσει τον αλγόριθμο XGBoost χρησιμοποιώντας κάθε φορά διάφορα μήκη παραθύρου. Το παρακάτω γράφημα περιγράφει πως μειώνεται ή αυξάνεται το σφάλμα RMSE των προβλέψεων του συγκεκριμένου αλγορίθμου, καθώς αλλάζει το μήκος του παραθύρου.

Εικόνα που περιέχει κείμενο, γραμμή, γράφημα, διάγραμμα

Περιγραφή που δημιουργήθηκε αυτόματα

Διάγραμμα που δείχνει την σχέση του RMSE (κατακόρυφος άξονας) με το μήκος του παραθύρου που εφαρμόζεται κάθε φορά στα ίδια δεδομένα (οριζόντιος άξονας).

Ο λόγος που επιλέχθηκε ο αλγόριθμος XGBoost για αυτή την βελτιστοποίηση είναι διότι παρουσιάζει την καλύτερη επίδοση, σε σχέση με τα υπόλοιπα μοντέλα (LSTM, Random Forest και Encoder Decoder) που χρησιμοποιήθηκαν στην έρευνα. Ταυτόχρονα, για την εύρεση των βέλτιστων υπερπαραμέτρων που θα ωφελήσουν το μοντέλο να παρουσιάσει υψηλή απόδοση στα συγκεκριμένα δεδομένα, έχει χρησιμοποιηθεί η μέθοδος Grid Search CV.

### **Αξιολόγηση των μοντέλων μηχανικής μάθησης**

Μετά από όλη αυτή την διαδικασία που περιεγράφηκε, η οποία συνοψίζεται στα επόμενα βήματα: διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης και ελέγχου, επιλογή των καλύτερων χαρακτηριστικών (σύμφωνα με την μήτρα συσχέτισης) που θα βοηθήσουν την εκπαίδευση του μοντέλου, εφαρμογή του συρόμενου παραθύρου στα δεδομένα μας με το βέλτιστο μήκος, χρήση της τεχνικής Grid Search CV για βελτιστοποίηση των υπερπαραμέτερων, αναπαρίστανται στα επόμενα δύο γραφήματα οι επιδόσεις του μοντέλου XGBoost.

**Συμπεράσματα**

**Προτάσεις για Βελτίωση**