ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΟΜΕΑΣ ΜΕΤΑΦΟΡΩΝ ΚΑΙ ΣΥΓΚΟΙΝΩΝΙΑΚΗΣ ΥΠΟΔΟΜΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΣΙΔΗΡΟΔΡΟΜΙΚΗΣ ΚΑΙ ΜΕΤΑΦΟΡΩΝ

Μέθοδοι πρόβλεψης διάρκειας δραστηριοτήτων με σκοπό τη βελτίωση των μοτίβων δραστηριοτήτων και την κυκλοφοριακή ανάλυση σε ελληνικές πόλεις

ΟΝΟΜΑ ΚΑΙ ΕΠΩΝΥΜΟ ΣΠΟΥΔΑΣΤΗ: ΚΑΤΣΑΪΤΗΣ ΔΙΟΝΥΣΙΟΣ
ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ ΣΠΟΥΔΑΣΤΗ: 01117058

Επιβλέπων: Κωνσταντίνος Γκιοτσαλίτης / Επίκουρος Καθηγητής, ΕΜΠ Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής, Σχολή Πολιτικών Μηχανικών, ΕΜΠ

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF CIVIL ENGINEERING
DEPT. OF TRANSPORTATION PLANNING AND ENGINEERING
RAILWAYS AND TRANSPORT LABORATORY

Methods for estimating activity durations improving travel activity patterns recognition and analysis in Greek cities

Dionysios Katsaitis

Cv17058

Supervisor: Konstantinos Gkiotsalitis / Assistant Professor, NTUA Dept. of Transportation Planning and Engineering, School of Civil Engineering, NTUA

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κύριο Κωνσταντίνο Γκιοτσαλίτη, ο οποίος ήταν πάντα πρόθυμος να με βοηθήσει και να μου παρέχει τις επιστημονικές γνώσεις του και τις συμβουλές του με σκοπό το καλύτερο αποτέλεσμα της διπλωματικής μου εργασίας. Θα ήθελα επίσης να ευχαριστήσω τον Δημήτρη Ριζόπουλο, υποψήφιο διδάκτορα στον τομέα Μεταφορών και Συγκοινωνιακής Υποδομής του ΕΜΠ, για την πολύτιμη βοήθεια που μου έδινε στη συνεχή επικοινωνία που είχαμε αυτό το διάστημα. Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου για τη στήριξη που μου παρείχε σε αυτό το ταξίδι ως την εκπόνηση της διπλωματικής μου εργασίας.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Μέθοδοι πρόβλεψης διάρκειας δραστηριοτήτων με σκοπό τη βελτίωση των μοτίβων δραστηριοτήτων και την κυκλοφοριακή ανάλυση σε ελληνικές πόλεις

Διονύσιος Κατσαΐτης[1]

[a] Εθνικό Μετσόβιο Πολυτεχνείο, Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής, Εργαστήριο Σιδηροδρομικής και Μεταφορών, Ηρώων Πολυτεχνείου 5, 15773 Αθήνα, Ελλάδα

1.Εισαγωγή

Ο τομέας των μεταφορών επηρεάζεται από τις συνεχείς αλλαγές που βιώνουμε. Υπάρχει άμεση ανάγκη για αναβάθμιση των συστημάτων μεταφοράς έτσι ώστε να ανταποκρίνονται στις ανάγκες των ανθρώπων και στον απαιτούμενο σεβασμό στο περιβάλλον. Στα πλαίσια αυτής της διπλωματικής εργασίας αναλύονται τρόποι εκτίμησης διάρκειας δραστηριοτήτων με σκοπό τη βελτίωση της αναγνώρισης μοτίβων δραστηριοτήτων των μετακινούμενων και της κυκλοφοριακής ανάλυσης στις ελληνικές πόλεις. Για τους σκοπούς της εργασίας επιλέχθηκε να γίνει ανάλυση στην πόλη της Αθήνας και χρησιμοποιήθηκε το AthensPop, ένα Agent Based μοντέλο που χρησιμοποιώντας την PAM (Βιβλιοθήκη Python) συνθέτει πληθυσμό του οποίου καταγράφει τα μοτίβα δραστηριοτήτων ενώ ταυτόχρονα παράγει και κυκλοφοριακή ανάλυση. Η στόχευση γίνεται στον εμπλουτισμό της έρευνας – βάσης δεδομένων με σκοπό τα ακριβέστερα συνθετικά δεδομένα και ακολούθως μοτίβα δραστηριοτήτων. Όσο το δυνατόν καλύτερα αποτελέσματα μας δίνουν στοιχεία για τη συμπεριφορά των μετακινούμενων, ένα πολύπλοκο πρόβλημα πολλών μεταβλητών, και μας προσφέρει μία ευχέρεια και ευελιξία να μπορούμε να ελέγξουμε αλλαγές, νέες πολιτικές και στρατηγικές με το δυνατόν μικρότερο ρίσκο.

Στη σημερινή εποχή στις μεγαλουπόλεις οι ρυθμοί και οι υποχρεώσεις αυξάνονται ραγδαία. Ταυτόχρονα καλούμαστε να αντιμετωπίσουμε συνεχώς αλλαγές στην καθημερινότητα μας αφού έχει φτάσει η στιγμή που το να δράσουμε για το περιβάλλον δεν είναι πρόληψη αλλά αναγκαιότητα. Σύμφωνα με την απογραφή της ΕΛΣΤΑΤ[2] του 2021 το 50.1% του πληθυσμού της χώρας αντιστοιχεί σε 4.3% της έκτασής της και

[1] *cv17058@ntua.gr*
[2] https://www.statistics.gr/statistics/pop

συγκεντρώνεται κατά κύριο λόγο στην Αττική και στη Θεσσαλονίκη μαζί με κάποιες ευρύτερες περιοχές της . Είναι ξεκάθαρο πώς το φαινόμενο της αστικοποίησης είναι αρκετά έντονο και δημιουργεί προβληματισμό.

Γίνεται ευκόλως κατανοητό πως δε θα μπορούσε να μείνει ανεπηρέαστος και ο τομέας των μεταφορών από αυτό. Η Αθήνα σύμφωνα με το Traffic Index Ranking 2023 του TomTom[3] βρίσκεται στην 31^{η} θέση στις πόλεις με τη χειρότερη κίνηση σε όλον τον κόσμο (μέτρηση μέσου απαιτούμενου χρόνου ανά 10 χιλιόμετρα). Μάλιστα, παρατηρείται αύξηση 10 δευτερολέπτων στον σχετικό χρόνο σε σχέση με το 2022, κάτι που μας δείχνει την αναγκαιότητα επεμβάσεων. Η δημιουργία συστημάτων μεταφορών που θα σέβεται τις ανάγκες των πολιτών αλλά και το περιβάλλον αποτελεί επιτακτική ανάγκη και με δεδομένη τη συνεχή αύξηση των ήδη ιλιγγιωδών ρυθμών απαιτούνται λύσεις ως προς το κυκλοφοριακό σύστημα.

Έτσι, καταλαβαίνουμε πως η κατά το δυνατόν εγκυρότερη πρόβλεψη των μετακινήσεων αποτελεί την πρόκληση της εποχής αφού κάτι τέτοιο θα έλυνε πολλά προβλήματα ως προς τον σχεδιασμό και την ευελιξία των συστημάτων μεταφοράς. Ένα ακόμη πολύ χρήσιμο εργαλείο που θα μπορούσε να βοηθήσει είναι τα μοτίβα των ταξιδιών (travel patterns). Θεωρούνται πολύ σημαντικά καθώς σε αναλυτική μορφή δίνουν επιπλέον πληροφορίες από ότι ένας απλός αριθμός μετακινήσεων, όπως ο σκοπός της μετακίνησης, η διάρκεια, η απόσταση, η χρονική και χωρική κατανομή καθώς και η επιλογή του μέσου κ.α.. Στην πραγματικότητα, η προσπάθεια να προσομοιώσουμε τη συμπεριφορά των μετακινούμενων είναι πολύ δύσκολη αλλά και πολύ σημαντική καθώς αυτόματα μπορούμε να διερευνήσουμε την επίδραση νέων πολιτικών στο κυκλοφοριακό σύστημα και στους μετακινούμενους, αλλά και γενικότερα να αντιμετωπίσουμε το κυκλοφοριακό σύστημα σαν έναν ζωντανό οργανισμό και να παρατηρήσουμε πως συμπεριφέρεται σε αλλαγές έτσι ώστε να είμαστε σε θέση να παίρνουμε σωστές και ασφαλείς αποφάσεις. Στη χώρα μας βρισκόμαστε σε πρώιμο στάδιο όσον αφορά στη συλλογή δεδομένων καθώς δεν υπάρχουν ούτε στοιχεία απογραφών ούτε μεγάλες έρευνες που να συγκεντρώνουν κυκλοφοριακά δεδομένα. Κάτι τέτοιο κατευθύνει την πορεία της έρευνας σε συνθετικά δεδομένα καθώς είναι η μοναδική αξιόπιστη λύση όταν οι βάσεις δεδομένων είναι πολύ μικρές. Με ανάλογες παραδοχές και με τη χρήση των μοντέλων συντίθεται πληθυσμός

[3] https://www.tomtom.com/traffic-index/ranking/

και οι μετακινήσεις του σε ένα πλαίσιο προσομοίωσης που βοηθάει στον σχεδιασμό, στην πρόβλεψη και στη γενικότερη διαχείριση των συστημάτων μεταφορών. Επίσης, δεν έχουν γίνει ακόμα εκτεταμένες μελέτες πάνω στη μοντελοποίηση της ζήτησης καθώς γενικότερα ο τομέας της μοντελοποίησης είναι σε στάδιο ανάπτυξης. Για την πόλη της Αθήνας, υπάρχει το AthensPop ένα Agent Based μοντέλο, που αφορά τη μητροπολιτική Αθήνα, το οποίο δεχόμενο κάποια δεδομένα συνθέτει έναν πληθυσμό αντιπροσωπευτικό του πραγματικού. Στη συνέχεια γύρω από αυτόν τον πληθυσμό, αναπαράγεται μία υπολογιστική μέθοδος προσομοίωσης στραμμένη γύρω από τις αποφάσεις καθενός πράκτορα/ανθρώπου (agent), μέσα σε ένα συγκεκριμένο περιβάλλον. Από αυτό το μοντέλο γεννώνται οι μετακινήσεις καθώς και διάφορα χαρακτηριστικά τους.

Στη συγκεκριμένη εργασία δίνεται έμφαση στις διάρκειες δραστηριοτήτων. Πιο συγκεκριμένα, εξετάζεται πως η πρόβλεψή τους στις περιπτώσεις που χρειάζεται μπορεί να ωφελήσει στην ακρίβεια της εκτίμησης των μοτίβων δραστηριοτήτων και στην κυκλοφοριακή ανάλυση. Το μοντέλο που θα χρησιμοποιηθεί βασίζεται στη βελτίωση και επέκταση του απλού Cox μοντέλου του Lee (2000) με απώτερο σκοπό τη χρήση του στο AthensPop και την ακριβέστερη μέσα από την εκτίμηση διάρκειας δραστηριοτήτων (activity duration). Σε περιπτώσεις που έχουμε είτε ελλιπή δεδομένα είτε μη λογικά δεδομένα θα προσπαθήσουμε βάσει των ατομικών χαρακτηριστικών να προβλέψουμε τις όσο το δυνατόν πιο κοντινές διάρκειες στην πραγματικότητα, εμπλουτίζοντας έτσι τη βάση δεδομένων και παίρνοντας ασφαλέστερα αποτελέσματα εν συνεχεία.

Γίνεται σαφές πως το να μπορούμε να προβλέπουμε τέτοιου είδους στοιχεία με ακρίβεια είναι αρκετά δύσκολο καθώς το θέμα της συμπεριφοράς των μετακινούμενων αποτελεί ένα πολύ πολυπαραγοντικό πρόβλημα. Κατά την προσπάθεια να το αναλύσουμε μπορούν να εξαχθούν πολύ χρήσιμα συμπεράσματα μέσα από συγκεκριμένη ανάλυση με σκοπό τη βελτίωση των συστημάτων μεταφοράς ή την εξέταση αλλαγών.

Στη συνέχεια θα αναπτυχθεί η βιβλιογραφική ανασκόπηση στο 2ο κεφάλαιο, δηλαδή εργασίες που μελετήθηκαν σχετικά με το θέμα της εργασίας και κατηγοριοποιήθηκαν αναλόγως τη μεθοδολογία τους, τον τύπο των αποτελεσμάτων που μπορούσαν να εξάγουν κλπ και για τα μοτίβα δραστηριοτήτων και για τα durations. Στο 3ο κεφάλαιο θα παρουσιαστούν οι συμβολισμοί, τα σύνολα, οι μεταβλητές, οι παράμετροι αλλά και η συλλογιστική των durations models ώστε φτάσουν σε

πρόβλεψη διαρκειών.

Στο 4ο κεφάλαιο θα αναλυθεί η δομή και η λειτουργία του AthensPop πιο διεξοδικά, πώς συνθέτει πληθυσμό αλλά και πώς γεννάει τις μετακινήσεις μέσω του συνθετικού πληθυσμού αλλά και πως συνδέεται με τη λειτουργία των cox μοντέλων και πως λειτουργεί μετά την επέμβαση.

Στο 5ο κεφάλαιο καταγράφεται η εφαρμογή του μοντέλου με τις αλλαγές που έχουν γίνει καθώς και η ανάλυση τους μέσα από τα στατιστικά τεστ αλλά και τα διαγράμματα που εξάχθηκαν.

Τέλος, στο 6ο κεφάλαιο παρουσιάζονται τα συμπεράσματα που συγκεντρώνονται από την παραπάνω ανάλυση και γενικότερη ανασκόπηση της επέκτασης που έγινε, τι βοήθεια δίνει για μελλοντικές παρεμβάσεις και τι περιθώρια για την ανάπτυξη της έρευνας στον συγκεκριμένο τομέα.

2. Βιβλιογραφική Ανασκόπηση

2.1. Συλλογή Δεδομένων

Από τα παλαιότερα χρόνια οι πιο δημοφιλείς πηγές δεδομένων όπως προαναφέρθηκε ήταν οι έρευνες νοικοκυριού (Household travel surveys στη συγκεκριμένη περίπτωση). Ως διαδικασία οι ερωτηθέντες καλούσαν να απαντήσουν με λεπτομέρεια σχετικά με τις μετακινήσεις τους για κάποιο τυπικό διάστημα (μία μέρα, μία εβδομάδα κλπ.). Ως μέθοδος μπορεί να προσδώσει πολύ υψηλή ακρίβεια καθώς οι ερωτήσεις μπορούν να γίνουν πολύ αναλυτικές αλλά και το δείγμα μπορεί να είναι αρκετά μεγάλο. Επίσης, μία παρόμοια πηγή δεδομένων που χρησιμοποιείται ακόμη και σήμερα είναι οι εθνικές απογραφές. Πολύ σταθερές πηγές και οι δύο, το μοναδικό τους πρόβλημα είναι ότι γίνονται με χαμηλή συχνότητα διότι απαιτούν και πολύ χρόνο για να ολοκληρωθούν αλλά και πολύ μεγάλο εργατικό δυναμικό άρα και χρήματα. Βέβαια είναι ακόμη αναγκαίες καθώς είναι μια τεράστια βάση πάνω στην οποία προστίθενται πλέον νέα δεδομένα με πιο σύγχρονους τρόπους άντλησης.

Στην εποχή που η τεχνολογία έχει γίνει κομμάτι της ζωής μας και της

καθημερινότητας μας τα δεδομένα αυξάνονται διαρκώς. Σύμφωνα με το Canalys[4] στην Ευρώπη περίπου το 82% των ανθρώπων χρησιμοποιούν smartphones ενώ σύμφωνα με το PewResearch[5] 3 στους 4 χρήστες έχουν ανοιχτό το GPS στο κινητό τους αφού αποτελεί προϋπόθεση χρήσης κάποιων εφαρμογών. Παρόμοια είναι και τα ποσοστά στην Αμερική ενώ γενικότερα σε αναπτυσσόμενες χώρες μειώνονται αλλά παραμένουν στατιστικά σημαντικά τα ποσοστά των χρηστών smart devices. Αυτά τα δεδομένα τοποθεσίας αντλούνται και είναι διαθέσιμα προς εκμετάλλευση από κάθε εφαρμογή για ερευνητικούς σκοπούς. Η τεράστια διαφορά με τα προηγούμενα είναι πως τα δεδομένα υπάρχουν σε ζωντανό χρόνο, παρέχουν πολλές λεπτομέρειες και για την κίνηση των μετακινούμενων και για τις ώρες αιχμής και για την κίνηση στον δρόμο.

Ένας άλλος τύπος δεδομένων που προέρχεται πάλι από τα smartphones είναι τα δεδομένα από τα μέσα κοινωνικής δικτύωσης (Social media). Όπως αναφέρεται στο DataReportal[6] τα ποσοστά των χρηστών Social media είναι κοντά στο 80% των κατοίκων αναλόγως την εξεταζόμενη περιοχή. Από αυτούς τους χρήστες το 30% περίπου μοιράζει την τοποθεσία του μέσω των check in. Γίνεται σαφές πως δημιουργείται μία νέα μεγάλη πηγή δεδομένων προς αξιοποίηση καθώς θα μπορούσε να δίνει και ακόμη μεγαλύτερες λεπτομέρειες αναλόγως της δηλωθείσας τοποθεσίας για τον σκοπό μετακίνησης. Τα τελευταία χρόνια είναι αρκετές οι εργασίες που βασίζονται στην ανάλυση κυκλοφοριακών μεγεθών με βάση δεδομένα από social media (π.χ. Γκιοτσαλίτης 2015)

Στα λεγόμενα μεγάλα δεδομένα (big data) προστίθενται επίσης και τα δεδομένα που αντλούνται από τα δίκτυα κινητής τηλεφωνίας. Χρησιμοποιήθηκαν από τους Jingjun Li et al. (2023) σε μια προσπάθεια προσομοίωσης για τη σύνθεση ενός πληθυσμού και κατ' επέκταση τη δημιουργία ενός μοντέλου βασισμένου σε πράκτορες (ABM) αλλά και από τους C. Chen et al (2016). Οι τοποθεσίες συλλέγονται μέσω των κεραιών κινητής τηλεφωνίας που υπάρχουν σε ικανοποιητικό αριθμό ώστε να μπορούν να εξυπηρετούν τον σκοπό τους. Με τις κατάλληλες παραδοχές είναι κι αυτά πολύ χρήσιμα δεδομένα έτοιμα να χρησιμοποιηθούν ώστε να εξαχθούν πολύ σημαντικά συμπεράσματα για τα συστήματα μεταφορών και τη συμπεριφορά των μετακινουμένων.

[4] https://www.canalys.com/newsroom/europe-smartphone-market-Q1-2024

[5] https://www.pewresearch.org/internet/2012/05/11/three-quarters-of-smartphone-owners-use-location-based-services/

[6] https://datareportal.com/reports/digital-2024-deep-dive-5-billion-social-media-users

Νέου είδους στοιχεία που μπορούμε να εκμεταλλευτούμε είναι ακόμα δεδομένα από τα έξυπνα πάσα (smart cards) που χρησιμοποιούν τα Μέσα Μαζικής Μεταφοράς όπως χρησιμοποιήθηκε από τους M.A. Munizaga, C. Palma (2012) στην προσπάθεια να δημιουργήσουν έναν πίνακα O-D. Πάλι με συγκεκριμένες παραδοχές μπορούν να γίνουν καλές εκτιμήσεις και να εξαχθούν πληροφορίες σχετικά με τα χαρακτηριστικά των μετακινήσεων.

Τέλος, στοιχεία αντλούνται από φωρατές, δηλαδή αισθητήρες τοποθετημένους σε σταθερά σημεία στο οδόστρωμα τα οποία μπορούν να αναγνωρίσουν διάφορα χαρακτηριστικά μετακίνησης όπως τον τύπο οχήματος, την ταχύτητα και μπορούν να βοηθήσουν σε ασφαλέστερες εκτιμήσεις για την ροή των οχημάτων.

2.2 Τρόποι Μοντελοποίησης

2.2.1. Μοντελοποίηση μοτίβων δραστηριοτήτων

Το δεύτερο σημαντικό σκέλος που εξετάστηκε και μελετήθηκε κατά τη βιβλιογραφική ανασκόπηση αφορά τους διαφορετικούς τρόπους μοντελοποίησης. Το δύσκολο έργο της μοντελοποίησης κι αυτό που προσπαθούμε είναι η πρόβλεψη των μετακινήσεων. Όσον αφορά δε τη δημιουργία μοτίβων δραστηριοτήτων (travel activity patterns) εκεί η πρόβλεψη γίνεται ακόμη πιο δύσκολη καθώς η συμπεριφορά των μετακινουμένων είναι πολύπλοκη και επηρεάζεται από πολλές διαφορετικές μεταβλητές. Οι διαφορετικές μέθοδοι μοντελοποίησης δίνουν ευελιξία στην έρευνα καθώς αναλόγως τον τύπο των δεδομένων και γενικότερα την περίπτωση εφαρμόζουν καλύτερα και μπορούν να πλησιάσουν με μεγαλύτερη ακρίβεια την πραγματικότητα.

Παραδοσιακά χρησιμοποιείται το μοντέλο των 4 βημάτων (Trip Generation – Trip Distribution – Mode choice – Trip Assignment). Όπως αναλύθηκε και από τον Ahmed, B. (2012) για την πόλη της Ντάκα γίνεται κατανοητό πως οι παραδοχές είναι πιο απλοϊκές όσον αφορά ίσως αυξητικούς συντελεστές με σκοπό την πρόβλεψη και η δημιουργία μοτίβων δραστηριοτήτων είναι δύσκολο να εκτιμηθεί καθώς το είδος των δεδομένων δεν περιέχει τόση λεπτομέρεια όση θα χρειαζόταν.

Έτσι, με την ανάπτυξη της τεχνολογίας εφαρμόζονται πλέον νεότεροι τρόποι δημιουργίας μοντέλων που μπορούν να αφομοιώσουν περισσότερα δεδομένα και να

εξάγουν πολύπλευρα συμπεράσματα. Ένας τέτοιος τρόπος είναι τα agent-based μοντέλα. Τα συγκεκριμένα μοντέλα επικεντρώνονται στη δημιουργία ενός συγκεκριμένου περιβάλλοντος με συγκεκριμένους κανόνες στα οποία ερευνάται η συμπεριφορά των πρακτόρων (ανθρώπων). Στην πλειοψηφία των περιπτώσεων συντίθεται ένας πληθυσμός βάσει δημογραφικών και κοινωνικοοικονομικών στοιχείων που αντλούνται από απογραφές και έρευνες, ο οποίος στόχος είναι να αντιπροσωπεύει όσο το δυνατόν καλύτερα τον πραγματικό πληθυσμό. Όσο πιο πολύ οι πράκτορες εκπροσωπούν τον πραγματικό πληθυσμό τόσο καλύτερο το μοντέλο και οι αποφάσεις που θα εξαχθούν από αυτό. Τέτοιες ερευνητικές προσπάθειες αναλύθηκαν και στη συγκεκριμένη ανασκόπηση, όπως των Ç. Tozluoglu, S. Dhamal and S. Yeh et al. (2023), των Dominik Ziemke et al. (2019), των Sebastian Hörl & Milos Balac (2021), των B.Y. He et al. (2021), των (Dominik Ziemke et al. 2019) καθώς και των Jingjun Li et al. (2023). Κάθε εργασία από αυτές στοχεύει και σε κάτι διαφορετικό ως προς την ανάλυση της. Για παράδειγμα, στο Παρίσι η ανάλυση γίνεται πάνω στον έλεγχο του τύπου των μετακινήσεων με στόχο τη δημιουργία activity travel patterns. Αντίθετα σε εργασίες όπως αυτή στις Βρυξέλλες η ανάλυση επικεντρώνεται στην επιλογή του μέσου και στον ρυθμό συμμετοχής των agents σε δραστηριότητες. Στη Νέα Υόρκη δε, η ανάλυση εκτός από την επιλογή του μέσου υπάρχει και χωρική αλλά και χρονική ανάλυση των μετακινήσεων με απώτερο σκοπό την εξέταση συγκεκριμένων πολιτικών όπως διόδια, που άλλωστε είναι και ο στόχος όλων αυτών των ερευνών. Τέλος, συναντούμε και τη χρήση Bayesian Networks και για τη σύνθεση του πληθυσμού στον J.W. Joubert (2018). Πιο διεξοδικά, στη συγκεκριμένη μελέτη αναλύεται η επεξεργασία των δεδομένων (απογραφή και δημογραφικά, κοιν/οικ.) για τη σύνθεση πληθυσμού ο οποίος προορίζεται για προσομοίωση μέσω MATSim (Agent-based modelling) ενώ τα Bayesian Networks χρησιμοποιούνται κατά κύριο λόγο βοηθητικά ως προς τη βελτίωση της συσχέτισης των χαρακτηριστικών και την αποτύπωση τους στον συνθετικό πληθυσμό.

Αρκετά σημαντική συσχέτιση με τις μετακινήσεις έχει παρατηρηθεί επίσης και για τις χρήσεις γης. Γνωρίζοντας καλά το πώς χρησιμοποιείται κάθε περιοχή (π.χ. κατοικήσιμη συνοικία, βιομηχανική συνοικία κλπ.) μπορούμε να δημιουργήσουμε σημεία ενδιαφέροντος γύρω από τα οποία να κατευθύνουμε και τις μετακινήσεις. Με αυτήν τη λογική αναπτύχθηκαν τα LUTI μοντέλα (Land use transportation). Το συγκεκριμένο είδος μοντέλων καταγράφει τη δυναμική αλληλεπίδραση μεταξύ χρήσης γης και μεταφορών, λαμβάνοντας υπόψη πώς οι αλλαγές στη χρήση γης επηρεάζουν

τη ζήτηση μεταφορών και πώς οι μεταφορές επηρεάζουν την ανάπτυξη και την κατανομή της χρήσης γης. Ακόμη, επιτρέπει την προσομοίωση διαφόρων σεναρίων πολιτικής και σχεδιασμού, όπως η κατασκευή νέων υποδομών μεταφορών ή οι αλλαγές στις ζώνες χρήσης γης. Τέτοιου είδους εργασίες αναλύθηκαν από τους M. Aljoufie et al. (2013) αλλά και των Y. Wang et al. (2015).
σε μια προσπάθεια να ποσοτικοποιήσουν τη διαθεσιμότητα στις μετακινήσεις βάσει των χρήσεων γης με διαφορετικές παραμέτρους σε κάθε εργασία και άλλη στόχευση. Ακόμα στην εργασία των Luis A. Guzman et al. (2017) για τη Μπογκοτά η ανάλυση στοχεύει στη δημιουργία μοτίβων δραστηριοτήτων και πιο συγκεκριμένα προσπαθεί να προβλέψει τα HWH & HOH tours (Home-Work-Home & Home-Other-Home).

Σημαντική προσέγγιση θεωρείται επίσης αυτή της μηχανικής εκμάθησης (machine learning). Κατά τη βιβλιογραφική ανασκόπηση αναλύθηκαν οι εργασίες του Γκιοτσαλίτη (2020) όπου η μηχανική εκμάθηση επιτυγχάνεται με Kullback-Leibler Divergence και τον αλγόριθμο Naïve Bayes καθώς και των Aurore Sallard et al. (2023) που χρησιμοποιούν τα Bayesian δίκτυα. Κατά αυτήν τη διαδικασία τα δεδομένα διαχωρίζονται σε training και testing κατά τέτοιον τρόπο ώστε ένα μέρος τους να εκπαιδεύεται και να παρέχει ακριβέστερες προβλέψεις βάσει στατιστικών μεγεθών ή αλγορίθμων. Τα μεγέθη που εξετάζονται στην ανάλυση αφορούν πάλι τα χαρακτηριστικά των μετακινήσεων όπως σκοπός μετακινήσεων, επιλογή μέσου, ώρες αλλά δύναται να γίνει και στατιστική ανάλυση.

Σε κάποιες περιπτώσεις επίσης καταγράφηκαν συνδυασμοί διαφορετικών τρόπων μοντελοποίησης. Πιο αναλυτικά, στην έρευνα των B.Y. He et al. (2020) συνδυάστηκαν tour based & trip based μοντέλα για την εύρεση της επιλογής μέσου. Επίσης εξάχθηκαν συμπεράσματα για κάθε σενάριο ξεχωριστά για τη χρονική και χωρική κατανομή των μετακινήσεων. Τα tour based μοντέλα θα μπορούσαμε να πούμε ότι ανήκουν στην κατηγορία των 4 βημάτων ενώ τα tour based Σε αντίθεση με τα Trip-Based Models, τα οποία αναλύουν μεμονωμένα ταξίδια, λαμβάνουν υπόψη τη συσχέτιση μεταξύ των ταξιδιών κατά τη διάρκεια της ημέρας. Με παρόμοιο τρόπο λειτουργεί και το μοντέλο των G. Jovicic, C.O. Hansen (2003) που θεωρείται ένα από τα μοντέλα με τη μεγαλύτερη επιτυχία καθώς κατάφερε να προβλέψει τη μελλοντική ζήτηση της κυκλοφορίας με μεγάλη επιτυχία. Εκείνο βέβαια, έχει μια προσέγγιση περισσότερο προς το μοντέλο 4 βημάτων και κυρίως υπολογίζει αριθμό μετακινήσεων. Ένα άλλο μοντέλο που μελετήθηκε αφορά την κατηγορία των Spatial – Temporal Interaction Models και πιο συγκεκριμένα είναι ένα spatial statistic model το οποίο

προσομοιώνει την αλληλεπίδραση ταξιδιών μεταξύ περιοχών σε διάφορες χρονικές στιγμές. Επίσης συνδυάζει τη χωροχρονική διάσταση με σκοπό την κατανόηση πώς οι μεταφορές εξελίσσονται με την πάροδο του χρόνου. Yixiao Li et al. (2019) γίνεται η προσπάθεια να αναλυθούν activity travel patterns χωρικά και χρονικά καθώς και να συσχετιστούν διάφοροι παράγοντες που επηρεάζουν τη συμπεριφορά των μετακινούμενων. Το μοντέλο χρησιμοποιεί μέση τιμή, τυπική απόκλιση, Three Modes Gaussian Function, densel kensity model και spillover commuting για τη δημιουργία των activity travel patterns.

2.2.2. Μοντελοποίηση διάρκειας δραστηριοτήτων

Όσον αφορά την εκτίμηση διάρκειας δραστηριοτήτων η βιβλιογραφική μελέτη στήθηκε κυρίως γύρω από μοντέλα κινδύνου (hazard based) αφού αυτή είναι η πιο διαδεδομένη μέθοδος υπολογισμού διαρκειών. Στο εξής η αναφορά σε μοντέλα κινδύνου θα γίνεται με τον όρο hazard based μοντέλα. Τα hazard based μοντέλα υπολογίζουν την πιθανότητα να τερματίσει ένα γεγονός (στη συγκεκριμένη περίπτωση τα activities που μελετάμε) ανά διαφορετική χρονική στιγμή t. Αρχικά εξετάστηκαν παραμετρικά hazard based μοντέλα. Πιο αναλυτικά, πρόκειται για μοντέλα τα οποία η συνάρτηση του κινδύνου να διακοπεί ένα activity περιγράφεται από την καμπύλη μιας κατανομής η οποία έχει προεπιλεχθεί από τον χρήστη (συνήθως Weibull ή Exponential). Κάποια από αυτά ήταν όπως των Sreela P.Ka et al. (2013) το οποίο σχεδιάστηκε αμιγώς για shopping durations και εστιάζει στην επιρροή των μεταβλητών σε αυτά. Ακόμα, ο Chandra R. Bhat (1996) και ο J. Urban Plann (1998), εστιάζουν σε shopping durations κυρίως και κατά δεύτερον λόγο σε άλλα purposes και προσπαθούν να εκτιμήσου εάν υπάρχουν παράγοντες μη παρατηρήσιμοι από τα υπάρχοντα δεδομένα που επηρεάζουν τα αποτελέσματα και πως πρέπει εκεί να ληφθούν υπόψιν κατάλληλα. Τέλος, ένα ακόμα παραμετρικό hazard based μοντέλο είναι του Annesha Enam et al. (2020). Στη συγκεκριμένη προσπάθεια αντλούνται δεδομένα από GPS και γίνεται η προσπάθεια συσχέτισης του purpose τους βάσει αυτών και στη συνέχεια του duration. Μένοντας στα hazard based μοντέλα εξετάστηκαν επίσης ημί – παραμετρικά. Η βασική διαφορά τους με τα παραμετρικά είναι ότι στα συγκεκριμένα δεν υποτίθεται η ύπαρξη κάποιας κατανομής για να περιγράφει τη συνάρτηση κινδύνου. Η βασική επικινδυνότητα παραμένει απροσδιόριστη και εκτιμάται μέσα από τα δεδομένα

(ατομικά χαρακτηριστικά). Τέτοια μοντέλα ήταν των Chunguang Liu et al. (2023) ο οποίος προσπάθησε να καλύψει ένα κενό στη βιβλιογραφία αναλύοντας τη διάρκεια δραστηριοτήτων σε επείγουσες καταστάσεις (Covid 19) χρησιμοποιώντας κοιν/οικ. ΄και δημογραφικά χαρακτηριστικά αλλά και τις χρήσεις γης ως μεταβλητή έντονης επιρροής των διαρκειών δραστηριοτήτων. Επίσης cox ημί παραμετρικό μοντέλο είναι και του J.L.Yee (2000) ο οποίος εργάζεται πάνω σε 5ετή έρευνα δίνοντας μια νέα οπτική στη δυναμική του χρόνου στις διάρκειες δραστηριοτήτων. Σε καμία από αυτές τις προσπάθειες βέβαια δεν εξετάστηκε η επιρροή της μη παρατηρήσιμης ετερογένειας (unobserved heterogeneity). Ακόμα τελευταίας μορφής hazard based που μελετήθηκε είναι του Pauline van den Berg et al.(2012). Στη συγκεκριμένη εργασία χρησιμοποιείται ένα accelerated hazard model το οποίο χρησιμοποιείται έτσι ώστε να εκτιμηθούν οι συσχετίσεις μεταξύ των social activities duration και των σχέσεων μεταξύ των ανθρώπων που συμμετέχουν καθώς και λεπτομέρειες σχετικά με τη συχνότητα αυτών των δραστηριοτήτων ή τις συνθήκες που κανονίστηκαν. Η κύρια διαφορά είναι πως με τη συγκεκριμένη μορφή του μοντέλου οι μεταβλητές επηρεάζουν άμεσα τον χρόνο διάρκειας μιας δραστηριότητας και όχι όπως στα προηγούμενα όπου επηρεαζόταν έμμεσα μέσα από τους συντελεστές των μοντέλων που επηρέαζαν τον κίνδυνο πρωτίστως.

Βέβαια, εκτός από τα hazard based μοντέλα μελετήθηκαν επίσης ένα discrete continuous choice μοντέλο του Y.-L. Chu (2019). Πιο αναλυτικά, στη συγκεκριμένη εργασία μοντελοποιούνται ταυτόχρονα δύο επίπεδα επιλογής. Το διακριτό που αφορά τη συμμετοχή σε μια maintenance activity καθώς και τη συνεχή μεταβλητή που δε θα μπορούσε να είναι άλλη από τη διάρκεια της δραστηριότητας. Έτσι δίνεται έμφαση στην αλληλεξάρτηση αυτών των δύο μερών του μοντέλου.

Τέλος όσον αφορά τις διάρκειες δραστηριοτήτων, μία ακόμη εναλλακτική προσέγγιση που εξετάστηκε αφορά τη δουλειά των Nebiyou Tilahun και David Levinson (2009) όπου μέσω ενός path model έδωσαν έμφαση στις αιτιακές σχέσεις μεταξύ των μεταβλητών με σκοπό να συμπεριλάβουν τόσο άμεσες όσο και έμμεσες επιδράσεις στις διάρκειες δραστηριοτήτων.

Πίνακας 1 . Βιβλιογραφική Ανασκόπηση

Επιστημονική Εργασία (έτος)	**Μεταβλητές Απόφασης**	**Στόχος**	**Τύπος Μοντέλου**	**Μέθοδος Επίλυσης**
Sebastian Hörl & Milos Balac (2021)	Χαρ/κά μετακίνησης	Δημιουργία συνθετικής ζήτησης μετακίνησης με open data με σκοπό τη διευκόλυνση επανάληψης μεθοδολογίας	Διακριτό μη Γραμμικό Μοντέλο (Agent based model)	Ευρετικός τρόπος
Bayes Ahmed (2012)	Πληθυσμός, Εισόδημα, Τιμή γης, Ανεργία	Πρόβλεψη O-D πινάκων σε 10 χρόνια	Γραμμικό Μοντέλο	Ακριβής Μέθοδος Επίλυσης
Brian Y., et al. (2013)	Επιλογή μέσου, κόστος διαδρομής, διάρκεια διαδρομής	Χωρική και χρονική κατανομή μετακινήσεων και επιλογή μέσου βάσει συγκεκριμένων σεναρίων	Διακριτό μη Γραμμικό Μοντέλο (Tour based model)	Ευρετικός τρόπος
Mohammed Aljoufie et al. (2013)	Χρήσεις γης, διαθεσιμότητα, κόστος διαδρομής	O-D πίνακες και εκτίμηση προσβασιμότητας	Διακριτό μη Γραμμικό Μοντέλο (LUTI model)	Μετα-ευρετικός τρόπος
Brian Yueshuai et. (2014)	Επιλογή μέσου, χωρική κατανομή	Εξέταση σεναρίων πολιτικής διοδίων βάσει διαφορετικών τιμολογιακών πολιτικών	Διακριτό μη Γραμμικό Μοντέλο (Agent based model)	Ευρετικός τρόπος
Luis A. Guzman, Ana M. Gomez, Carlos Rivera (2017)	Εισόδημα νοικοκυριού, πληθυσμός, χρήσεις γης, ιδιοκτησία αυτοκινήτου	Πρόβλεψη ζήτησης μετακινήσεων για HWH- HOH ταξίδια	Διακριτό μη Γραμμικό Μοντέλο (LUTl model)	Μετα-ευρετικός τρόπος
Marcela A. Munizaga, Carolina Palma (2012)	Τοποθεσία & διάρκεια μεταξύ διαδοχικών πληρωμών	Καταγραφή O-D πινάκων μέσα από δεδομένα από MMM	Διακριτό μη Γραμμικό Μοντέλο (Trip based model)	Ευρετικός τρόπος
Goran Jovicic, Christian Overgaard Hansen (2003)	Πληθυσμός, χρήσεις γης, ιδιοκτησία αυτοκινήτου,	Εκτίμηση αριθμού μετακινήσεων και σκοπό διαδρομών για την εξέταση πολιτικής διοδίων	Διακριτό μη Γραμμικό Μοντέλο (Tour based model)	Ευρετικός τρόπος

Jingju et al. (2023)	Πληθυσμός, Τοποθεσία, Ώρα έναρξης διαδρομών	Εκτίμηση αλυσιδών δραστηριοτήτων μέσω προσομοίωσης από ABM	Διακριτό μη Γραμμικό Μοντέλο (Agent Based model)	Ευρετικός τρόπος
Dr. Konstantinos Gkiotsalitis and Dr. Antony Stathopoulos (2015)	Ώρα μετακίνησης, τύπος διαδρομής	Πρόβλεψη διανυθεισών αποστάσεων μετακινήσεων και μοτίβων μετακίνησης	Διακριτό μη Γραμμικό Μοντέλο (machine learning)	Μετα-ευρετικός τρόπος
Yixiao Li, Zhaoxin Dai Lining Zhu and Xiaoli Liu (2019)	Τοποθεσία, ώρες έναρξης και ταχύτητα μετακίνησης	Δημιουργία μοτίβων δραστηριοτήτων μετακινήσεων	Διακριτό μη Γραμμικό Μοντέλο (spatial statistic model)	Μετα-ευρετικός τρόπος
Azalden Alsger et al. (2011)	Χωρικές, χρονικές μεταβλητές, κοιν/οικ. χαρ/κά, χρήσεις γης	Εκτίμηση σκοπών μετακινήσεων	Διακριτό μη Γραμμικό Μοντέλο (rule based model)	Ευρετικός τρόπος
ÇaglarTozluoglu et al. (2023)	Κοιν/οικ. & χωρικά χαρ/κά	Εκτίμηση μοτίβων δραστηριοτήτων μετακινήσεων και χωροχρονική κατανομή αυτών	Διακριτό μη Γραμμικό Μοντέλο (rule based model)	Ευρετικός τρόπος
Aurore Sallarda, Milos Balac (2023)	Κοιν/οικ. χαρ/κά & σκοποί μετακίνησης	Καταγραφή μοτίβων δραστηριοτήτων μετακινήσεων	Διακριτό μη Γραμμικό Μοντέλο (machine learning)	Μετα-Ευρετικός τρόπος
Chunguang Liu, et al. (2023)	Κοιν/οικ. χαρ/κά, χρήσεις γης, αποστάσεις μετακίνησης	Συσχετίσεις μεταξύ διάρκειας δραστηριοτήτων και μεταβλητών	Semi parametric model (cox)	Ακριβής μέθοδος επίλυσης
Jeffrey P.Kharoufeh,Konstadinos G.Goulias	Κοιν/οικ. χαρ/κά / φύλο	Συσχετίσεις μεταξύ διάρκειας δραστηριοτήτων και μεταβλητών, έμφαση στην οικογενειακή κατάσταση	Μη γραμμικό (Kernel density estimator)	Ακριβής μέθοδος επίλυσης
N.Golshani et al.	Κοιν/οικ. χαρ/κά & ώρες έναρξης ταξιδιών	Συσχετίσεις μεταξύ διάρκειας δραστηριοτήτων και μεταβλητών, ανά ξεχωριστή ενοποίηση με ώρες έναρξης	Μη γραμμικό (copula joint based model)	Ακριβής μέθοδος επίλυσης

Sreela P. et al. (2013)	Κοιν/οικ. χαρ/κά & χαρ/κά μετακίνησης	Συσχετίσεις μεταξύ διάρκειας δραστηριοτήτων και μεταβλητών	Μη γραμμικό (Weibull parametric)	Ακριβής μέθοδος επίλυσης
Pauline Van den Berg et al. (2012)	Κοιν/οικ. χαρ/κά & χαρ/κά μετακίνησης	Συσχετίσεις μεταξύ διάρκειας δραστηριοτήτων και μεταβλητών ανά latent class	Μη γραμμικό (Weibull parametric)	Ακριβής μέθοδος επίλυσης
Y.-L. Chu (2019)	Κοιν/οικ. χαρ/κά & χαρ/κά μετακίνησης	Συσχετίσεις μεταξύ διάρκειας δραστηριοτήτων και μεταβλητών ανά χρονική περίοδο και ανά σενάριο	Διακριτό μη γραμμικό (discrete choice model)	Ακριβής μέθοδος επίλυσης
Nebiyou Tilahun, David Levinson (2009)	Κοιν/οικ. χαρ/κά & χαρ/κά μετακίνησης	Συσχετίσεις μεταξύ διάρκειας δραστηριοτήτων και μεταβλητών	Μη γραμμικό (path model)	Ακριβής μέθοδος επίλυσης
Chandra R. Bhat (1996)	Κοιν/οικ. χαρ/κά & χαρ/κά μετακίνησης	Συσχετίσεις μεταξύ διάρκειας δραστηριοτήτων και μεταβλητών & επιρροή ετερογένειας	Μη γραμμικό (Weibull parametric vs non parametric)	Ακριβής μέθοδος επίλυσης
Annesha Enam et al. (2020)	Κοιν/οικ. οιχαρ/κά & χαρ/κά μετακίνησης	Εύρεση σκοπών μετακίνησης και συσχετίσεις μεταξύ διάρκειας δραστηριοτήτων και μεταβλητών ανά σκοπό μετακίνησης	Μη γραμμικό (Weibull parametric hazard based)	Ακριβής μέθοδος επίλυσης
J. Urban Plann (1998)	Κοιν/οικ. χαρ/κά & χαρ/κά μετακίνησης	Συσχετίσεις μεταξύ διάρκειας δραστηριοτήτων και μεταβλητών ανά σκοπό μετακίνησης με ετερογένεια	Μη γραμμικό (Weibull parametric hazard)	Ακριβής μέθοδος επίλυσης
J.L. Yee, D.A. Niemeier (2000)	Κοιν/οικ. χαρ/κά & χαρ/κά μετακίνησης	Συσχετίσεις μεταξύ μεταβλητών και διάρκειας δραστηριοτήτων ανά σκοπό μετακίνησης	Μη γραμμικό (Cox parametric hazard based)	Ακριβής μέθοδος επίλυσης

3.Μεθοδολογία

3.1 Πρότυπο μοντέλο

Στο πρακτικό μέρος της εργασίας θα χρησιμοποιηθεί ένα hazard based μοντέλο έτσι ώστε να προβλέπει διάρκειες δραστηριοτήτων (activity durations) σε περιπτώσεις που χρειάζεται.

Ως πρότυπο θα χρησιμοποιηθεί το μοντέλο των J.L. Yee & D.A. Niemeier (2000), δηλαδή ένα ημιπαραμετρικό cox hazard based μοντέλο. Κατά τη λειτουργία του cox ισχύει η παρακάτω συνάρτηση η οποία εκφράζει τον κίνδυνο να τερματίσει μία δραστηριότητα:

$$h_i(t) = ho(t) * \exp(\beta_c^T X_i)$$

Η βασική ιδιότητα του Cox είναι πως δεν προαποφασίζεται κάποια κατανομή να εκφράζει τη συνάρτηση κινδύνου h(t) όπως στα παραμετρικά.

3.2 Μαθηματικό Μοντέλο

3.2.1. Συμβολισμοί πρότυπου μοντέλου

Ως βάση θα αναλυθεί το μοντέλο του J.L.Yee (2000). Στη συγκεκριμένη μελέτη αναλύθηκε ένα απλό Cox μοντέλο το οποίο καλούταν να εκτιμήσει διάρκειες μετακινήσεων για διάφορους σκοπούς μετακίνησης. Η συνάρτηση που το εκφράζει είναι:

$$h_i(t) = ho(t) * \exp(\beta^T X_i)$$

Όπου,

- ho(t) η βασική συνάρτηση κινδύνου
- $\beta^T X$ το εσωτερικό γινόμενο των συντελεστών β και των χαρακτηριστικών X για κάθε άτομο i

Εν συνεχεία η συνάρτηση επιβίωσης υπολογίζεται από τη σχέση

$$S(t) = S_0(t)^{\exp(\beta^T X_i)}$$

Όπου,

- So(t) η βασική συνάρτηση επιβίωσης που υπολογίζεται από τη σχέση

$S_o(t) = \exp(-H_o(t))$ που $H_o(t) = \int_0^t h_o(u)du$ και

- $\beta^T X$ το εσωτερικό γινόμενο των συντελεστών β και των χαρακτηριστικών X για κάθε άτομο i

Για τον υπολογισμό των διαρκειών επιλέγουμε συνήθως το median duration δηλαδή την τιμή του t που επιβεβαιώνει τη συνάρτηση S(t) = 0.5 και εκφράζει τη χρονική στιγμή t που το 50% έχει λήξει τη δραστηριότητα ενώ το άλλο 50% τη συνεχίζει.

$$t_{med} = S_o^{-1}(0.5) * \exp(-\beta^T X)$$

Πίνακας 2. Συμβολισμοί βασικού μοντέλου

Σύνολα	
K	Το σύνολο των χαρακτηριστικών που χρησιμοποιούνται ως προβλεπτικοί παράγοντες
I	Το σύνολο των παρατηρήσεων
Παράμετροι	
β_κ	Συντελεστές για κάθε μεταβλητή X_κ
$ho(t)$	Βασική συνάρτηση κινδύνου
Μεταβλητές	
$activity\ dur_i$	Η διάρκεια δραστηριότητας (survival time) για κάθε παρατήρηση i.
$X_{i,k}$	Το διάνυσμα των προβλεπτικών μεταβλητών για κάθε παρατήρηση
$status_i$	Η μεταβλητή λογοκρισίας (1 = το γεγονός συνέβη, 0 = censored).

Όπου X_i από τα δεδομένα της έρευνας έχουμε:

- Ηλικία
- Ζώνη Κατοικίας
- Ζώνη Άφιξης
- Ζώνη Αναχώρηση;

- Μέσο
- Εισόδημα
- Ιδιοκτησία Οχήματος
- Επαγγελματική κατάσταση
- Εκπαίδευση
- Φύλο
- Ώρα Έναρξης Μετακίνησης

3.2.2. Επέκταση Μοντέλου και μαθηματικοί συμβολισμοί

Για την επέκταση του μοντέλου αποφασίστηκε η προσθήκη ετερογένειας. Πιο αναλυτικά, στο προτεινόμενο μοντέλο προστίθενται δύο μεταβλητές ($frailty_1$ & $frailty_2$) που αντιπροσωπεύουν δύο διαφορετικές λανθάνουσες κατηγορίες (Από εδώ και στο εξής θα αναφέρονται ως latent classes). Οι latent classes είναι δύο υποομάδες των παρατηρήσεων οι οποίες δεν παρατηρούνται άμεσα αλλά υπολογίζονται μέσω στατιστικών μοντέλων. Κατά τη διαδικασία υπολογισμού των διαρκειών τα δεδομένα χωρίζονται σε 2 διαφορετικές ομάδες βάσει των χαρακτηριστικών τους. Έτσι, κάθε μία από αυτές τις ομάδες φανερώνει ένα κρυφό μοτίβο το οποίο ναι μεν δεν περιγραφόταν από τα δεδομένα και τις μεταβλητές που αντλούνταν από αυτά, είναι ικανό δε να επηρεάσει την τελική διάρκεια των δραστηριοτήτων. Κάτι τέτοιο για παράδειγμα θα μπορούσε να είναι μετακινήσεις ατόμων που έχουν μικρά παιδιά ή μετακινήσεις ατόμων που αποφεύγουν την κίνηση κλπ.) Αφορά δηλαδή κρυφά χαρακτηριστικά ή συμπεριφορές οι οποίες είναι πράγματι κάτι το οποίο μπορεί να επηρεάσει τις διάρκειες. Κάθε παρατήρηση μπορεί να ανήκει είτε στη μία ομάδα είτε στην άλλη βάσει και των ατομικών χαρακτηριστικών που συγκεντρώνει αλλά και των ατομικών χαρακτηριστικών που συγκεντρώνουν οι υπόλοιπες και αυτό υπολογίζεται με πιθανότητες όπως θα αναλυθεί και στη συνέχεια.

Το μοντέλο βασίζεται στην εκτίμηση ενός Cox μοντέλου με frailty, χρησιμοποιώντας μη παραμετρική βασική επικινδυνότητα.

$$h_i(t) = ho(t) \sum_{c=1}^{2} Z_{ic} * \exp\left(\beta_c^T X_{iK} + frailty_c\right)$$

Πίνακας 3. Συμβολισμοί μοντέλου μετά την επέκταση

Σύνολα	
K	Το σύνολο των χαρακτηριστικών που χρησιμοποιούνται ως προβλεπτικοί παράγοντες
C	Το σύνολο των λανθανουσών κατηγοριών
I	Το σύνολο των παρατηρήσεων
Παράμετροι	
$\beta_{c,\kappa}$	Συντελεστές για κάθε χαρακτηριστικό k και λανθάνουσα κατηγορία c.
$z_{frailty}$	Οι λανθάνουσες μεταβλητές frailty για κάθε κατηγορία.
$\sigma_{frailty}$	Η τυπική απόκλιση των frailty terms.
$class_{probs}$	Οι πιθανότητες συμμετοχής σε κάθε λανθάνουσα κατηγορία.
Μεταβλητές	
$activity\ dur_i$	Η διάρκεια δραστηριότητας (survival time) για κάθε παρατήρηση i.
$X_{i,k}$	Το διάνυσμα των προβλεπτικών μεταβλητών για κάθε παρατήρηση
$status_i$	Η μεταβλητή λογοκρισίας (1 = το γεγονός συνέβη, 0 = censored).

Όπου X_i από τα δεδομένα της έρευνας όπως πριν:

- $ho(t)$: Η μη παραμετρική βασική επικινδυνότητα.
- Z_{ic} : Η πιθανότητα του ατόμου i να ανήκει στη λανθάνουσα κατηγορία c.
- $frailty_c$: Ο όρος frailty για την κατηγορία c που υπολογίζεται από τη σχέση

$$frailty_c = z_{frailty_c} * \sigma_{frailty}$$

Όπου,

- $z_{frailty_c}$ = είναι μη κεντραρισμένος τυχαίος παράγοντας από 0 έως 1
- $\sigma_{frailty}$ = τυπική απόκλιση

Πιθανότητες ένταξης σε λανθάνουσα κατηγορία (Latent class): Για τον διαχωρισμό

των παρατηρήσεων στις 2 latent classes χρησιμοποιείται η πιθανότητα Z_{ic} [7]η οποία εκφράζει την εκ των υστέρων (a posteriori) πιθανότητα της παρατήρησης *i* να ανήκει στη latent class *c* όπου το $class_{probs}$ προέρχεται από μια Dirichlet εκ των προτέρων (a priori) κατανομή και πρέπει να ισχύει

$$\sum_{c \in c} classprobs_c = 1, classprobs_c \geq 0$$

Και το Z_{ic} υπολογίζεται από τη σχέση

$$Z_{ic} = classprobs_c * likelihood_c * \frac{1}{\sum_{c'} classprobs_{c'} * likelihood_{c'}}$$

Όπου, $classprobs_c$ είναι η αρχική πιθανότητα ένταξης στη latent class *c*, $likelihood_c$ εκφράζει το πόσο καλά ταιριάζουν τα δεδομένα με τις παραμέτρους της latent class *c* και περιλαμβάνει δύο περιπτώσεις.

$$likelihood_c = \begin{cases} Weibull\ PDF, & status_i = 1 \\ Weibull\ CCDF, & status_i = 0 \end{cases}$$

Και

$$Weibull\ PDF(t) = \frac{k}{\lambda} * \left(\frac{t}{\lambda}\right)^{\kappa - 1} \exp\left(-\frac{t}{\lambda}\right)^{\kappa},$$

όπου k, λ είναι οι παράμετροι της Weibull και ισούνται με:

$$\mathrm{k} = \frac{1}{\sigma_{frailty}} \quad \& \lambda = \exp\left(X_i * \beta_c + frailty_c\right)$$

Αντίστοιχα, $Weibull\ CCDF(t) = \exp\left(-\frac{t}{\lambda}\right)^{\kappa}$ με τις ίδιες παραμέτρους.

Τέλος, ο όρος $\sum_{c'} classprobs_{c'} * likelihood_{c'}$ αφορά την κανονικοποίηση των πιθανοτήτων ώστε το άθροισμα τους να ισούται με 1.

Αξίζει να σημειωθεί η μεγάλη διαφορά με το απλό cox, κι αυτή έγκειται στη δημιουργία του frailty μοντέλου και στη συγχώνευσή του με το απλό cox ώστε να υπάρχουν ενιαίοι συντελεστές για τις μεταβλητές και τους frailty όρους. Στη συνέχεια του κειμένου όπου λανθάνουσα κατηγορία θα αναφέρεται ως latent class.

[7] https://discourse.mc-stan.org/t/using-the-posterior-predictive-distribution-of-my-model-to-estimate-the-probability-that-a-future-observation-is-greater-than-15/4245?utm_source=chatgpt.com

Συνάρτηση επιβίωσης για κάθε latent class c:

$$S_{i,c}(t) = \exp\left(-\int_0^t h_o(u)\exp(X_i\beta_c + frailty_c)\,du\right)$$

Όπου,

- $ho(t)$: Η μη παραμετρική βασική επικινδυνότητα
- $X_i\beta_c$: : Το εσωτερικό γινόμενο των συντελεστών β_c και των μεταβλητών X για κάθε άτομο i
- $frailty_c$: ο όρος για κάθε latent class c

Συνάρτηση Συνολικής Επιβίωσης

$$S_i(t) = \pi_1 * S_{i_1}(t) + \pi_2 S_{i_2}(t)$$

Όπου,

- $\pi_1 = Z_{i1}$ (πιθανότητες ένταξης στην latent class 1)
- $\pi_2 = Z_{i2}$ (πιθανότητες ένταξης στην latent class 2)

Τελευταίο βήμα για τις median durations, ψάχνουμε το t_{ned} που επιβεβαιώνει τη σχέση

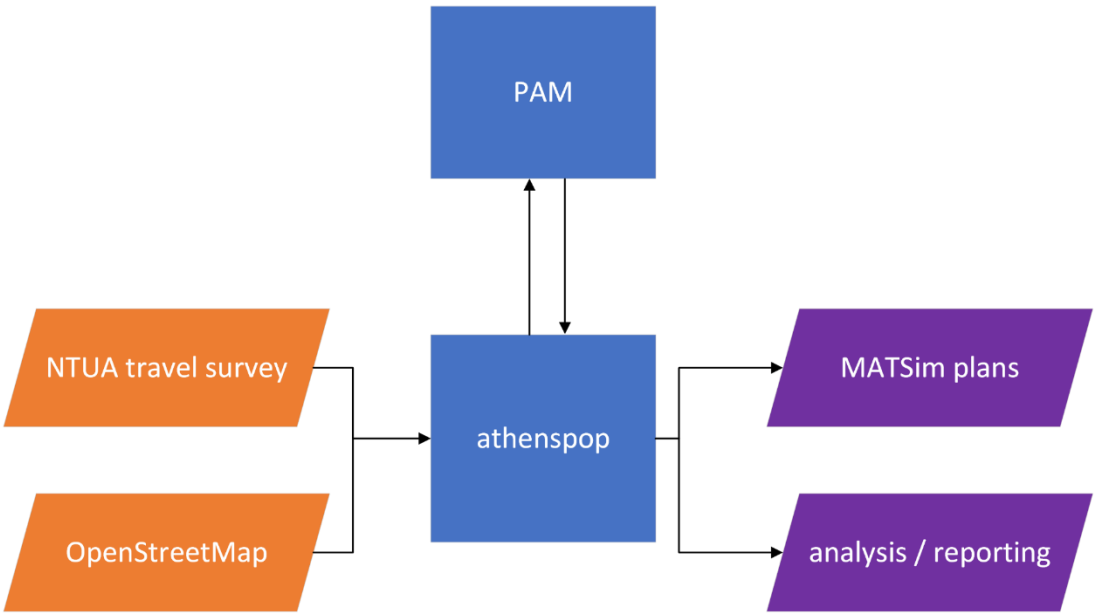
$$S_i(t_{med}) = 0.5$$

4.Μέθοδος Επίλυσης

4.1.Τρόπος λειτουργίας AthensPop πριν την επέμβαση

Για το πρακτικό κομμάτι της εργασίας αποφασίστηκε η εύρεση των μοτίβων δραστηριοτήτων (travel activity patterns) να γίνει στην πόλη της Αθήνας με τη βοήθεια του μοντέλου AthensPop[8]. Το AthensPop όπως αναφέρθηκε και παραπάνω είναι ένα Agent Based μοντέλο που συνθέτει πληθυσμό και από αυτόν καταγράφει τα μοτίβα δραστηριοτήτων μέσω της PAM μιας ανοιχτής βιβλιοθήκης της Python (έκδοση 3.8). Για αυτόν τον λόγο η επίλυση του προβλήματος κατατάσσεται σε ευρετική καθώς η διαδικασία σε κανόνες και δεδομένα για να καθοδηγήσει τη δημιουργία μετακινήσεων. Η έρευνα αποτελεί θεμέλιο λίθο καθώς πάνω σε αυτήν στηρίζονται τα δεδομένα που παράγει το μοντέλο (πληθυσμός και μετακινήσεις με τα χαρακτηριστικά τους).

[8] https://github.com/Theodore-Chatziioannou/athenspop



Εικόνα 1. Workflow AthensPop πριν την επέμβαση. Πηγή: https://github.com/Theodore-Chatziioannou/athenspop

Από το παραπάνω σχήμα φαίνεται η λειτουργία του AthensPop. Ως δεδομένα εισόδου το AthensPop δέχεται την έρευνα και τους χάρτες. Οι χάρτες βοηθούν στη χωρική κατανομή των μετακινήσεων. Όσον αφορά την έρευνα, έχουμε μία έρευνα ερωτηματολογίου με daily activity chains καθώς και τα ξεχωριστά ατομικά χαρακτηριστικά που αναφέρθηκαν και παραπάνω. Με τον ανάλογο κώδικα αυτά τα στοιχεία επεξεργάζονται και με τη βοήθεια της PAM[9] παράγουν τον συνθετικό πληθυσμό και τα ταξίδια του. Έτσι ως παραγόμενα δεδομένα έχουμε τα συνθετικά δεδομένα έτοιμα για να ενταχθούν σε προσομοίωση MATSim [10], μοτίβα δραστηριοτήτων αλλά και μία ανάλυση και οπτικοποίηση των δεδομένων η οποία θα αναλυθεί περαιτέρω αργότερα.

Παρακάτω θα αναλυθεί εκτενώς η δομή της έρευνας και όλα τα mappings για κάθε μεταβλητή καθώς στη συνέχεια θα χρησιμοποιηθεί κάθε μία ξεχωριστή βαθμίδα κάθε μεταβλητής.

[9] https://github.com/arup-group/pam
[10] https://matsim.org/

Πίνακας 4. Μεταβλητές Έρευνας και mappings

Κατηγορία	Όνομα Μεταβλητής	Τύπος	Mappings
Δημογραφικά	Gender	Κατηγορική	(1: male, 0: female)
	Age	Συνεχής	Ηλικία σε Έτη
	Education	Κατηγορική	1: Primary School, 2: High School, 3: Bachelor, 4: Master or Phd
	Employment	Κατηγορική	1: Inactive, 2: Unemployed, 3: Student, 4: Active
	Income	Κατηγορική	0: No income, 1: 750 or less, 2: 750-1500, 3: 1500-2500, 4: 2500 or more
	Car_own	Κατηγορική	0: No, 1: Yes
Χωρικές Μεταβλητές	Dest	Κατηγορική	1: Central Athens, 2: West Athens, 3: East Attica, 4: South Athens, 5: North Athens, 6: Piraeus
	Home	Κατηγορική	1: Central Athens, 2: West Athens, 3: East Attica, 4: South Athens, 5: North Athens, 6: Piraeus
Μέσο Μετακίνησης	Mode	Κατηγορική	1: car, 2: taxi,3: bus, 4: train, 5: motorcycle, 6: bicycle, 7: walk, 8: E-scooter
Χρονικές Μεταβλητές	Time	Συνεχής	Χρόνος Έναρξης Μετακίνησης σε ώρες
Απόσταση Μετακίνησης	Dist	Συνεχής	Απόσταση Μετακίνησης σε m

Το μοντέλο παράγει πληθυσμό τον οποίον κατατάσσει σε νοικοκυριά. Στα αποτελέσματα συντίθεται ένα ποσοστό του συνολικού πληθυσμού της Αττικής , το οποίο επιλέγεται και του οποίου τα χαρακτηριστικά θεωρούνται αντιπροσωπευτικά όλου του πληθυσμού. Στη συγκεκριμένη περίπτωση 0.01% του πληθυσμού της Μητροπολιτικής Αττικής. Οι παραγόμενες μετακινήσεις τους επιδέχονται μιας τυχαίας

διασποράς (jittering) δηλαδή μια τεχνική κατά την οποία ορίζονται τα χρονικά πλαίσια μίας τυχαιότητας για τις ώρες έναρξης των μετακινήσεων. Έτσι, σε κάθε ώρα έναρξης μιας μετακίνησης που υπάρχει στην έρευνα προστίθεται μία τιμή είτε θετική είτε αρνητική σε ένα ορισμένο εύρος (+/- 30 λεπτά). Με αυτόν τον τρόπο καταφέρνουμε η κάθε παρατήρηση όταν πολλαπλασιαστεί και γίνει η πηγή πολλών να έχει διαφορετική ώρα έναρξης μετακίνησης και να παράγονται πιο αληθοφανή αποτελέσματα. Πιο αναλυτικά, αν μια μετακίνηση υποτίθεται πως ξεκινάει στις 8 με ένα jittering 30 λεπτών έχει ίσες πιθανότητες να ξεκινάει οποιαδήποτε χρονική στιγμή μεταξύ 07:30 και 08:30. Έτσι, προσδίδεται μία ρεαλιστικότητα στα χρονολόγια των μετακινήσεων καθώς είναι σαφές πως οι απαντήσεις δεν είναι απαραίτητα ακριβείς. Όλα τα υπόλοιπα χαρακτηριστικά των μετακινήσεων μένουν ως έχουν αφού αντλούνται από την έρευνα πολλαπλασιασμένα αναλόγως το ποσοστό του πληθυσμού που θέλουμε να πετύχουμε.

Προσοχή πρέπει να δοθεί σε σφάλματα ή μη λογικά δεδομένα που βρίσκονται κατά την επεξεργασία. Καθώς αυτό που μας ενδιαφέρει είναι να οπτικοποιηθούν τα μοτίβα των δραστηριοτήτων μίας τυπικής ημέρας γίνεται κατανοητό πως κατά τη συμπλήρωση της έρευνας εάν υπάρχουν μετακινήσεις οι οποίες ξεπερνούν τις 12 το βράδυ φροντίζεται μέσα από το μοντέλο και τη συνάρτηση *stepday* αυτές να περνούν στην επόμενη μέρα και εν τέλει να αγνοούνται στο τελικό χρονολόγιο των δραστηριοτήτων καθώς ενδιαφερόμαστε για ένα 24ωρο. Μία ακόμη συνάρτηση στην οποία και θα εστιάσουμε είναι η *fix_nobackhome*. Ένα πρόβλημα που αντιμετωπίζει το μοντέλο έχει να κάνει με τις απαντήσεις της έρευνας. Πιο συγκεκριμένα, πολλοί από τους ερωτηθέντες συμπλήρωναν κάποιες μετακινήσεις αλλά ξεχνούσαν να αναφέρουν την επιστροφή τους στο σπίτι. Μιας και η διάρκεια των δραστηριοτήτων υπολογίζεται ως η διαφορά των ωρών έναρξης δύο διαδοχικών μετακινήσεων ($time_{i+1} - time_i$, όπου i ο αριθμός του ταξιδιού) καταλαβαίνουμε πως με το να αγνοείται η επιστροφή στο σπίτι η αμέσως προηγούμενη μετακίνηση διαρκεί μέχρι το τέλος της ημέρας. Κάτι τέτοιο θα μπορούσε να συμβαίνει μόνο για λόγους ψυχαγωγίας. Για αυτόν τον λόγο, όσοι μετακινούμενοι δεν αναφέρουν την επιστροφή τους στο σπίτι και η τελευταία τους μετακίνηση δεν έχει σκοπό την ψυχαγωγία καλούμε τη συγκεκριμένη συνάρτηση η οποία έχει στόχο να προσθέσει την επιστροφή στο σπίτι αλλά και να προβλέψει τι ώρα έγινε αυτό, και κατ' επέκταση τη διάρκεια της αμέσως προηγούμενης δραστηριότητας. Για να γίνει αυτό απαραίτητες είναι επίσης οι συναρτήσεις που δημιουργούν δείγμα διάρκειας δραστηριοτήτων βασισμένες στην υπόθεση κάποιας κατανομής. Έτσι, έχουν οριστεί οι συναρτήσεις *create duration sampler gaussian* & *create duration sampler*

empirical. Αποτελεί επιλογή του χρήστη ποια από τις δύο θα επιλέξει. Στην ουσία αυτές οι συναρτήσεις συλλέγουν τα απαραίτητα στατιστικά στοιχεία των διαρκειών των δραστηριοτήτων και είτε υποθέτοντας ότι ακολουθούν την κανονική κατανομή είτε την εμπειρική, προβλέπουν τη διάρκεια της αμέσως προηγούμενης δραστηριότητας. Έτσι, υπολογίζεται αυτόματα και η ώρα της επιστροφής στο σπίτι.

Καθώς όπως προαναφέρθηκε η συνάρτηση fix_nobackhome καλείται σε περίπτωση που δεν αναφέρεται επιστροφή στο σπίτι εάν η προηγούμενη μετακίνηση αφορά έναν από τους παραπάνω σκοπούς. Σε αυτήν την περίπτωση προστίθεται μία μετακίνηση με σκοπό return home και πρέπει να προβλεφθεί η ώρα επιστροφής άρα και η διάρκεια της ακριβώς προηγούμενης διαδρομής. Όπως φαίνεται και στις παρακάτω φωτογραφίες (Εικόνες 2 & 3) αρχικά δεν υπάρχει μετακίνηση επιστροφής στο σπίτι ενώ φαίνεται η τελευταία μετακίνηση να αφορά εργασιακούς σκοπούς και να ξεκινάει στις 18:00. Μετά την παρέμβαση της συνάρτησης προστίθεται η μετακίνηση επιστροφής στο σπίτι και μέσω του μοντέλου εκτιμάται η διάρκεια της μετακίνησης για εργασιακούς σκοπούς και εν συνεχεία εκτιμάται και η ώρα επιστροφής στο σπίτι. Με αυτόν τον τρόπο παρέχουμε μια πληρέστερη έρευνα η οποία λειτουργεί ως δεδομένα εισόδου στο Agent based μοντέλο που θα χρησιμοποιηθεί για την εξαγωγή των μοτίβων δραστηριοτήτων.

Στόχος της έρευνας είναι να βελτιώσουμε την ακρίβεια των μετακινήσεων που προβλέπονται. Μία προσέγγιση ως προς την επίλυση αυτού του προβλήματος αποτελεί η εκτίμηση τους στις περιπτώσεις που αυτή κρίνεται απαραίτητη μέσω του cox μοντέλου που μελετήθηκε καθώς ο συνηθέστερος τρόπος εκτίμησης διαρκειών όπως μελετήθηκε κατά τη βιβλιογραφική ανασκόπηση είναι τα hazard based μοντέλα και σίγουρα όχι η υπόθεσης μιας στατιστικής κατανομής.

Για αυτόν τον λόγο αντιμετωπίζουμε την ως πρότινος ενιαία διάρκεια δραστηριοτήτων σε ξεχωριστές διάρκειες ανά σκοπό μετακίνησης. Με αυτόν τον τρόπο αφαιρούμε μεγάλο μέρος της πολυπλοκότητας της διάρκειας δραστηριοτήτων ως μία μεταβλητή για όλους τους τύπους μετακίνησης.

pid	gender	age	education	employment	income	car_own	home	dest1	purp1	mode1	time1	dest2	purp2	mode2	time2	dest3	purp3	mode3	time3
111	1: male	21	2: high school	3: student	1: 750 or less	1: yes	12	22	1: work	4: train	13	18	1: work	4: train	18				

Εικόνα 2. Ελλιπής έρευνα

pid	gender	age	education	employment	income	car_own	home	dest1	purp1	mode1	time1	dest2	purp2	mode2	time2	dest3	purp3	mode3	time3
111	1: male	21	2: high school	3: student	1: 750 or less	1: yes	12	22	1: work	4: train	13	18	1: work	4: train	18	12	2: return home	4: train	22

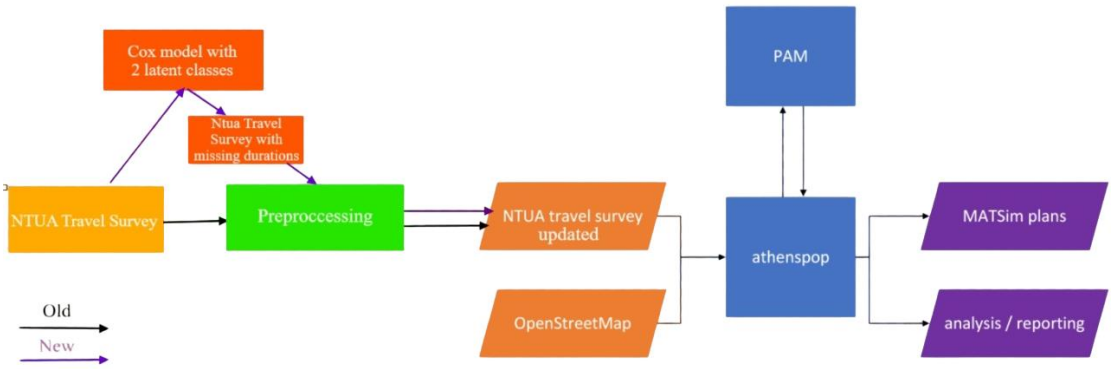
Εικόνα 3. Συμπληρωμένη έρευνα με fix_nobackhome

4.2 Τρόπος Λειτουργίας AthensPop μετά την επέμβαση

Για την ανάλυση της εργασίας τα μοντέλα δημιουργήθηκαν ανά ξεχωριστό σκοπό και πιο συγκεκριμένα για σκοπούς μετακίνησης:

- Εργασία
- Εκπαίδευση
- Ψώνια
- Λοιπές μετακινήσεις και υπηρεσίες

Με την έμφαση να δίνεται στις διάρκειες δραστηριοτήτων, το νέο workflow μοιάζει κάπως έτσι:



Εικόνα 4. Workflow AthensPop μετά την επέμβαση

Με μαύρο βελάκι απεικονίζεται η διαδικασία πριν την παρέμβαση ενώ με μωβ βελάκι μετά την παρέμβαση.

Αρχικά αποτελείται από μία έρευνα ερωτηματολογίου ημερήσιας αλληλουχίας μετακινήσεων. Σε κάθε μία παρατήρηση γίνεται έλεγχος για τις 3 συγκεκριμένες περιπτώσεις που θα κληθούν τα μοντέλα να εκτιμήσουν κάποια διάρκεια. Πιο συγκεκριμένα οι 3 περιπτώσεις είναι:

- Διάρκειες που λείπουν (Στην έρευνα συναντήθηκε πολλές φορές η απουσία κάποιας ώρας έναρξης. Αντί να αφαιρεθεί η παρατήρηση και να οδηγήσει σε συρρίκνωση της βάσης δεδομένων εκτιμάται η διάρκεια της μετακίνησης και συμπληρώνεται η ώρα έναρξης)
- Τυπογραφικά λάθη (Αρκετές φορές βρέθηκαν από την πρώτη κιόλας

μετακίνηση αρνητικές διάρκειες)

- Fix no back home (Όπως προαναφέρθηκε σε περιπτώσεις μη επιστροφής στο σπίτι)

Εκμεταλλευόμαστε τα cox μοντέλα που έχουμε δημιουργήσει και ακολουθώντας τις συναρτήσεις που αναφέρθηκαν στο κεφάλαιο της μεθοδολογίας και γνωρίζοντας τα ατομικά χαρακτηριστικά κάθε παρατήρησης υπολογίζουμε τις συγκεκριμένες διάρκειες για κάθε παρατήρηση. Αυτές πλέον περνούν στην έρευνα η οποία είναι εμπλουτισμένη και έτοιμη να χρησιμοποιηθεί ως δεδομένο εισόδου. Αλλάζοντας και τον preprocessing κώδικα δε χρειάζεται πλέον να υποτεθεί κάποια κατανομή (κανονική ή Weibull) για να εκτιμηθούν οι διάρκειες που λείπουν. Έτσι, στη συνέχεια χωρίς κάποια αλλαγή παίρνουμε ως δεδομένα εξόδου και τα μοτίβα δραστηριοτήτων καθώς και την κυκλοφοριακή ανάλυση όπως θα παρουσιαστεί και παρακάτω.

5.Εφαρμογή και Αποτελέσματα

Στη συγκεκριμένη ενότητα θα σχολιαστεί αναλυτικά η λειτουργία των cox μοντέλων, τα παραγόμενα δεδομένα αλλά και πώς αυτά επηρεάζουν τα μοτίβα δραστηριοτήτων αλλά και τη γενικότερη ανάλυση. Όπως έχει προαναφερθεί, όλα τα μοντέλα έχουν υπολογιστεί per purpose καθώς είναι ένας τρόπος να μειωθεί η πολυπλοκότητα των μεταφορών. Έτσι, έχουμε συντελεστές μεταβλητών και frailty όρων, ανάλυση και οπτικοποιήσεις δεδομένων ανά ξεχωριστό σκοπό μετακίνησης.

5.1 Σύγκριση cox μοντέλων

Όλα τα μοντέλα όπως έχει προαναφερθεί υπολογίζονται ξεχωριστά per purpose. Κατά τη σύγκριση θα παρουσιαστούν τα αποτελέσματα για ένα απλό cox μοντέλο όπως του J.L.Yee (2000) και του μοντέλου που μελετάται, δηλαδή του cox μοντέλου με 2 latent classes για μη παρατηρήσιμες μεταβλητές μέσα από τα δεδομένα της έρευνας.

Ως μέτρο σύγκρισης θα χρησιμοποιηθεί το AIC για το απλό cox μοντέλο ενώ για το cox με frailty όρους θα χρησιμοποιηθούν το WAIC καθώς το frailty μοντέλο αναπτύσσεται μέσω Bayesian Networks και εκείνος είναι ο αντίστοιχος όρος λόγω

πολυπλοκότητας. Αυτή θα είναι μία πρώτη σύγκριση μεταξύ των μοντέλων, στην οποία θα συγκριθεί στην ουσία πόσο καλά περιγράφει το μοντέλο τα δεδομένα της έρευνας, πόσο μεγάλη προβλεπτική ακρίβεια έχει και όλα αυτά λαμβάνοντας υπόψιν την πολυπλοκότητα των μοντέλων. Για τη σύγκριση έχουν χρησιμοποιηθεί είτε όλες οι μεταβλητές είτε λιγότερες αναλόγως το μοντέλο. Πιο αναλυτικά, αφαιρέθηκαν μεταβλητές οι οποίες είχαν υψηλό δείκτη VIF (>10), δηλαδή μεταβλητές που είχαν ισχυρή συσχέτιση μεταξύ τους. Κάτι τέτοιο θα μπορούσε να μειώσει την αξιοπιστία του μοντέλου. Για να συγκριθούν βέβαια επί ίσοις όροις, όποιες μεταβλητές χρησιμοποιήθηκαν στο απλό cox χρησιμοποιήθηκαν και στο πιο πολύπλοκο με τους frailty όρους. Επισημαίνεται πως κάθε μοντέλο έχει ίδια ακριβώς βάση δεδομένων και ίδιες μεταβλητές για να δώσει μία αξιοπιστία στο συγκεκριμένο τεστ καθώς τα δύο metrics υπολογίζονται με διαφορετικό τρόπο.

$$AIC = 2k - 2\log(L)$$

Όπου, k = αριθμός παραμέτρων

L = Η πιθανότητα να ταιριάζουν τα αποτελέσματα με τα δεδομένα (Likelihood)

Το WAIC είναι μεν παρόμοιο metric, αφορά δε μοντέλα που χρησιμοποιούν Bayesian Networks. Πιο συγκεκριμένα, εκφράζεται από τη σχέση:

$$WAIC = 2 \cdot p_{WAIC} - 2\,\mathrm{mean}(\log(L))$$

Όπου,

- p_{waic} είναι η ποινή για την πολυπλοκότητα και εκτιμάται ως η διακύμανση του log – likelihood
- $mean(\log(L))$ είναι ο μέσος όρος της λογαριθμικής πιθανότητας

Και τα 2 metrics έχουν ως στόχο να συγκρίνουν τα δύο duration models ως προς έναν συνδυασμό ακρίβειας και απλότητας καθώς στα συγκεκριμένα τεστ η πολυπλοκότητα που δεν προσφέρει, τιμωρείται με αύξηση του metric. Στόχος όσο πιο μικρό metric γίνεται. Παρακάτω φαίνονται οι συντελεστές των μοντέλων για κάθε μεταβλητή όπως και τα υπολογισμένα metrics που χρησιμοποιήθηκαν για τη σύγκριση (Με πράσινο κελί

τα καλύτερα τεστ κάθε φορά). Στις εικόνες 5, 6, 7, 8 παρουσιάζονται οι συντελεστές Xi των μεταβλητών β ανά σκοπό μετακίνησης για να ακολουθηθεί η διαδικασία που περιγράφεται στις ενότητες 3.2.1. και 3.2.2. Θετικοί συντελεστές σημαίνει πως αυξάνεται η διάρκεια της δραστηριότητας σε σχέση με τη μεταβλητή αναφοράς που για κάθε κατηγορία έχει προαποφασιστεί αυτόματα από την R ποια θα είναι. Σε κάθε πίνακα παρουσιάζονται οι συντελεστές και για το απλό cox και για το ενισχυμένο cox με frailty. Επίσης, όσον αφορά τα στατιστικά τεστ, οι μικρότερες τιμές είναι και οι καλύτερες.

Simple Cox for work trips		**CoxFrailty model for work trips**		
Variable		Variable	Latent Class 1 Coefficient	Latent Class 2 Coefficient
gender1	0.06607	gender1	1.50340399	0.03035941
income1	0.70821	income1	0.52587342	-0.02361092
income2	0.29834	income2	1.486609309	-0.002202576
income3	0.23001	income3	-0.04952334	-0.01334346
income4	0.49772	income4	-0.1409907	0.5271416
car_own1	0.04747	car_own1	-0.03310816	-0.43971262
dest2	0.07214	mode2	-0.08449414	0.0635485
dest3	-0.26045	mode3	-0.04265314	0.02530859
dest4	0.23836	mode4	-0.07100655	-0.04079558
dest5	0.04472	mode5	-0.04650754	0.01337012
dest6	-0.09533	mode6	-0.05007086	0.007570051
dist	0.07205	mode7	0.05182202	0.02531785
time	1.04976	dest2	0.02284334	-0.03119678
age	0.09896	dest3	0.006381445	-0.033065584
mode3	-0.46617	dest4	0.07264133	-0.09132006
mode2	-0.18209	dest5	-0.01190793	-0.21142202
mode5	-0.41842	dest6	0.008215312	0.00513783
moder4	-0.29365	time	0.01213347	-0.017837
mode7	-0.08208	age	0.0568433	-0.01392283
		dist	0.48945862	-0.01027033
AIC	3162.45	**WAIC**	1590.606	

Εικόνα 5. Σύγκριση AIC & WAIC Cox μοντέλων για εργασιακές μετακινήσεις

Το WAIC του ενισχυμένου cox μοντέλου ισούται με 1590.606, πολύ μικρότερο από το AIC του απλού cox που ισούται με 3162.45 οπότε μπορούμε με ασφάλεια να ισχυριστούμε πως το ενισχυμένο είναι αρκετά πιο ακριβές στις προβλέψεις του.

Simple Cox for other trips		CoxFrailty model for other trips		
Variable		Variable	Latent Class 1 Coefficient	Latent Class 2 Coefficient
gender1	-0.3773	gender1	0.4505691	0.1588863
education3	1.5055	education3	0.4236803	0.1201465
education4	0.1457	education4	0.4325918	0.2242078
income1	-1.2281	income1	0.1164397	0.2399594
income2	0.3316	income2	0.1099807	0.0379478
income3	0.2898	income3	0.39382638	0.04058856
income4	-0.5531	income4	0.4022377	-0.3144901
car_own1	0.5254	car_own1	0.3486657	-0.3328318
time	0.1469	time	0.3470449	-0.1634363
age	1.3705	age	0.3274456	-0.1608515
dist	0.6622	dist	0.3352419	-0.03707611
AIC	171.8394	WAIC	166.1761	

Εικόνα 6. Σύγκριση AIC & WAIC Cox μοντέλων για μετακινήσεις για λοιπές μετακινήσεις

Για την κατηγορία των λοιπών μετακινήσεων ναι μεν το WAIC είναι μικρότερο αλλά οι 5 μονάδες διαφοράς από το AIC του απλού Cox δεν είναι αρκετές από μόνες τους για ασφαλή συμπεράσματα οπότε θεωρείται ότι από το συγκεκριμένο τεστ δεν υπάρχουν έντονες διαφορές και αναμένονται και τα υπόλοιπα για τη λήψη ακριβέστερων συμπερασμάτων

Simple Cox for market trips		CoxFrailty model for market trips		
Variable		Variable	Latent Class 1 Coefficient	Latent Class 2 Coefficient
gender1	0.78161	gender1	0.427822	-0.2127371
education2	0.90052	education2	0.2962413	-0.1157026
education4	-0.182	education4	0.24319871	-0.00856842
mode2	-0.07168	car_own1	0.007776561	-0.364065484
mode5	0.58311	mode2	0.84870383	-0.02662871
mode4	-0.17005	mode3	0.083510248	0.008921855
mode7	0.55415	mode4	0.209935575	-0.007859116
car_own1	0.11415	mode5	0.78954	-0.277325
dist	-0.13073	mode6	-0.1551232	-0.01105033
time	0.93292	mode7	-0.099894802	0.007521191
age	0.44967	dist	0.1548397	-0.206963
		time	0.20750502	-0.05706515
		age	0.4520385	-0.1132496
AIC	285.441	WAIC	103.1039	

Εικόνα 7. Σύγκριση AIC & WAIC Cox μοντέλων για μετακινήσεις για ψώνια

Για τις μετακινήσεις για ψώνια το WAIC ισούται με 103.1039 και είναι ξεκάθαρα καλύτερο από το AIC του απλού cox μοντέλου που ισούται με 285.441

Simple Cox for educational trips		CoxFrailty model for educational trips		
Variable		Variable	Latent Class 1 Coefficient	Latent Class 2 Coefficient
gender1	0.351893	gender1	0.66523762	0.04720101
employment2	0.724891	employment2	0.51990633	0.01293419
employment3	0.160933	employment3	0.599233147	-0.008544966
employment4	0.149458	employment4	0.3400249	0.1732747
car_own1	0.041381	car_own1	0.4476138	0.2218874
dest2	-1.18387	dest2	0.37737315	-0.02846846
dest3	0.722663	dest3	0.429522	-0.0720768
dest4	0.130873	dest4	0.31449912	0.04710893
dest5	-0.74865	dest5	0.42144398	0.03287825
dest6	0.473	dest6	0.1348655	-0.203553
dist	-0.02695	dist	0.1174283	-0.2416647
time	1.093069	time	0.24707234	0.03521755
age	0.003572	age	0.24960111	0.02201722
AIC	257.6539	WAIC	229.4468	

Εικόνα 8. Σύγκριση AIC & WAIC Cox μοντέλων για εκπαιδευτικές μετακινήσεις

Τέλος, για εκπαιδευτικές μετακινήσεις το WAIC ισούται με 229.4468 και είναι και αυτό καλύτερο από το AIC του απλού cox μοντέλου που ισούται με 257.6539

5.2 Split train/test comparison

Μία ακόμα πολύ σημαντική σύγκριση που δε θα μπορούσε να λείπει αφορά τη διαδικασία όπου χωρίζουμε τα δεδομένα μας σε training και testing. Στη συγκεκριμένη περίπτωση τα δεδομένα χωρίστηκαν 80% training και 20% testing. Με αυτόν τον τρόπο αποφεύγουμε την πιθανότητα του overfitting καθώς όταν στο μοντέλο τοποθετούνται όλα τα δεδομένα υπάρχει ο κίνδυνος το μοντέλο να προσαρμοστεί σε αυτά. Κάτι τέτοιο είναι κακό για την προσαρμοστικότητα του σε νέα δεδομένα και μας οδηγεί σε παραπλανητικά αποτελέσματα και συμπεράσματα. Ξανά, τα τεστ έγιναν στα μοντέλα per trip purpose.

Στη συνέχεια υπολογίζουμε mean & median durations. Για τα εργασιακά και τα υπολειπόμενα ταξίδια θα χρησιμοποιηθούν οι median durations καθώς έχουν μεγαλύτερο εύρος στις πιθανές διάρκειες ενώ για τα εκπαιδευτικά ταξίδια και για αυτά για ψώνια θα χρησιμοποιηθούν mean durations. Έχοντας λοιπόν όλες τις ζητούμενες διάρκειες για το 20% του testing data θα ελέγξουμε εάν οι τιμές που προβλέφθηκαν είναι κοντά με τις πραγματικές τιμές και ποιο από τα δύο μοντέλα έκανε καλύτερη δουλειά. Για αυτόν τον έλεγχο θα χρησιμοποιήσουμε το MAE και το RMSE, δύο στατιστικούς δείκτες, καθώς και το IBS.

Σχετικά με το MAE, εκφράζεται από τη σχέση:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - y_{Pi}|$$

Όπου, y_i η πραγματική διάρκεια για κάθε παρατήρηση i και

y_{pi} η προβλεπόμενη διάρκεια για κάθε παρατήρηση i.

Χρησιμοποιείται για να υπολογίζει το μέσο μέγεθος του σφάλματος στις προβλέψεις.

Αναλόγως λειτουργεί και το RMSE το οποίο περιγράφεται από την παρακάτω σχέση:

$$RMSE = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_{Pi})^2}$$

Με τους ανάλογους συμβολισμούς δίνει έμφαση στα πιθανά outliers αφού υψώνονται στο τετράγωνο και τονώνει πιθανά λάθη.

Και στους δύο δείκτες θέλουμε να είναι όσο το δυνατόν πιο κοντά στο 0, κάτι που δείχνει πως οι προβλεπόμενες τιμές διάρκειας δραστηριοτήτων είναι κοντά στις αληθινές.

Ένα άλλο μέτρο σύγκρισης που χρησιμοποιήθηκε αφορά το τεστ IBS. Πιο αναλυτικά, το BS (Brier Score) είναι η μέση τετραγωνική διαφορά μεταξύ της πραγματικής κατάστασης (event ή no event) και της προβλεπόμενης πιθανότητας από το μοντέλο:

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i(t) - S_i(t)\right)^2$$

Όπου,

Συμβολισμός	*Περιγραφή*
T	Χρονική στιγμή
N	Αριθμός παρατηρήσεων
$\delta_i(t)$	δ_i(t) Δείκτης που δείχνει αν το άτομο

	i έχει υποστεί το γεγονός μέχρι τη στιγμή t: 1 αν το γεγονός έχει συμβεί. 0 αν δεν έχει συμβεί.
$S_i(t)$	Η εκτιμώμενη πιθανότητα επιβίωσης από το μοντέλο για το άτομο i στη χρονική στιγμή t.

Το IBS αποτελεί το ολοκληρωμένο BS που εκφράζεται από τη σχέση:

$$IBS = \frac{1}{tmax}\int_0^{tmax} BS(t)dt$$

Στην ουσία το συγκεκριμένο τεστ εκτιμά τη δυνατότητα πρόβλεψης των μοντέλων επιβίωσης, παίρνει τιμές που κυμαίνονται από 0 έως 1, με το 0 να σημαίνει τέλεια πρόβλεψη ενώ το 1 το απόλυτο σφάλμα

Οι συγκρίσεις έγιναν για work, market & other trips καθώς τα education data ήταν ήδη λιγοστά και η μείωσή τους κατά 80% αφού εξετάζεται μόνο το 20% test data έκανε τόσο μικρό το δείγμα που δε θα μπορούσαν να αντληθούν αξιόπιστα συμπεράσματα. Παρακάτω, οι πίνακες των συγκρίσεων:

Πίνακας 5.Στατιστικά τεστ για εργασιακές μετακινήσεις

Work trips models comparison		
Train/Test (80/20)		
Metric	Simple Cox	Cox with frailty
MAE	5.865	4.387
RMSE	6.796	5.161
IBS	0.149	0.029

Πίνακας 6.Στατιστικά τεστ για μετακινήσεις για ψώνια

Market trips models comparison		
Train/Test (80/20)		
Metric	Simple Cox	Cox with frailty
MAE	2.317	1.5
RMSE	2.89	2.425
IBS	0.361	0.018

Πίνακας 7. Στατιστικά τεστ για λοιπές μετακινήσεις

Other trips models comparison		
Train/Test (80/20)		
Metric	Simple Cox	Cox with frailty
MAE	3.677	2.28
RMSE	4.257	2.858
IBS	0.135	0.099

Γίνεται ευκόλως κατανοητό πως το cox with frailty μοντέλο που μελετούμε είναι αποτελεσματικότερο στην πρόβλεψη των διαρκειών καθώς υπερισχύει σε όλα τα test και για όλα τα purposes.

5.3 Αληθινά vs συνθετικά δεδομένα

Συνεχίζοντας λοιπόν θα γίνει και μία ακόμα ευρύτερη σύγκριση. Πιο συγκεκριμένα, Η σύγκριση που γίνεται αφορά τα αληθινά vs συνθετικά δεδομένα. Ψάχνουμε λοιπόν να βρούμε ποια συνθετικά είναι πιο κοντά στα αληθινά μας δεδομένα (έρευνα ερωτηματολογίου). Τα συνθετικά τα οποία ουσιαστικά θα τεθούν σε σύγκριση αφορούν συνθετικά δεδομένα που παράγονται από το AthensPop με 3 διαφορετικούς τρόπους (3 διαφορετικές έρευνες ως δεδομένα εισόδου).

Οι 3 αυτοί τρόποι είναι:

1. Απλή έρευνα χωρίς καμία αλλαγή. Σε αυτήν την περίπτωση φροντίζει το AthensPop από μόνο του και υποθέτει την κανονική κατανομή για τα durations όταν χρειάζεται η συμπλήρωση καθώς λείπει.
2. Εμπλουτισμένη έρευνα με απλό cox μοντέλο χωρίς frailty
3. Εμπλουτισμένη έρευνα με cox μοντέλο με frailty (2 latent classes που αντικατοπτρίζουν μη παρατηρήσιμα μοτίβα που επηρεάζουν τα durations και δεν καλύπτουν οι μεταβλητές της έρευνας).

Έτσι, θα εξεταστούν αυτά τα 3 διαφορετικά συνθετικά δεδομένα και πιο συγκεκριμένα τα συνθετικά durations για να αποδείξουμε πως το cox μοντέλο με frailty όρους αυξάνει την πολυπλοκότητα, βοηθάει στις ακριβέστερες προβλέψεις και τελικά έχει ως αποτέλεσμα καλύτερα μοτίβα δραστηροτήτων.

Οι τρόποι σύγκρισης θα είναι ξανά το MAE και το RMSE αλλά και το Kolmogorov – Smirnov test το οποίο είναι ιδανικό για να συγκρίνει πως κατανέμονται οι διάρκειες σε σύγκριση με τα πραγματικά δεδομένα και αν ταιριάζουν από στατιστικής πλευράς. Διεξοδικότερα, θέλουμε όσο το δυνατόν μικρότερο D-Statistic που δείχνει ότι οι κατανομές των συγκρινόμενων δεδομένων είναι όσο το δυνατόν πιο κοντά. Ακόμα, όσο το δυνατόν μεγαλύτερο p-value μας δείχνει τη στατιστική σημαντικότητα της διαφοράς των δύο κατανομών. Τέλος, θα παρουσιαστεί και ένα boxplot που στην ουσία αποτελεί μια οπτική απεικόνιση των διαρκειών και αληθινών και προβλεπόμενων για να ολοκληρωθεί η σύγκριση. Στα συγκεκριμένα διαγράμματα θα εστιάσουμε σε 4 σημαντικά σημεία:

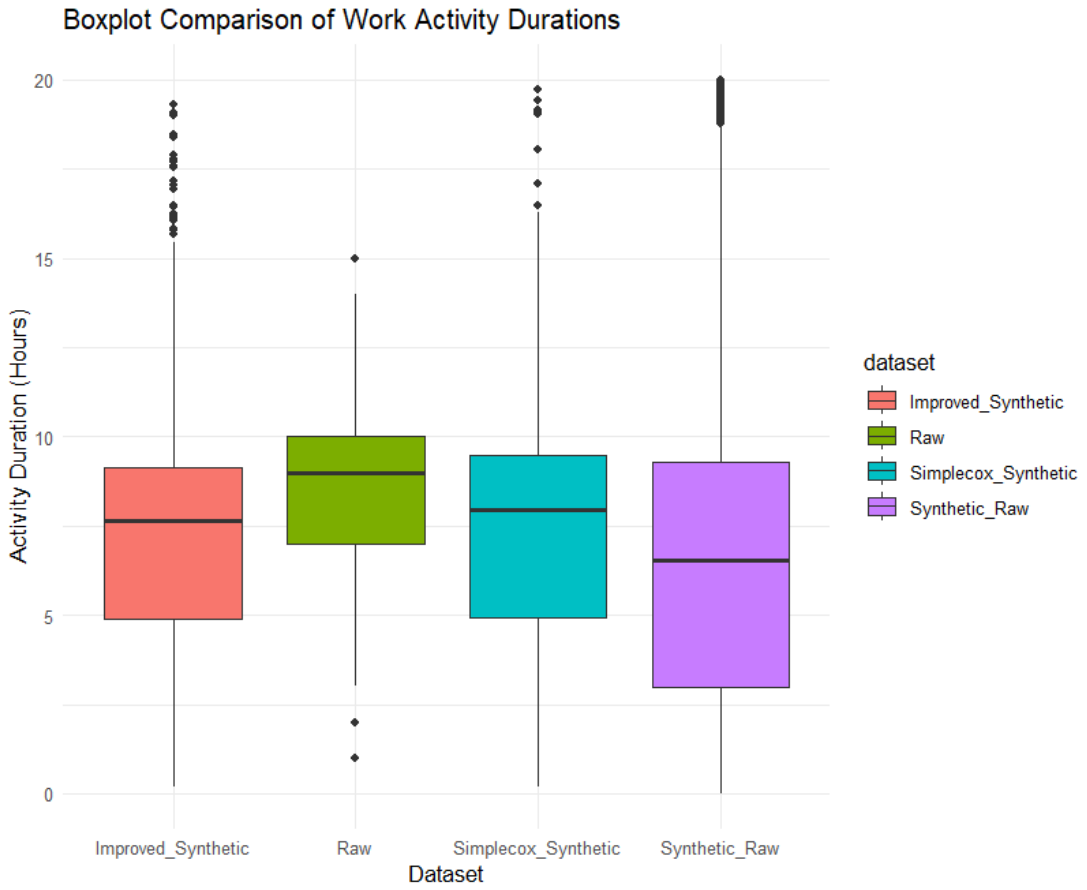
1. Θέση του median (μεσαία γραμμή bold γραμμή σε κάθε κουτί, εκφράζει την κεντρική τάση)
2. Διαστάσεις του κουτιού (IQR, εκφράζει τη διασπορά από το 25^{ο} έως το 75^{ο} ποσοστημόριο)
3. Άκρα (Εκφράζει τη συνολική διασπορά)
4. Ακραίες τιμές (Outliers, φαίνονται ως σημεία κατά τη διάσταση y).

Παρακάτω στους πίνακες φαίνονται τα αποτελέσματα των στατιστικών τεστ και στις φωτογραφίες το διάγραμμα boxplot.

5.3.1. Work durations Comparison

Πίνακας 8. Αληθινά vs συνθετικά δεδομένα για εργασιακές μετακινήσεις

Work Comparison real Vs synthetic data			
Metrics	Raw vs Synthetic Raw	Raw vs Improved Synthetic	Raw vs Simplecox_synthetic
K-S test			
D	0.2939	0.20822	0.25486
p-value	< 2.2e-16	3.46E-11	< 2.2e-16
MAE	4.572079	3.664099	3.707881
RMSE	5.609498	4.723046	4.739945



Εικόνα 9. Boxplot σύγκριση αληθινών vs συνθετικών δεδομένων για εργασιακές μετακινήσεις

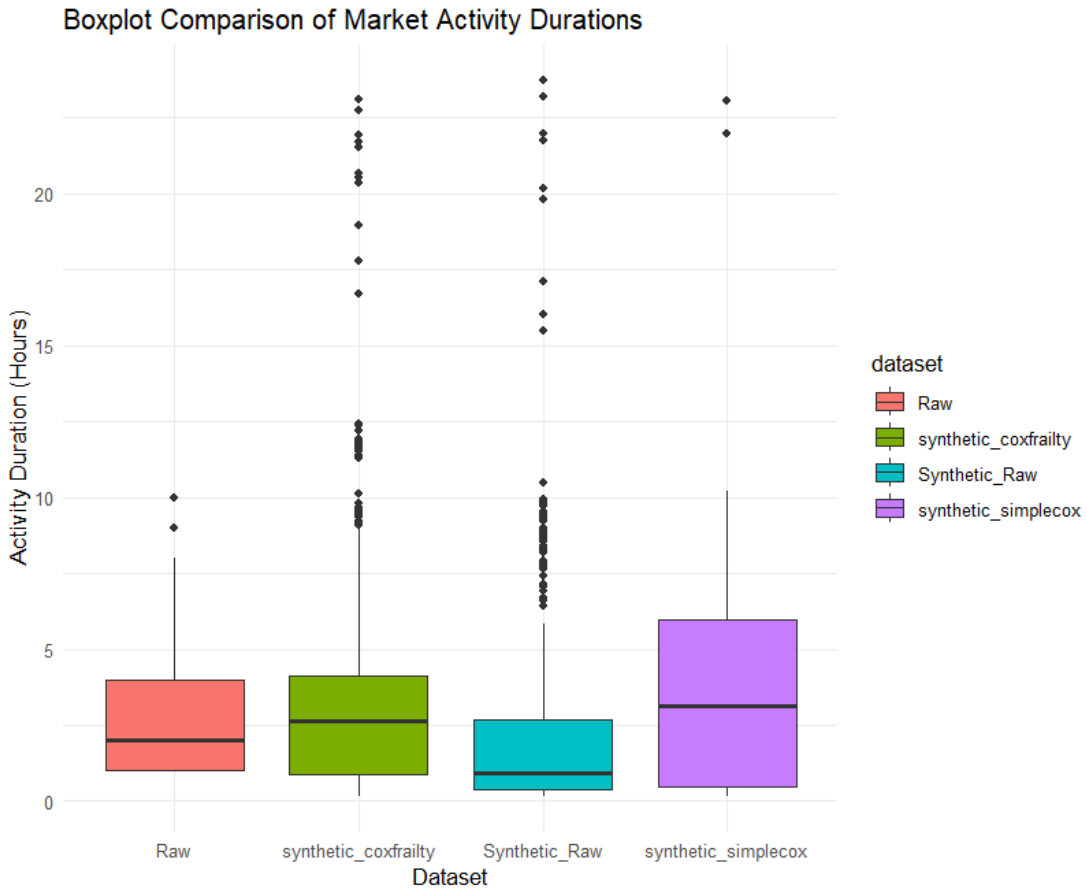
Όσον αφορά τη median τιμή (Διάμεσο) τα συνθετικά που προέρχονται από τα δύο Cox μοντέλα είναι πάρα πολύ κοντά με τα συνθετικά που προέρχονται από την έρευνα

εμπλουτισμένη με τις διάρκειες από το απλό Cox μοντέλο να είναι οριακά ψηλότερα και κοντινότερα στην πράσινη διάμεσο που είναι των πραγματικών δεδομένων. Σχετικά με το IQR το ύψος του μπλε δείχνει μια διασπορά λίγο χαμηλότερες διάρκειες σε σχέση με το ροζ και το πράσινο. Γενικά παρατηρούμε πολύ κοντινά boxplots αλλά λόγω των στατιστικών τεστ μπορούμε με σιγουριά να ισχυριστούμε πως τα συνθετικά που προέκυψαν από το cox with frailty μοντέλο είναι καλύτερα από τα συνθετικά με το απλό cox και τα συνθετικά που προέκυψαν μέσω κανονικής κατανομής.

5.3.2. Market Durations Comparison

Πίνακας 9. Αληθινά vs συνθετικά δεδομένα για μετακινήσεις για ψώνια

Market Comparison real Vs synthetic data			
Metrics	Raw vs Synthetic Raw	Raw vs Improved Synthetic	Raw vs Simplecox_synthetic
K-S test			
D	0.8619	0.4235	0.62409
p-value	< 2.2e-16	3.71E-07	1.73E-14
MAE	2.685109	2.407971	2.3248
RMSE	3.932881	3.591605	3.943628



Εικόνα 10. Boxplot σύγκριση αληθινών vs συνθετικών δεδομένων για μετακινήσεις για ψώνια

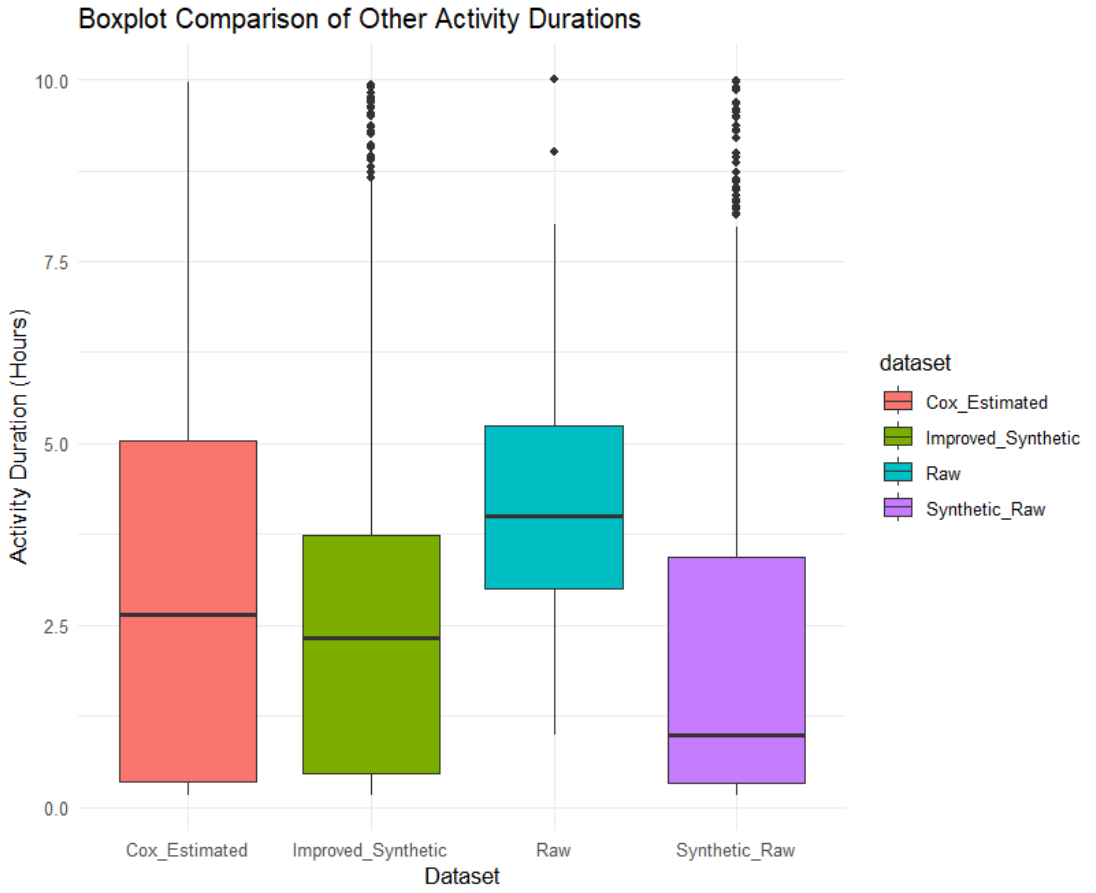
Στη συγκεκριμένη περίπτωση είναι πιο ξεκάθαρα τα πράγματα, τα συνθετικά μέσω Cox with Frailty είναι πιο κοντά στα αληθινά σε σχέση με όλα τα υπόλοιπα συνθετικά και από άποψη IQR και Median. Όσον αφορά τα outliers παρατηρούμε στα συνθετικά ακραίες και μη λογικές τιμές οι οποίες βέβαια είναι αποτέλεσμα της προσομοίωσης και όχι της στατιστικής λειτουργίας των cox μοντέλων

5.3.3. Other & Service Durations Comparison

Πίνακας 10. Αληθινά vs συνθετικά δεδομένα για λοιπές μετακινήσεις

Other Comparison real Vs synthetic data			
Metrics	Raw vs Synthetic Raw	Raw vs Improved Synthetic	Raw vs Simplecox_synthetic
K-S test			

D	0.47753	0.35482	0.2801
p-value	8.421e-08	0.0001707	0.005868
MAE	4.755	4.3898	4.5188
MSE	6.1339	5.8461	5.9359



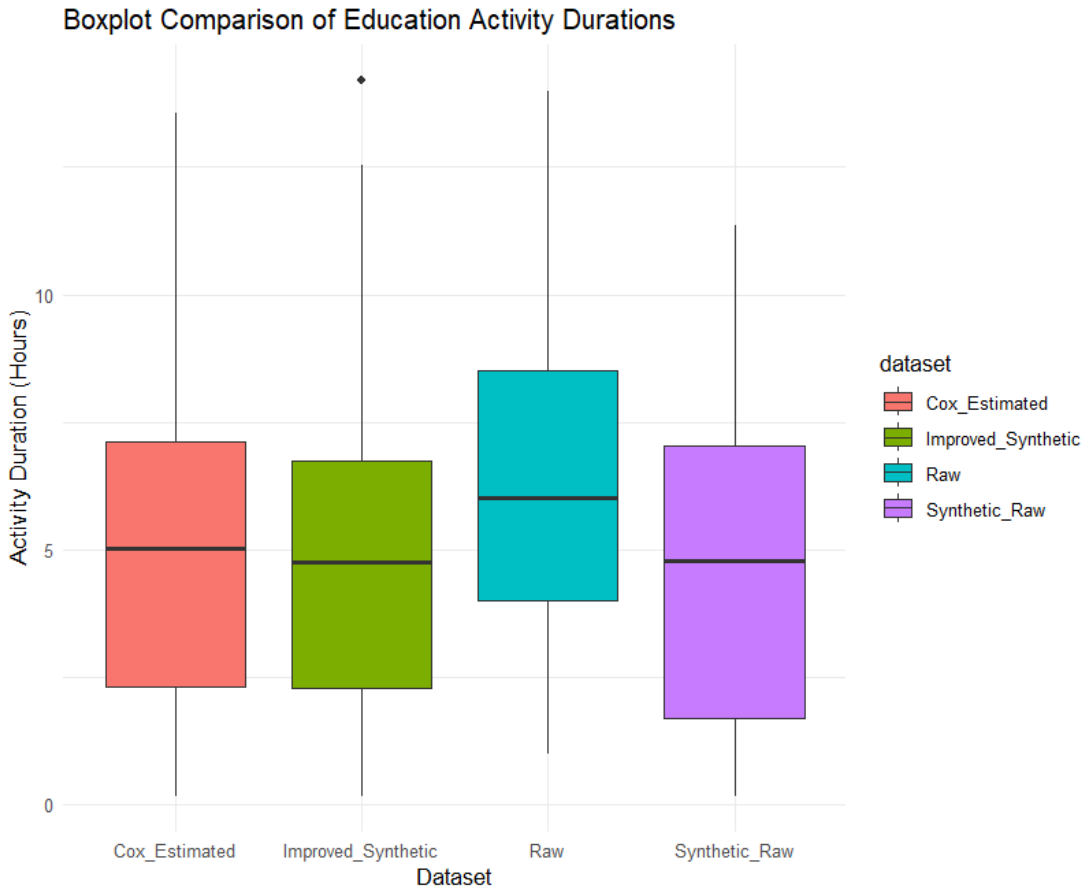
Εικόνα 11. Boxplot σύγκριση αληθινών vs συνθετικών δεδομένων για λοιπές μετακινήσεις

Ενώ τα στατιστικά τεστ είναι μοιρασμένα ως προς τα συνθετικά με τα cox μοντέλα από το boxplot φαίνεται ότι τα συνθετικά μέσω του απλού cox υπερισχύουν σε ακρίβεια σε σχέση με τα άλλα συνθετικά που εξετάζουμε αφού πρόκειται και για κοντινότερη τιμή διαμέσου, και για καλύτερη διασπορά και για μη ύπαρξη ακραίων τιμών – σφαλμάτων.

5.3.4. Education Durations Comparison

Πίνακας 11. Αληθινά vs συνθετικά δεδομένα για εκπαιδευτικές μετακινήσεις

Education Comparison real Vs synthetic data			
Metrics	Raw vs Synthetic Raw	Raw vs Improved Synthetic	Raw vs Simplecox synthetic
K-S test			
D	0.25078	0.18736	0.2253
p-value	0.000972	0.1031	0.02735
MAE	4.047721	3.951568	3.91789
RMSE	5.132352	5.040041	5.062376



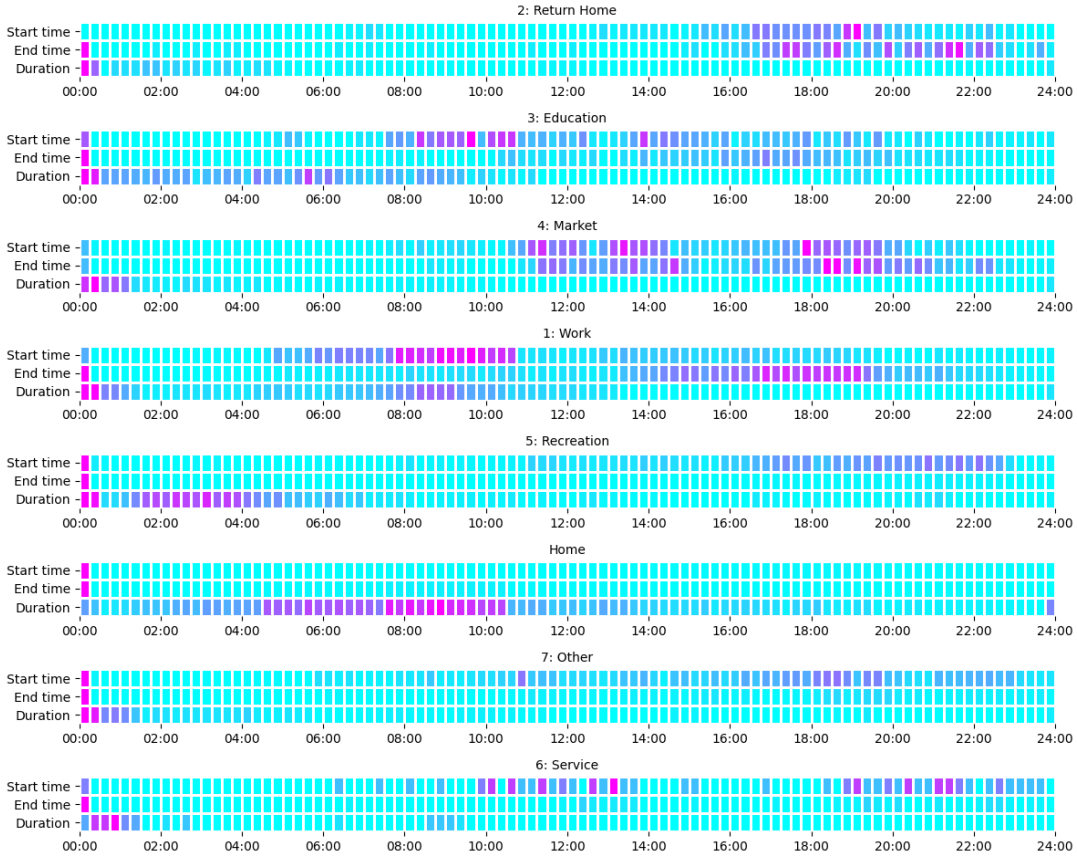
Εικόνα 12. Boxplot σύγκριση αληθινών vs συνθετικών δεδομένων για εκπαιδευτικές μετακινήσεις

Και στη συγκεκριμένη περίπτωση οι διαφορές που παρατηρούνται δεν είναι μεγάλες, κάτι αναμενόμενο καθώς οι επεμβάσεις σε εκπαιδευτικές μετακινήσεις ήταν λιγοστές. Παρ ’όλα αυτά με τα κριτήρια που θέσαμε για τα boxplot διαγράμματα όλα έχουν

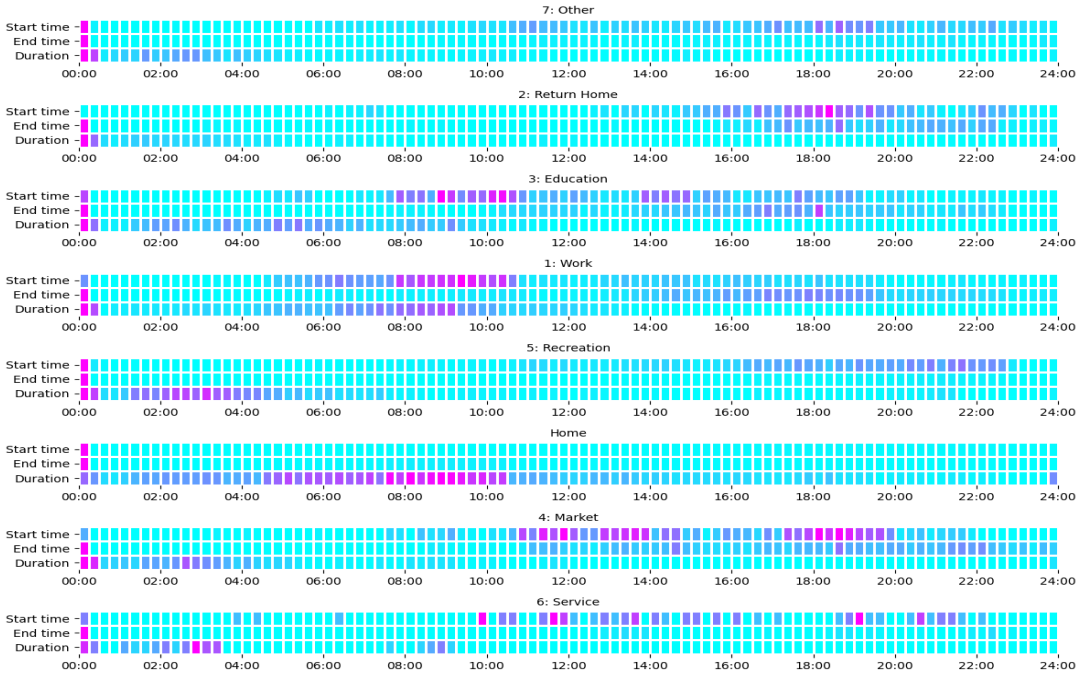
κοινές διαμέσους. Επίσης τα δύο συνθετικά που δημιουργήθηκαν με την έρευνα μέσω cox μοντέλων φαίνεται να ταιριάζουν καλύτερα στο IQR, δηλαδή στη διασπορά των διαρκειών. Τέλος, όσον αφορά τις μεγαλύτερες τιμές που δεν παρατηρούνται συχνά, τα συνθετικά μέσω απλού cox περιγράφουν καλύτερα τα αληθινά δεδομένα.
Αποδεικνύοντας λοιπόν πως το cox μοντέλο με frailty όρους με τις 2 latent classes για ετερογένεια είναι καλύτερο προχωράμε και στις αλλαγές που παρατηρήθηκαν και από την κυκλοφοριακή ανάλυση και από τα μοτίβα δραστηριοτήτων που εξάχθηκαν μέσω της PAM από το AthensPop. Πιο διεξοδικά, θα αναλυθούν οι διαφορές που παρατηρήθηκαν και σε επίπεδα κυκλοφοριακής ανάλυσης αλλά και στα μοτίβα δραστηριοτήτων. Τα δεδομένα που εξάγονται αφορούν, αριθμό ταξιδιών ανά φύλο/εισόδημα/ηλικιακή ομάδα, ποσοστιαίες κατανομές ανά ώρα, καθώς και την κατανομή των σκοπών μετακίνησης.

5.4. Μοτίβα Δραστηριοτήτων & Κυκλοφοριακή Ανάλυση

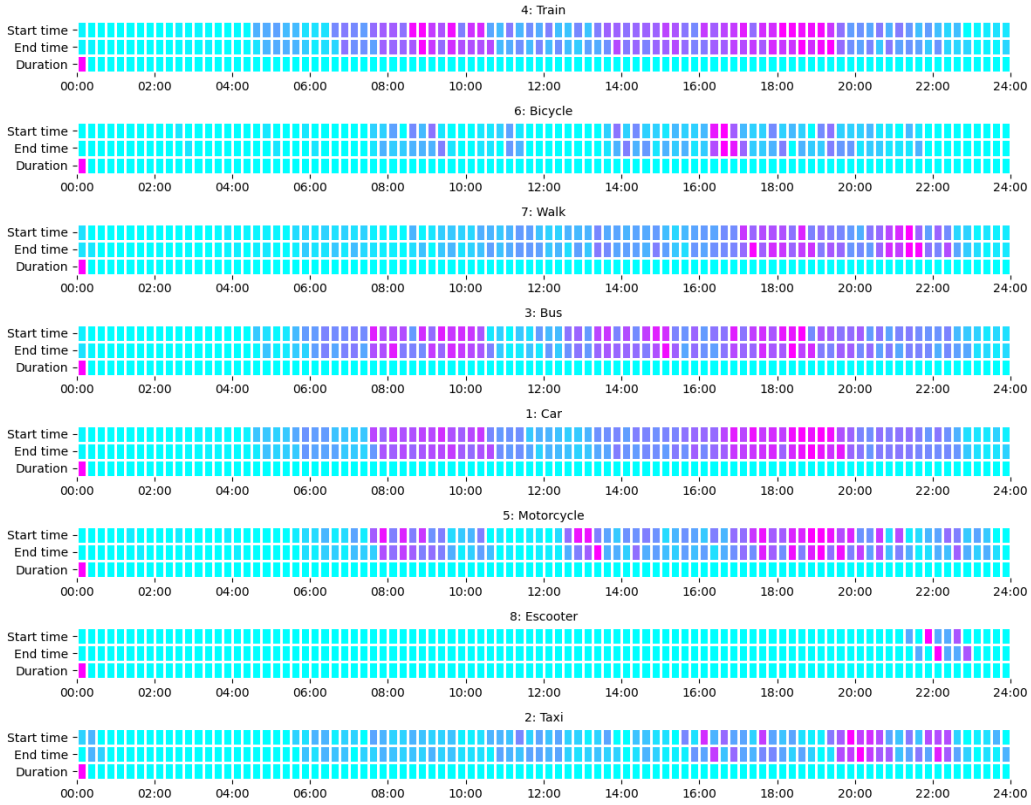
Στη συγκεκριμένη ενότητα θα παρουσιαστούν τα ημερήσια χρονοδιαγράμματα για όλους τους διαφορετικούς σκοπούς μετακινήσεων και θα παρουσιαστούν οι διαφορές των αποτελεσμάτων. Πιο διεξοδικά, παρακάτω απεικονίζονται οι ωριαίες συχνότητες μετακινήσεων ανά σκοπό μετακίνησης (Εικόνα 13 & 14) και ανά μέσο μετακίνησης (Εικόνα 15 & 16) (Διαφορετικά συνθετικά δεδομένα παραγόμενα από διαφορετικές έρευνες – δεδομένα εισόδου). Ανά γραμμή απεικονίζονται κατά σειρά: Ώρα έναρξης μετακίνησης, ώρα λήξης μετακίνησης και διάρκεια μετακίνησης. Όσο πιο γαλάζιο σημαίνει πως δεν παρατηρούνται μετακινήσεις για αυτόν τον σκοπό εκείνες τις ώρες, όσο πιο ροζ σημαίνει πυκνότερες μετακινήσεις.



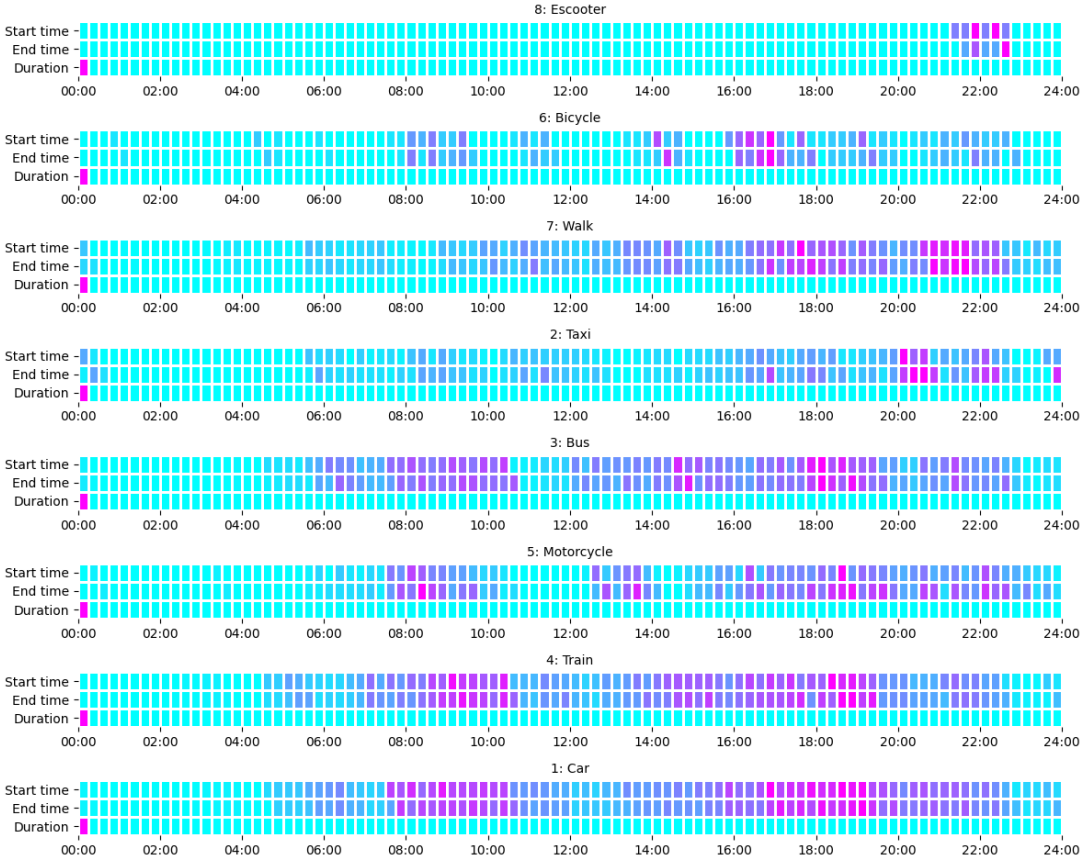
Εικόνα 13. Ημερήσιο χρονοδιάγραμμα ανά σκοπό μετακίνησης μέσω της ανεπεξέργαστης έρευνας



Εικόνα 14. Ημερήσιο χρονοδιάγραμμα ανά σκοπό μετακίνησης μέσω της έρευνας συμπληρωμένης με coxfrailty



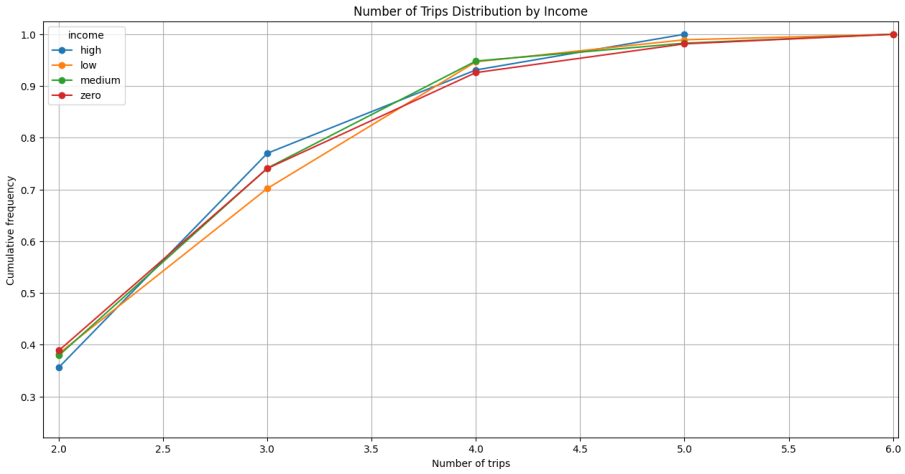
Εικόνα 15. Ημερήσιο χρονοδιάγραμμα χρήσης μέσων ανά σκοπό μετακίνησης μέσω της ανεπεξέργαστης έρευνας



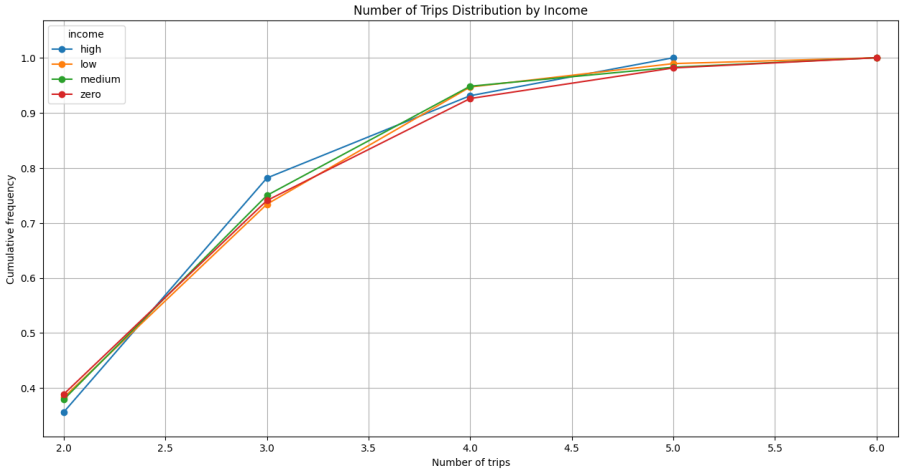
Εικόνα 16. Ημερήσιο χρονοδιάγραμμα χρήσης μέσων ανά σκοπό μετακίνησης μέσω της έρευνας συμπληρωμένης με coxfrailty

Από το μοντέλο εξάγονται και κάποιες οπτικοποιήσεις δεδομένων κυκλοφοριακής ανάλυσης. Πιο αναλυτικά, θα αποτυπωθούν οι διαφορές σε κάποια διαγράμματα κυκλοφοριακής ανάλυσης με τα αποτελέσματα που έχουν προέλθει από το AthensPop χωρίς κάποια παρέμβαση και το AthensPop βελτιωμένο με το cox μοντέλο με frailty όρους.

Ξεκινώντας, παρουσιάζεται ο αριθμός ημερήσιων ταξιδιών ανά κατηγορία εισοδήματος. Εδώ οι κατηγορίες εισοδήματος είναι χωρισμένες σε 4 κατηγορίες (Μηνιαίο εισόδημα μικρότερο των 750€, 750€-1500€, 1500€-2500€, 2500€ ή περισσότερα). Στον άξονα y υπολογίζεται η σωρευτική συχνότητα των παρατηρήσεων ανά εισοδηματική κατηγορία. Πιο αναλυτικά, αφορά το ποσοστό των ατόμων ανά κατηγορία που κάνουν τον αριθμό μετακινήσεων του άξονα x. Σε κάθε παραπάνω μετακίνηση το ποσοστό αυξάνεται έως ότου φτάσει στο 1 (100%) αφού κάθε άτομο που έχει μετακινηθεί 3 φορές έχει μετακινηθεί σίγουρα και 2 κ.ο.κ.

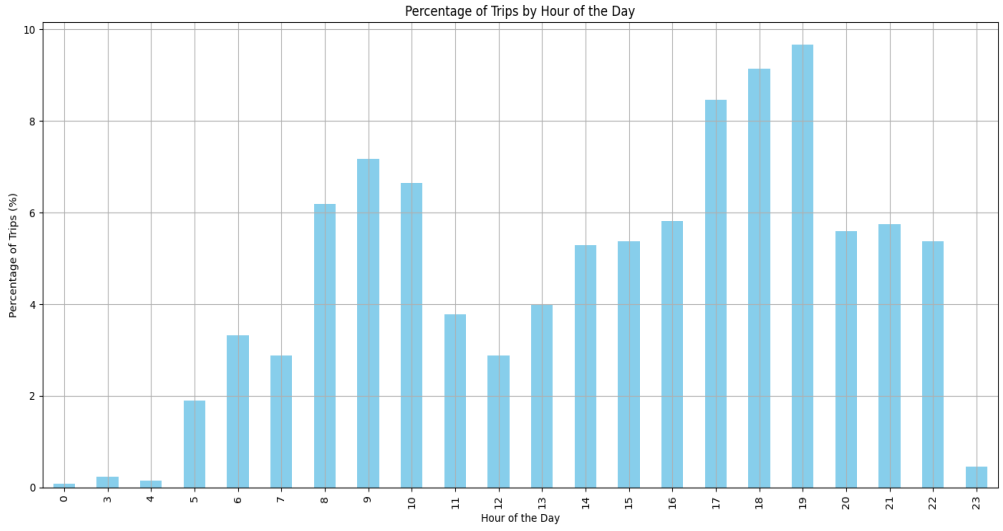


Εικόνα 17. Αριθμός ημερήσιων ταξιδιών ανά κατηγορία εισοδήματος πριν

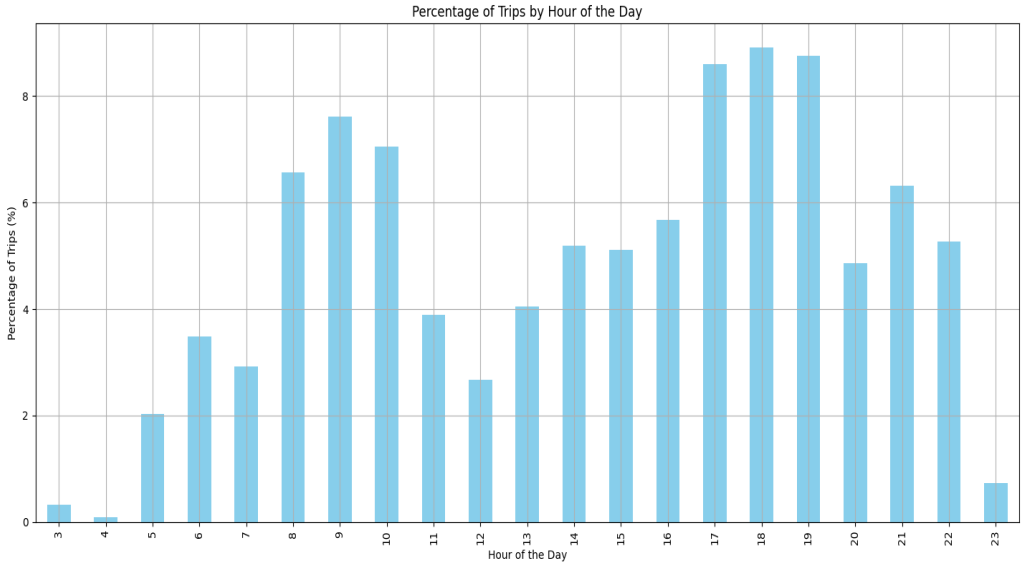


Εικόνα 18. Αριθμός Ταξιδιών ανά κατηγορία εισοδήματος μετά

Στη συνέχεια παρουσιάζεται ένα διάγραμμα bar με το ποσοστό ημερήσιων ταξιδιών ανά ώρα της ημέρας. Το συγκεκριμένο έχει πιο απλή ανάνγωση, στον άξονα y βρίσκεται το ποσοστό των συνολικών ταξιδιών μίας ημέρας ενώ στον άξονα x βρίσκονται οι ώρες της ημέρας. Οι διαφορές παρατηρούνται κατά κύριο λόγο στις απογευματινές και βραδινές ώρες καθώς η παρέμβαση του fix_no_back_home είναι η μεγαλύτερη αλλαγή στην έρευνα που χρησιμοποιείται ως δεδομένα εισόδου και αφορά αυτές τις ώρες.

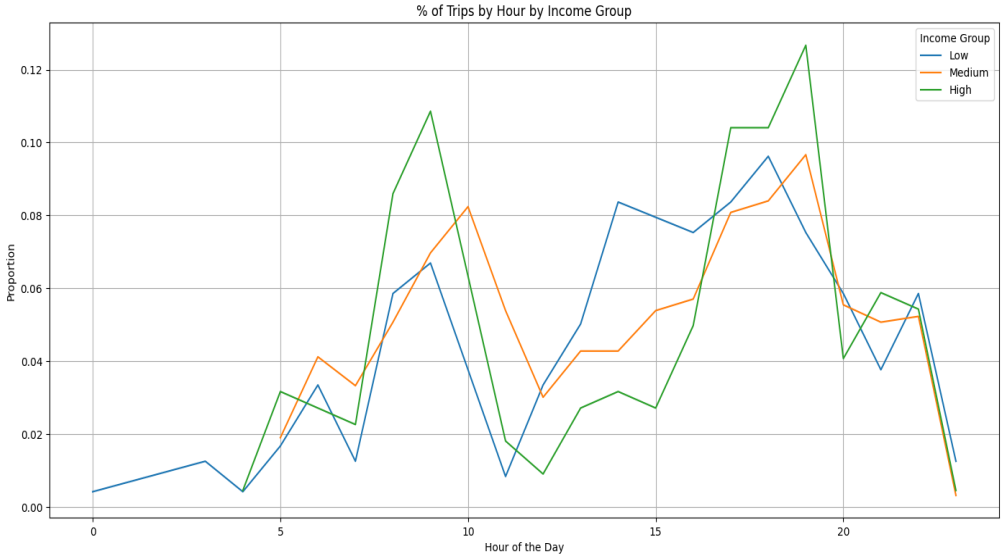


Εικόνα 19. Ποσοστό ημερήσιων ταξιδιών ανά ώρα έναρξης μετακινήσεων πριν

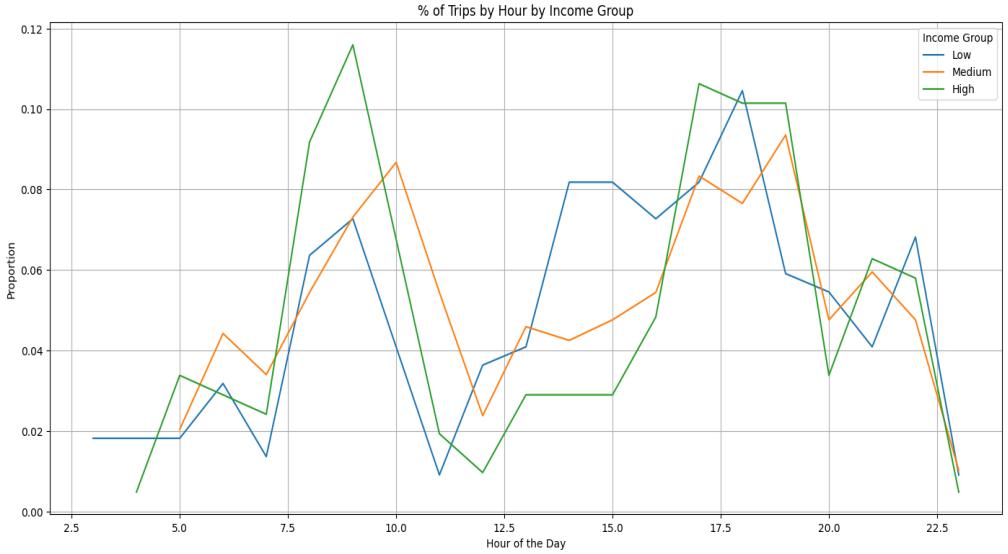


Εικόνα 20. Ποσοστό ημερήσιων ταξιδιών ανά ώρα έναρξης μετακινήσεων μετά

Στις εικόνες 21 και 22 αποεικονίζεται το ποσοστό ταξιδιών ανά ώρα ανά εισοδηματική κατηγορία. Στον άξονα y φαίνεται το ποσοστό των ημερήσιων ταξιδιών ανά ώρα ενώ κάθε καμπύλη αντιπροσωπεύει διαφορετική εισοδηματική κατηγορία (0-750€, 750-2500€, > 2500€).

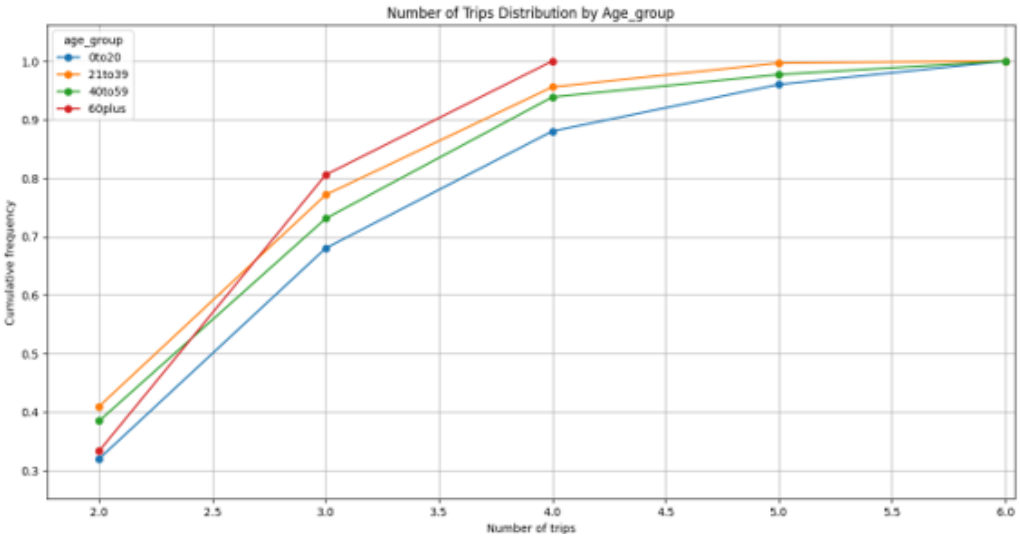


Εικόνα 21. Ποσοστό ημερήσιων ταξιδιών ανά ώρα έναρξης ανά κατηγορία εισοδήματος πριν

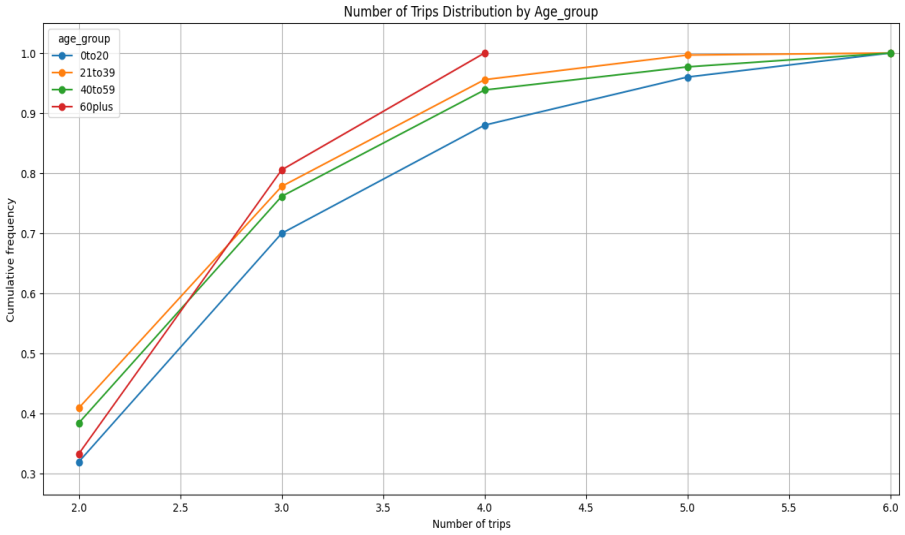


Εικόνα 22. Ποσοστό ημερήσιων ταξιδιών ανά ώρα έναρξης ανά κατηγορία εισοδήματος μετά

Ένα τελευταίο διάγραμμα που εξετάζεται αφορά την κατανομή του ημερήσιου αριθμού μετακινήσεων ανά ηλικιακή ομάδα (Εικόνες 23 & 24). Στον άξονα y παρουσιάζεται η σωρευτική συχνότητα όπως και προηγουμένως ενώ στον αξονα x ο αριθμός των ημερήσιων μετακινήσεων για κάθε άτομο.



Εικόνα 23. Αριθμός ημερήσιων μετακινήσεων ανά ηλικιακή κατηγορία



Εικόνα 24. Αριθμός ημερήσιων μετακινήσεων ανά ηλικιακή κατηγορία μετά

6. Συμπεράσματα

6.1. Συμπεράσματα από τα αποτελέσματα της διαδικασίας

Μπορούμε λοιπόν να ισχυριστούμε με σιγουριά πως οι διάρκειες που λείπουν συμπληρώνονται με μεγαλύτερη ακρίβεια μέσω του cox frailty με 2 λανθάνουσες κατηγορίες σε σύγκριση με ένα απλό Cox μοντέλο και σίγουρα πολύ καλύτερα σε σύγκριση με την υπόθεση της κανονικής ή της Weibull κατανομής. Επίσης, αποδεικνύεται, όπως ήταν αναμενόμενο, πως πιο ακριβείς έρευνες (βάσεις δεδομένων) βελτιώνουν τα εξαγόμενα συνθετικά και εν συνεχεία προσφέρουν πιο ακριβή μοτίβα δραστηριοτήτων. Γενικότερα στα ημερήσια χρονοδιαγράμματα μετακινήσεων ανά σκοπό μετακίνησης και χρήσης μέσων με όλους τους διαφορετικούς τρόπους που ακολουθήθηκαν δεν παρατηρούνται μεγάλες διαφορές στη μεγάλη εικόνα. Όμως, μπορούμε να διακρίνουμε σημαντικές λεπτομέρειες κυρίως προς το τέλος της ημέρας καθώς η πλειοψηφία των παρεμβάσεων αφορά διάρκειες σε activity chains που δεν αναφερόταν η επιστροφή στο σπίτι στο τέλος της ημέρας.

Κάποια συγκεκριμένα συμπεράσματα ως προς την κυκλοφοριακή συμπεριφορά των

μετακινούμενων είναι:

1) Στις μετακινήσεις για ψώνια παρατηρούμε πως το εύρος διάρκειας μεγαλώνει καθώς υπάρχουν αρκετές μετακινήσεις που διαρκούν 2-3 ώρες κάτι το οποίο είναι πολύ φυσιολογικό. Όσον αφορά τις ώρες έναρξης και λήξης των μετακινήσεων αυτών παρατηρείται μια σχετικά αυξημένη δράση τις απογευματινές ώρες μετά τις 18:00 και έως τις 19:30.
2) Στις εργασιακές μετακινήσεις παρατηρείται μεγαλύτερη διασπορά στο εύρος της διάρκειας αλλά και στο εύρος στις ώρες λήξης μετακινήσεων σε αντίθεση με την ανεπεξέργαστη έρευνα που η λήξη συγκεντρώνεται γύρω από τις απογευματινές ώρες.
3) Σε λοιπές μετακινήσεις και μετακινήσεις υπηρεσιών εμφανίζεται ένα εύρος που δεν υπήρχε πριν στη διάρκεια τους και πιο συγκεκριμένα παρατηρούνται περισσότερες τέτοιες μετακινήσεις διάρκειας 1.5 έως 3 ωρών. Όσον αφορά τις ώρες έναρξης και λήξης των μετακινήσεων αυτές συμβαίνουν κυρίες απογευματινές ώρες, όμως συγκεκριμένα για τις μετακινήσεις υπηρεσιών εμφανίζεται μία αύξηση τις πρωινές ώρες.
4) Οι εκπαιδευτικές μετακινήσεις δεν εμφανίζουν μεγάλες διαφορές, πολύ πιθανόν γιατί υπήρξε ελάχιστη επέμβαση στον συγκεκριμένο σκοπό μετακινήσεων. Επίσης παρουσιάζουν ένα ακανόνιστο μοτίβο και στις ώρες έναρξης και λήξης τους αλλά και στις διάρκειες.
5) Σχετικά εντονότερη χρήση αυτοκινήτων και τρένου/μετρό κατά τις απογευματινές ώρες αιχμής
6) Λίγο αραιότερη χρήση μοτοσυκλετών και λεωφορείων κατά τις απογευματινές ώρες
7) Υψηλότερη χρήση των E-scooter και ποδηλάτων τις βραδινές ώρες
8) Εντονότερη δραστηριότητα με περπάτημα τις απογευματινές ώρες
9) Αύξηση των ποσοστιαίων και των απόλυτων μετακινήσεων συνολικά τις απογευματινές ώρες αιχμής 17:00-20:00 και 22:00-00:00 και σχετική πτώση στις 16:00 και στις 20:00
10) Κατηγορία υψηλού εισοδήματος συνδέεται πιθανότατα με τις τυπικές ώρες εργασίας καθώς τα υψηλότερα τους ποσοστά συγκεντρώνονται και το πρωί και το απόγευμα αναλόγως. Η κατηγορία χαμηλού εισοδήματος παρουσιάζει και ένα peak τις μεσημεριανές ώρες που δεν εμφανίζουν οι άλλες δύο κατηγορίες.

11) Κατηγορία υψηλού εισοδήματος συνδέεται επίσης με λιγότερες μετακινήσεις. Οι υπόλοιπες κατηγορίες αντιδρούν παρόμοια
12) Μεγαλύτερη ηλικιακή ομάδα αντιστοιχεί σε λιγότερα ταξίδια με το αντιστρόφως ανάλογο να ισχύει για τις μικρότερες ηλικιακές ομάδες όπως αναμενόταν

6.2. Σημαντικότητα αποτελεσμάτων

Γίνεται ευκόλως κατανοητό πως η επέμβαση στην έρευνα αποτελεί μία επέμβαση στη ρίζα της διαδικασίας όπως φαίνεται και στις εικόνες 1 & 4. Η έρευνα είναι ο πρωταρχικός θεμέλιος λίθος για την ανάπτυξη των μοντέλων αυτού του είδους. Έτσι, μια καλύτερη, ακριβέστερη με λιγότερα λάθη έρευνα μπορεί να μας εξασφαλίσει καλύτερα συνθετικά δεδομένα. Κάτι πολύ χρήσιμο καθώς η ύπαρξη προσομοιόσεων με ακρίβεια βοηθάει στη συνεχή μελέτη και διερεύνηση θεμάτων σχετικών με τα συστήματα μεταφοράς. Είτε αυτό πρόκειται για μοτίβα δραστηριοτήτων και κυλοφοριακές αναλύσεις όπως στη συγκεκριμένη περίπτωση, είτε πρόκειται για διερεύνηση αλλαγών ή νέων πολιτικών η καλή απεικόνιση του πληθυσμού είναι ένα πολύ σημαντικό εργαλείο. Η τεχνολογία και η σύγχρονη εποχή ωθούν τις ερευνητικές μεθόδους προς αυτήν την κατεύθυνση καθώς τέτοιες μέθοδοι όπως οι προσομοιώσεις προσφέρουν λύσεις χωρίς να απαιτούν μεγάλα έξοδα όπως π.χ. οι έρευνες ερωτηματολογίου. Με αυτόν τον τρόπο γίνεται κατανοητό πως η σημαντικότητα των αποτελεσμάτων είναι κομβική στην ανάπτυξη της έρευνας σε έναν τομέα των μεταφορών σχετικά αχαρτογράφητο, όπως τα μοτίβα δραστηριοτήτων. Τέλος, η αποκωδικοποίηση της συμπεριφοράς των μετακινούμενων όπως αναφέρθηκε και κατά την εισαγωγή είναι ένα πολυδιάστατο πρόβλημα το οποίο είναι πολύ δύσκολο να αποτυπωθεί και να ποσοτικοποιηθεί. Τέτοιες ερευνητικές προσπάθειες κάνουν σημαντικά βήματα προς αυτήν την κατεύθυνση με σκοπό τη βελτίωση των συστημάτων μεταφοράς και τα λιγότερα δυνατά ρίσκα. Βέβαια, τα περιθώρια βελτίωσης και επέκτασης είναι πολύ μεγάλα και θα αναλυθούν περαιτέρω στην επόμενη ενότητα.

6.3. Επέκταση έρευνας

Είναι πολύ σημαντικό να αναφερθεί πως για τη συγκεκριμένη εργασία

χρησιμοποιήθηκε μία μικρή έρευνα του ΕΜΠ που αριθμούσε περίπου 500 παρατηρήσεις και διαφορετικά daily activity chains. Προσθέτοντας πολυπλοκότητα στο μοντέλο με τις 2 λανθάνουσες κατηγορίες, που στην ουσία αντιπροσωπεύουν 2 μεταβλητές των μοντέλων για μη παρατηρήσιμα μοτίβα που επηρεάζουν τις διάρκειες (κρυφές ομάδες), γίνεται κατανοητό πως όσο περισσότερα δεδομένα δοθούν στο μοντέλο ως εξάσκηση τόσο ασφαλέστερα αποτελέσματα θα δώσει πίσω. Ειδικότερα εφόσον όλα τα μοντέλα γίνονται ξεχωριστά ανά σκοπό μετακίνησης αυτό μειώνει μεν την πολυπλοκότητα, μειώνει δε και τα δεδομένα σε τεράστιο βαθμό που γίνεται δύσκολο να φανερωθούν αυτά τα κρυφά μοτίβα.

Με το νέο Στρατηγικό Σχέδιο Μεταφορών του ΟΑΣΑ για το 2024 που αναμένεται να ολοκληρωθεί τα επόμενα χρόνια θα εμπεριέχει μία τεράστια έρευνα η οποία θα αποδειχθεί πολύ βοηθητική. Πιο αναλυτικά, η έρευνα θα διεξαχθεί σε 36.600 νοικοκυριά (1% κάθε δήμου , 71.200 μετακινούμενους σε όλα τα μέσα μαζικής μεταφοράς και στις τοποθεσίες υποδομών τους. Επίσης αναμένονται έρευνες σε εταιρείες εμπορευματικών μεταφορών, οδηγούς ταξί αλλά και 1.200 ερωτηματολόγια μέσω ατομικών συνεντεύξεων δεδηλωμένης προτίμησης.

Το πρόβλημα των δεδομένων και της συλλογής τους όπως αναφέρθηκε και στην αρχή είναι υπαρκτό και μεγάλο. Τέτοιες έρευνες «λύνουν τα χέρια» των μηχανικών και δρουν υπέρ της σωστής πρόβλεψης και του ακριβή σχεδιασμού. Αξίζει να αναφερθεί πως τελευταία τέτοια έρευνα για τη μητροπολιτική Αθήνα συναντά κανείς το 2009 στο προηγούμενο Στρατηγικό Σχέδιο Μεταφορών Αττικής του ΟΑΣΑ. Στη συγκεκριμένη περίπτωση η διαφορά ανάμεσα σε 500 παρατηρήσεις της υπάρχουσας έρευνας και 70.000 του ΣΣΜΑ είναι προφανώς τεράστια. Γίνεται κατανοητό πως με μεγαλύτερες έρευνες δίνεται το περιθώριο της παρατήρησης αυτών των κρυφών ομάδων για τις οποίες επεκτάθηκε το cox μοντέλο. Βεβαίως υπάρχει το περιθώριο και της αύξησης αυτών των ομάδων (π.χ. από 2 σε 4) αλλά και της βελτίωσης των μεταβλητών με πληρέστερη έρευνα. (π.χ. στη συγκεκριμένη εργασία δε γίνεται κανένας διαχωρισμός σε μέρες της εβδομάδας / δεν υπάρχουν πληροφορίες νοικοκυριού).

Άξια αναφοράς είναι και η εφαρμογή αυτών των μοντέλων σε λογισμικό ανοιχτού κώδικα. Πιο αναλυτικά, όπως παρουσιάστηκε και στη διαδικασία λειτουργίας του AthensPop η βιβλιοθήκη PAM χρησιμοποιείται για τη σύνθεση του πληθυσμού και την παραγωγή των μετακινήσεων αυτού του πληθυσμού. Κάτι τέτοιο είναι αρκετά σημαντικό καθώς η χρήση σε λογισμικό ανοιχτού κώδικα διευκολύνει την

ενσωμάτωση του cox with frailty μοντέλου και γενικότερα σε agent based μοντέλα και στο AthensPop συγκεκριμένα. Επίσης, η ανοιχτή φύση του λογισμικού επιτρέπει την προσαρμογή των μοντέλων στις ανάγκες κάθε εφαρμογής ενώ παρέχει ευελιξία για μελλοντική εξέλιξη και επαναχρησιμοποίηση των μεθόδων. Τέτοιες προσπάθειες προάγουν τη διαφάνεια, τη συνεργασία και την καινοτομία στη μοντελοποίηση των μοτίβων δραστηριοτήτων και στην κυκλοφοριακή ανάλυση.

7. Βιβλιογραφικές Αναφορές

Ahmed, B. (2012). The Traditional Four Steps Transportation Modeling Using Simplified Transport Network: A Case Study of Dhaka City, Bangladesh

Agafonov, A., & Myasnikov, V. (2021). *Short-Term Traffic Flow Prediction in a Partially Connected Vehicle Environment*.

Annesha Enam et al. (2020) Hazard-Based Model of Activity Generation Using Vehicle Trajectory Data

Aurore Sallard et al. (2023). Travel demand generation using Bayesian Networks: an application to Switzerland

B.Y. He et al. (2021). A validated multi-agent simulation test bed to evaluate congestion pricing policies on population segments by time-of-day in New York City

B.Y. He et al. (2020). Evaluation of city-scale built environment policies in New York City with an emerging-mobility-accessible synthetic population

Chandra R Bhat (1996) A HAZARD-BASED DURATION MODEL OF SHOPPING ACTIVITY WITH NONPARAMETRIC BASELINE SPECIFICATION AND NONPARAMETRIC CONTROL FOR UNOBSERVED HETEROGENEITY

C. Chen et al (2016) The promises of big data and small data for travel behavior (aka human mobility) analysis

Ç. Tozluoglu, S. Dhamal and S. Yeh et al. (2023). A synthetic population of Sweden: datasets of agents, households, and activity-travel patterns

Chunguang Liu et al. (2023). Activity Duration under the COVID-19 Pandemic: A Comparative Analysis among Different Urbanized Areas Using a Hazard-Based Duration Model

Dominik Ziemke et al. (2019). The MATSim Open Berlin Scenario: A multimodal agent-based transport simulation scenario based on synthetic demand modeling and open data

Gruyer, D., Orfila, O., Glaser, S., Hedhli, A., Hautière, N., & Rakotonirainy, A. (2021). Are connected and automated vehicles the silver bullet for future transportation challenges? Benefits and weaknesses on safety, consumption, and traffic congestion. *Frontiers in Sustainable Cities, 2*, 607054.

G. Jovicic, C.O. Hansen (2003). A passenger travel demand model for Copenhagen

J. Urban Plann (1998) Integrated Modeling of Urban Shopping Activities

J.L. Yee, D.A. (2000) Niemeier Analysis of activity duration using the Puget sound transportation panel

Jingjun Li et al. (2023). A Stepwise Approach of Generating Agent-based Simulation Model for Brussels Using Ubiquitous Big Data

J.W. Joubert (2018). Synthetic populations of South African urban areas

Jingjun Li et al. (2023). A Stepwise Approach of Generating Agent-based Simulation Model for Brussels Using Ubiquitous Big Data

K. Gkiotsalitis, A. Stathopoulos (2015). A utility-maximization model for retrieving users' willingness to travel for participating in activities from big-data

Konstantinos Gkiotsalitis (2020). Predicting Traveling Distances and Unveiling Mobility and Activity Patterns of Individuals from Multisource Data

Luis A. Guzman et al. (2017). A Strategic Tour Generation Modeling within a Dynamic Land-Use and Transport Framework: A Case Study of Bogota, Colombia

M. Aljoufie et al. (2013). A cellular automata-based land use and transport interaction model applied to Jeddah, Saudi Arabia

M.A. Munizaga, C. Palma (2012). Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile

Nebiyou Tilahun David Levinson (2009) Contacts and Meetings: Location, Duration and Distance Traveled

Pauline van den Berg et al (2012) A latent class accelerated hazard model of social activity duration

Sreela P.Ka et al. (2013). Modeling of Shopping Participation and Duration of Workers in Calicut

Sebastian Hörl & Milos Balac (2021). Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data

Y. Wang et al. (2015). Assessing the accessibility impact of transport policy by a land-use and transport interaction model – The case of Madrid

Yixiao Li et al. (2019). Analysis of Spatial and Temporal Characteristics of Citizens' Mobility Based on E-Bike GPS Trajectory Data in Tengzhou City, China

Y.-L. Chu et al. (2019) Modeling workers' daily out-of-home maintenance activity participation and duration

8. Παράρτημα

Στο παράρτημα θα ακολουθήσουν ενδεικτικά κάθε ξεχωριστό script της διαδικασίας για έναν σκοπό μετακίνησης. Ο κώδικας θα ακολουθεί τη συλλογιστική πορεία από την επεξεργασία της έρευνας έως την εκτίμηση διαρκειών και τις συγκρίσεις και είναι γραμμένος σε R (έκδοση 4.4.2).

simplecox_other

```
# Load required libraries
library(survival)  # For survival analysis functions
library(dplyr)     # For data manipulation
library(car)       # For multicollinearity checks (VIF)
library(tidyr)     # For data tidying and reshaping

# Set file path and load data
file_path <- "D:/Downloads/other_data.csv"
data <- read.csv(file_path)

# Replace "NA" strings with actual NA values in the dataset
data[data == "NA"] <- NA

# Apply mappings for categorical variables
gender_map <- c('0: feamale' = 'female', '1: male' = 'male')
education_map <- c('1: primary school' = 'primary', '2: high school' = 'secondary',
                   '3: bachelor' = 'bachelor', '4: master or phd' = 'tertiary')
employment_map <- c('1: inactive' = 'inactive', '2: unemployed' = 'unemployed',
                    '3: student' = 'student', '4: active' = 'active')
car_own_map <- c('0: no' = 'no', '1: yes' = 'yes')
income_map <- c('0: no income' = 0, '1: 750 or less' = 1, '2: 750-1500' = 2,
                '3: 1500-2500' = 3, '4: 2500 or more' = 4)
modes_map <- c('1: car' = 'car', '2: taxi' = 'car', '3: bus' = 'bus',
               '4: train' = 'rail', '5: motorcycle' = 'motorcycle',
               '6: bicycle' = 'bike', '7: walk' = 'walk', '8: escooter' = 'escooter')
zone_map <- c('1' = 'Central Athens', '2' = 'Central Athens', '3' = 'Central Athens',
              '4' = 'Central Athens', '5' = 'Central Athens', '6' = 'Central Athens',
              '7' = 'Central Athens', '13' = 'Central Athens', '14' = 'Central Athens',
              '15' = 'Central Athens', '16' = 'Central Athens', '20' = 'Central Athens',
              '26' = 'Central Athens', '8' = 'West Athens', '9' = 'West Athens',
              '11' = 'West Athens', '17' = 'West Athens', '19' = 'West Athens',
              '32' = 'West Athens', '10' = 'East Attica', '36' = 'East Attica',
              '12' = 'East Attica', '18' = 'South Athens', '21' = 'South Athens',
              '23' = 'South Athens', '25' = 'South Athens', '28' = 'South Athens',
              '22' = 'North Athens', '24' = 'North Athens', '29' = 'North Athens',
              '34' = 'North Athens', '35' = 'North Athens', '27' = 'Piraeus',
```

```
              '30' = 'Piraeus', '31' = 'Piraeus', '33' = 'Piraeus')

# Apply mappings
data$gender <- gender_map[data$gender]
data$employment <- employment_map[data$employment]
print(data$education)
data$car_own <- car_own_map[data$car_own]
data$income <- income_map[data$income]
data$mode <- modes_map[data$mode]
data$dest <- zone_map[as.character(data$dest)]

# Ensure 'activity_dur' is numeric and replace blanks with NA
data$activity_dur[data$activity_dur == ""] <- NA
data$activity_dur <- as.numeric(data$activity_dur)

# Ensure 'dist' is numeric
data$dist <- as.numeric(data$dist)

# Convert 'income' to a categorical variable
data$income <- factor(data$income,
                      levels = c(0, 1, 2, 3, 4),
                       labels = c('no income', '750 or less', '750-1500', '1500-2500',
'2500 or more'))

# Convert 'dest' to a categorical variable
data$dest <- factor(data$dest,
                    levels = c('Central Athens', 'West Athens', 'East Attica',
                               'South Athens', 'North Athens', 'Piraeus'))

# Standardize continuous variables (dist remains numeric)
data$age <- scale(data$age, center = TRUE, scale = TRUE)
data$time <- scale(data$time, center = TRUE, scale = TRUE)
data$dist <- scale(data$dist, center = TRUE, scale = TRUE)

# Filter complete cases
data <- data %>%
  filter(!is.na(activity_dur) & !is.na(event)) %>%
  drop na(gender, education, employment, car own, income, dest, dist, mode)

# Check processed data
str(data)
summary(data)

# Fit Cox proportional hazards model
cox_model <- coxph(Surv(activity_dur, event) ~ gender + education + income +
                     car_own + dist + time + age,
                   data = data)

# Display model summary
summary(cox_model)

# Calculate AIC for model comparison
aic_no_frailty <- AIC(cox_model)
cat("AIC without frailty (with time and categorical dest):", aic_no_frailty, "\n")

# Check multicollinearity using VIF
vif_values <- vif(cox_model)
cat("Variance Inflation Factors (VIF):\n")
print(vif_values)
```

Cox with frailty for work trips

```r
# Load required libraries
library(survival)  # For survival analysis functions
library(dplyr)     # For data manipulation
library(car)       # For multicollinearity checks (VIF)
library(tidyr)     # For data tidying and reshaping

# Set file path and load data
file_path <- "D:/Downloads/other_data.csv"
data <- read.csv(file_path)

# Replace "NA" strings with actual NA values in the dataset
data[data == "NA"] <- NA

# Apply mappings for categorical variables
gender_map <- c('0: feamale' = 'female', '1: male' = 'male')
education_map <- c('1: primary school' = 'primary', '2: high school' = 'secondary',
                   '3: bachelor' = 'bachelor', '4: master or phd' = 'tertiary')
employment_map <- c('1: inactive' = 'inactive', '2: unemployed' = 'unemployed',
                    '3: student' = 'student', '4: active' = 'active')
car_own_map <- c('0: no' = 'no', '1: yes' = 'yes')
income_map <- c('0: no income' = 0, '1: 750 or less' = 1, '2: 750-1500' = 2,
                '3: 1500-2500' = 3, '4: 2500 or more' = 4)
modes_map <- c('1: car' = 'car', '2: taxi' = 'car', '3: bus' = 'bus',
               '4: train' = 'rail', '5: motorcycle' = 'motorcycle',
               '6: bicycle' = 'bike', '7: walk' = 'walk', '8: escooter' = 'escooter')
zone_map <- c('1' = 'Central Athens', '2' = 'Central Athens', '3' = 'Central Athens',
              '4' = 'Central Athens', '5' = 'Central Athens', '6' = 'Central Athens',
              '7' = 'Central Athens', '13' = 'Central Athens', '14' = 'Central Athens',
              '15' = 'Central Athens', '16' = 'Central Athens', '20' = 'Central Athens',
              '26' = 'Central Athens', '8' = 'West Athens', '9' = 'West Athens',
              '11' = 'West Athens', '17' = 'West Athens', '19' = 'West Athens',
              '32' = 'West Athens', '10' = 'East Attica', '36' = 'East Attica',
              '12' = 'East Attica', '18' = 'South Athens', '21' = 'South Athens',
              '23' = 'South Athens', '25' = 'South Athens', '28' = 'South Athens',
              '22' = 'North Athens', '24' = 'North Athens', '29' = 'North Athens',
              '34' = 'North Athens', '35' = 'North Athens', '27' = 'Piraeus',
              '30' = 'Piraeus', '31' = 'Piraeus', '33' = 'Piraeus')

# Apply mappings
data$gender <- gender_map[data$gender]
data$employment <- employment_map[data$employment]
print(data$education)
data$car_own <- car_own_map[data$car_own]
data$income <- income_map[data$income]
data$mode <- modes_map[data$mode]
data$dest <- zone_map[as.character(data$dest)]

# Ensure 'activity_dur' is numeric and replace blanks with NA
data$activity_dur[data$activity_dur == ""] <- NA
data$activity_dur <- as.numeric(data$activity_dur)

# Ensure 'dist' is numeric
data$dist <- as.numeric(data$dist)

# Convert 'income' to a categorical variable
data$income <- factor(data$income,
                      levels = c(0, 1, 2, 3, 4),
                       labels = c('no income', '750 or less', '750-1500', '1500-2500',
'2500 or more'))

# Convert 'dest' to a categorical variable
data$dest <- factor(data$dest,
                    levels = c('Central Athens', 'West Athens', 'East Attica',
                               'South Athens', 'North Athens', 'Piraeus'))

# Standardize continuous variables (dist remains numeric)
data$age <- scale(data$age, center = TRUE, scale = TRUE)
data$time <- scale(data$time, center = TRUE, scale = TRUE)
data$dist <- scale(data$dist, center = TRUE, scale = TRUE)

# Filter complete cases
data <- data %>%
  filter(!is.na(activity_dur) & !is.na(event)) %>%
  drop_na(gender, education, employment, car_own, income, dest, dist, mode)
```

```r
# Check processed data
str(data)
summary(data)

# Fit Cox proportional hazards model
cox_model <- coxph(Surv(activity_dur, event) ~ gender + education + income +
                     car_own + dist + time + age,
                   data = data)

# Display model summary
summary(cox_model)

# Calculate AIC for model comparison
aic_no_frailty <- AIC(cox_model)
cat("AIC without frailty (with time and categorical dest):", aic_no_frailty, "\n")

# Check multicollinearity using VIF
vif_values <- vif(cox_model)
cat("Variance Inflation Factors (VIF):\n")
print(vif_values)
```

Stan file for frailty model

```
data {
  int<lower=0> N;                  // Number of observations
  int<lower=1> K;                  // Number of predictors
  vector[N] time;                  // Survival times
  int<lower=0, upper=1> status[N]; // Censoring status (1 = event, 0 = censored)
  matrix[N, K] X;                  // Predictor matrix
}

parameters {
  vector[K] beta[2];               // Coefficients for each latent class
  vector[2] z_frailty;             // Latent variables for frailty terms (non-centered)
  real<lower=0> sigma_frailty;     // Standard deviation of frailty terms
  simplex[2] class_probs;          // Mixing proportions for the latent classes
}

transformed parameters {
  vector[2] frailty;               // Non-centered frailty terms
  frailty = z_frailty * sigma_frailty;  // Transform latent frailty terms
}

model {
  // Priors
  for (k in 1:2) {
    beta[k] ~ normal(0, 1);  // Weakly informative prior for coefficients
  }
  z_frailty ~ normal(0, 1);  // Standard normal prior for non-centered frailty
  sigma_frailty ~ normal(0, 1);  // Prior for frailty standard deviation
  class_probs ~ dirichlet(rep_vector(2.0, 2));  // Stronger prior for mixing proportions

  // Likelihood
  for (i in 1:N) {
    real log_lik_1 = weibull_lpdf(time[i] | 1 / sigma_frailty, exp(dot_product(X[i],
beta[1]) + frailty[1])) * status[i] +
                     weibull_lccdf(time[i] | 1 / sigma_frailty, exp(dot_product(X[i],
beta[1]) + frailty[1])) * (1 - status[i]);
    real log_lik_2 = weibull_lpdf(time[i] | 1 / sigma_frailty, exp(dot_product(X[i],
beta[2]) + frailty[2])) * status[i] +
                     weibull_lccdf(time[i] | 1 / sigma_frailty, exp(dot_product(X[i],
beta[2]) + frailty[2])) * (1 - status[i]);

    // Contribution to the log posterior (numerically stable log-sum-exp)
    target += log_sum_exp(
      log(class_probs[1]) + log_lik_1,
      log(class_probs[2]) + log_lik_2
    );
  }
}
```

```
generated quantities {
  vector[N] log_lik;  // Per-observation log-likelihood

  for (i in 1:N) {
    real log_lik_1 = weibull_lpdf(time[i] | 1 / sigma_frailty, exp(dot_product(X[i],
beta[1]) + frailty[1])) * status[i] +
                     weibull_lccdf(time[i] | 1 / sigma_frailty, exp(dot_product(X[i],
beta[1]) + frailty[1])) * (1 - status[i]);
    real log_lik_2 = weibull_lpdf(time[i] | 1 / sigma_frailty, exp(dot_product(X[i],
beta[2]) + frailty[2])) * status[i] +
                     weibull_lccdf(time[i] | 1 / sigma_frailty, exp(dot_product(X[i],
beta[2]) + frailty[2])) * (1 - status[i]);

    // Log-likelihood using log-sum-exp for numerical stability
    log_lik[i] = log_sum_exp(
      log(class_probs[1]) + log_lik_1,
      log(class_probs[2]) + log_lik_2
    );
  }
}
```

Estimating durations simple cox

```
# Load required libraries
library(dplyr)
library(survival)
library(pracma)

# Define paths (update with your actual paths)
file_path <- "D:/Downloads/market_data.csv"  # Data file
output_file <- "D:/Documents/simplecox_market_predicted_durations1.csv"  # Output file

# Load the dataset
data <- read.csv(file_path)

# Ensure all required variables exist
required_vars <- c("pid", "gender", "education", "income", "car_own", "dist", "time",
"activity_dur", "event")
if (!all(required_vars %in% colnames(data))) {
  stop("Dataset is missing required variables: ", paste(setdiff(required_vars,
colnames(data)), collapse = ", "))
}

# Replace "NA" strings and blanks with actual NA values
data[data == "NA"] <- NA
data[data == ""] <- NA

# Apply mappings for categorical variables
gender_map <- c('0: feamale' = 'female', '1: male' = 'male')
education_map <- c('2: high school' = 'secondary', '3: bachelor' = 'tertiary', '4:
master or phd' = 'tertiary')
income_map <- c('0: no income' = 0, '1: 750 or less' = 1, '2: 750-1500' = 2, '3: 1500-
2499' = 3, '4: 2500 or more' = 4)
car_own_map <- c('0: no' = 'no', '1: yes' = 'yes')

# Map and align variable levels
data$gender <- factor(gender_map[data$gender], levels = c("female", "male"))
data$education <- factor(education_map[data$education], levels = c("secondary",
"tertiary"))
data$income <- factor(income_map[data$income], levels = c(0, 1, 2, 3, 4))
data$car_own <- factor(car_own_map[data$car_own], levels = c("no", "yes"))

# Ensure continuous variables are numeric
data$activity_dur <- as.numeric(data$activity_dur)
data$dist <- as.numeric(data$dist)
data$time <- as.numeric(data$time)

# Handle missing data
data$gender[is.na(data$gender)] <- "female"  # Default gender
data$education[is.na(data$education)] <- "secondary"  # Default education
data$income[is.na(data$income)] <- 2  # Median income level
data$car_own[is.na(data$car_own)] <- "no"  # Default car ownership
```

```
data$dist[is.na(data$dist)] <- median(data$dist, na.rm = TRUE)
data$time[is.na(data$time)] <- median(data$time, na.rm = TRUE)

# Define PIDs for prediction
pids_to_predict <- c(113, 121, 173, 205, 206, 219, 226, 232, 244, 276, 280, 295, 308,
312, 318, 329,
                     334, 368, 403, 421, 429, 443, 450, 458, 463, 487, 488, 489, 494,
497, 511,
                     514, 515, 532, 533, 548, 558, 562, 573, 576, 577, 578, 580, 601)

# Filter data for prediction
data_to_predict <- data %>% filter(pid %in% pids_to_predict)

# Refactor variables to match model levels
data_to_predict$gender <- factor(data_to_predict$gender, levels = c("female", "male"))
data_to_predict$education <- factor(data_to_predict$education, levels = c("secondary",
"tertiary"))
data_to_predict$income <- factor(data_to_predict$income, levels = c(0, 1, 2, 3, 4))
data_to_predict$car_own <- factor(data_to_predict$car_own, levels = c("no", "yes"))

# Load the Cox model
cox_model <- coxph(Surv(activity_dur, event) ~ gender + education + income + car_own +
dist + time, data = data)

# Extract baseline hazard
baseline hazard <- basehaz(cox model, centered = FALSE)

# Predict linear predictor
linear_predictor <- predict(cox_model, newdata = data_to_predict, type = "lp")

# Calculate predicted durations
durations <- numeric(nrow(data to predict))
for (i in seq_len(nrow(data_to_predict))) {
  cat("\nProcessing PID:", data_to_predict$pid[i], "\n")

  if (is.na(linear_predictor[i])) {
    durations[i] <- NA
    next
  }

  risk_score <- exp(linear_predictor[i])
  cumulative_hazard <- baseline_hazard$hazard * risk_score
  survival_probs <- exp(-cumulative_hazard)
  duration_index <- which.max(survival_probs < 0.5)

  if (length(duration_index) == 0) {
    durations[i] <- NA
  } else {
    durations[i] <- baseline_hazard$time[duration_index]
  }
}

# Add predicted durations to the dataset
data_to_predict$predicted_activity_dur <- durations

# Save the results to a CSV file
write.csv(data_to_predict, file = output_file, row.names = FALSE)

cat("Predicted durations saved to:", output_file, "\n")
```

Estimating activity durations cox with frailty

```r
# Load required libraries
library(dplyr)
library(pracma)

# Load the dataset
file_path <- "D:/Downloads/education_data.csv"  # Replace with your dataset path
data <- read.csv(file_path)

# Ensure all required variables exist
required_vars <- c("pid", "gender", "education", "employment", "income", "car_own",
                   "mode", "dest", "time", "dist")
if (!all(required_vars %in% colnames(data))) {
  stop("Dataset is missing required variables:", paste(setdiff(required_vars,
colnames(data)), collapse = ", "))
}

# Filter rows with complete cases for required variables
data <- data[complete.cases(data[required_vars]), ]

# Scale continuous variables
data$time <- scale(data$time, center = TRUE, scale = TRUE)
data$dist <- scale(data$dist, center = TRUE, scale = TRUE)

# Define the PIDs for which you want predictions
pids_to_predict <- c(137, 142, 212, 297, 363, 399)  # Replace with your list of PIDs

# Subset the data for the given PIDs
data_to_predict <- data %>%
  filter(pid %in% pids_to_predict)

# Check if all PIDs exist in the data
missing_pids <- setdiff(pids_to_predict, data_to_predict$pid)
if (length(missing_pids) > 0) {
  cat("Warning: The following PIDs are not in the dataset:", missing_pids, "\n")
}

# Prepare the model matrix for the selected PIDs
X_predict <- model.matrix(~ gender + education + employment + income + car_own + mode +
dest + time + dist - 1, data = data_to_predict)

# Define beta coefficients dynamically to align with X_predict
beta_means <- matrix(c(
  # Example coefficients for latent class 1
  0.5319583,  0.4255160,  0.4036381,  0.34068976,  0.17002396,  0.12647604,  0.31377163,
0.295925816,
  0.45879381, 0.3200252, 0.35329534, 0.23834686, 0.32311831, 0.2695578,

  # Example coefficients for latent class 2
  0.2539367,  0.2320291,  0.0839978,  0.03289959,  0.04622145,  0.0242167,  0.01040276,  -
0.00621423,
  -0.002336634, -0.0224325, -0.004259964, -0.004207541, 0.08984353, 0.1015274
), nrow = 2, byrow = TRUE)

# Compute linear predictors for both latent classes
linear_predictor_class1 <- as.vector(X_predict %*% beta_means[1, ])
linear_predictor_class2 <- as.vector(X_predict %*% beta_means[2, ])

# Define frailty values
frailty_means <- c(0.1286201, 0.1003085)

# Add frailty term
linear_predictor_class1 <- linear_predictor_class1 + frailty_means[1]
linear_predictor_class2 <- linear_predictor_class2 + frailty_means[2]

# Load the baseline hazard
baseline_hazard <- read.csv("baseline_hazard_education.csv")  # Replace with the path
to your baseline hazard file
baseline_survival <- exp(-cumsum(baseline_hazard$hazard))
baseline_times <- baseline_hazard$time

# Compute adjusted hazards for each observation at all time points
adjusted_hazard_class1 <- sapply(linear_predictor_class1, function(lp) {
  exp(lp) * baseline_survival
})
adjusted_hazard_class2 <- sapply(linear_predictor_class2, function(lp) {
  exp(lp) * baseline_survival
```

```
})

# Compute probabilities for each class by normalizing adjusted hazards
class_prob1_adjusted           <-           colSums(adjusted_hazard_class1)           /
(colSums(adjusted_hazard_class1) + colSums(adjusted_hazard_class2))
class_prob2_adjusted <- 1 - class_prob1_adjusted

# Assign latent class dynamically based on highest probability
data_to_predict$latent_class <- ifelse(class_prob1_adjusted > class_prob2_adjusted, 1,
2)

# Compute survival probabilities for each PID
survival_probabilities <- sapply(1:nrow(data_to_predict), function(i) {
  exp(-baseline_survival    *    exp(ifelse(data_to_predict$latent_class[i]    ==    1,
linear_predictor_class1[i], linear_predictor_class2[i])))
})

# Calculate mean and median durations
mean_durations <- apply(survival_probabilities, 2, function(surv_curve) {
  trapz(baseline_times, surv_curve)
})
median_durations <- apply(survival_probabilities, 2, function(surv_curve) {
  if (any(surv_curve <= 0.5)) {
    baseline_times[which.min(abs(surv_curve - 0.5))]
  } else {
    NA
  }
})

# Add predicted durations to the dataset
data_to_predict$mean_duration <- mean_durations
data_to_predict$median_duration <- median_durations

# Save results to a CSV file
output_file <- "predicted_durations_education.csv"
write.csv(data_to_predict, output_file, row.names = FALSE)
cat("Predicted durations saved to:", output_file, "\n")

# Compute probabilities for each class by normalizing adjusted hazards
class_prob1_adjusted           <-           colSums(adjusted_hazard_class1)           /
(colSums(adjusted_hazard_class1) + colSums(adjusted_hazard_class2))
class_prob2_adjusted <- 1 - class_prob1_adjusted

# Print class probabilities for the first 10 PIDs
cat("Class probabilities for the first 10 PIDs:\n")
if (nrow(data_to_predict) >= 10) {
  print(data_to_predict[1:10, c("pid")])  # Print PIDs for reference
  probabilities <- data.frame(
    pid = data_to_predict$pid[1:10],
    class_prob1 = class_prob1_adjusted[1:10],
    class_prob2 = class_prob2_adjusted[1:10]
  )
} else {
  print(data_to_predict[, c("pid")])  # If fewer than 10 rows, print available rows
  probabilities <- data.frame(
    pid = data_to_predict$pid,
    class_prob1 = class_prob1_adjusted,
    class_prob2 = class_prob2_adjusted
  )
}

print(probabilities)
```

Comparison Real vs Synthetic data

```
# Load Required Libraries
library(dplyr)
library(ggplot2)
library(readr)

# Define File Paths
raw_file <- "D:/Downloads/raw_data_work.csv"     # Replace with path to raw data
synthetic_raw_file <- "D:/Documents/book12.csv"  # Replace with path to improved data
improved_synthetic_file  <-  "D:/Downloads/frailty_work_data_synthetic.csv"  #  Replace
with path to synthetic durations
cox_estimated_file <- "D:/Downloads/synthetic_work_data_simplecox.csv" # Replace with
path to Cox-estimated data

# Load Data
raw_data <- read_csv(raw_file)
synthetic_raw <- read_csv(synthetic_raw_file)
improved_synthetic <- read_csv(improved_synthetic_file)
cox estimated <- read csv(cox estimated file)

# Ensure Consistent Column Name for Duration
names(raw_data)[names(raw_data) == "duration_column_in_raw"] <- "activity_dur"
names(synthetic_raw)[names(synthetic_raw)   ==   "duration_column_in_synthetic_raw"]   <-
"activity_dur"
names(improved_synthetic)[names(improved_synthetic)                                    ==
"duration_column_in_improved_synthetic"] <- "activity_dur"
names(cox_estimated)[names(cox_estimated)      ==      "duration_column_in_cox"]      <-
"activity_dur"

# Remove Non-positive Durations
raw_data <- raw_data %>% filter(activity_dur > 0)
synthetic_raw <- synthetic_raw %>% filter(activity_dur > 0)
improved_synthetic <- improved_synthetic %>% filter(activity_dur > 0)
cox_estimated <- cox_estimated %>% filter(activity_dur > 0)

# Ensure 'activity_dur' is numeric in all datasets
raw_data$activity_dur <- as.numeric(raw_data$activity_dur)
synthetic_raw$activity_dur <- as.numeric(synthetic_raw$activity_dur)
improved_synthetic$activity_dur <- as.numeric(improved_synthetic$activity_dur)
cox_estimated$activity_dur <- as.numeric(cox_estimated$activity_dur)

# Add Dataset Labels
raw_data$dataset <- "Raw"
synthetic_raw$dataset <- "Synthetic_Raw"
improved_synthetic$dataset <- "Improved_Synthetic"
cox_estimated$dataset <- "Simplecox_Synthetic"

# Combine Data
combined_data <- bind_rows(raw_data, synthetic_raw, improved_synthetic, cox_estimated)

# Filter data to include only durations of up to 20 hours
combined_data <- combined_data %>%
  filter(activity_dur <= 20)

# Open text file to save results
sink("statistical_metrics.txt")

# === Step 2: Descriptive Analysis ===
cat("\nSummary Statistics:\n")
cat("Raw Data:\n")
print(summary(raw_data$activity_dur))
cat("\nSynthetic Raw Data:\n")
print(summary(synthetic raw$activity dur))
cat("\nImproved Synthetic Data:\n")
print(summary(improved_synthetic$activity_dur))
cat("\nCox-Estimated Data:\n")
print(summary(cox_estimated$activity_dur))

# === Step 3: Kolmogorov-Smirnov Tests ===
cat("\nKolmogorov-Smirnov Test:\n")
cat("Raw vs Synthetic Raw:\n")
ks_test_raw_synthetic <- ks.test(raw_data$activity_dur, synthetic_raw$activity_dur)
print(ks_test_raw_synthetic)
```

```
cat("\nRaw vs Improved Synthetic:\n")
ks_test_raw_improved <- ks.test(raw_data$activity_dur, improved_synthetic$activity_dur)
print(ks_test_raw_improved)

cat("\nRaw vs Cox-Estimated:\n")
ks_test_raw_cox <- ks.test(raw_data$activity_dur, cox_estimated$activity_dur)
print(ks_test_raw_cox)

cat("\nImproved Synthetic vs Cox-Estimated:\n")
ks_test_improved_cox            <-            ks.test(improved_synthetic$activity_dur,
cox_estimated$activity_dur)
print(ks_test_improved_cox)

# === Step 4: MAE and RMSE ===
# Define functions for MAE and RMSE
mae <- function(actual, predicted) {
  mean(abs(actual - predicted))
}

rmse <- function(actual, predicted) {
  sqrt(mean((actual - predicted)^2))
}

# Calculate and save MAE and RMSE
cat("\nError Metrics:\n")
cat("Raw vs Synthetic Raw:\n")
cat("MAE:", mae(raw_data$activity_dur, synthetic_raw$activity_dur), "\n")
cat("RMSE:", rmse(raw_data$activity_dur, synthetic_raw$activity_dur), "\n")

cat("\nRaw vs Improved Synthetic:\n")
cat("MAE:", mae(raw_data$activity_dur, improved_synthetic$activity_dur), "\n")
cat("RMSE:", rmse(raw data$activity dur, improved synthetic$activity dur), "\n")

cat("\nRaw vs Cox-Estimated:\n")
cat("MAE:", mae(raw_data$activity_dur, cox_estimated$activity_dur), "\n")
cat("RMSE:", rmse(raw_data$activity_dur, cox_estimated$activity_dur), "\n")



# === Step 5: Visualizations ===
# Density Plot
ggplot(combined_data, aes(x = activity_dur, fill = dataset)) +
  geom_density(alpha = 0.5) +
  labs(
    title = "Comparison of Activity Durations",
    x = "Activity Duration (Hours)",
    y = "Density",
    fill = "Dataset"
  ) +
  theme_minimal()

# Boxplot
ggplot(combined_data, aes(x = dataset, y = activity_dur, fill = dataset)) +
  geom_boxplot() +
  labs(
    title = "Boxplot Comparison of Work Activity Durations",
    x = "Dataset",
    y = "Activity Duration (Hours)"
  ) +
  theme_minimal()
```

Split train/test comparison simple cox

```
 # Load required libraries
library(survival)
library(pec)
library(dplyr)
library(pracma) # For numerical integration

# Set file path and load data
file_path <- "D:/Downloads/other_data.csv"
data <- read.csv(file_path)

# Apply mappings
gender_map <- c('0: feamale' = 0, '1: male' = 1)
education_map <- c('1: primary school' = 1, '2: high school' = 2, '3: bachelor' = 3,
'4: master or phd' = 4)
employment_map <- c('1: inactive' = 1, '2: unemployed' = 2, '3: student' = 3, '4: active'
= 4)
car_own_map <- c('0: no' = 0, '1: yes' = 1)
income_map <- c('0: no income' = 0, '1: 750 or less' = 1, '2: 750-1500' = 2, '3: 1500-
2500' = 3, '4: 2500 or more' = 4)
modes_map <- c('1: car' = 1, '2: taxi' = 1, '3: bus' = 2, '4: train' = 3, '5: motorcycle'
= 4, '6: bicycle' = 5, '7: walk' = 6, '8: escooter' = 7)
zones_map <- c(
  "1" = 1, "2" = 1, "3" = 1, "4" = 1, "5" = 1, "6" = 1, "7" = 1, "13" = 1, "14" = 1,
"15" = 1, "16" = 1, "20" = 1, "26" = 1, # Central Athens
  "8" = 2, "9" = 2, "11" = 2, "17" = 2, "19" = 2, "32" = 2, # West Athens
  "10" = 3, "12" = 3, "36" = 3, # East Attica
  "18" = 4, "21" = 4, "23" = 4, "25" = 4, "28" = 4, # South Athens
  "22" = 5, "24" = 5, "29" = 5, "34" = 5, "35" = 5, # North Athens
  "27" = 6, "30" = 6, "31" = 6, "33" = 6  # Piraeus
)

# Apply mappings to the data
data$gender <- gender_map[data$gender]
data$education <- education_map[data$education]
data$employment <- employment_map[data$employment]
data$car_own <- car_own_map[data$car_own]
data$income <- income_map[data$income]
data$mode <- modes_map[as.character(data$mode)]
data$dest <- zones_map[as.character(data$dest)]

# Standardize continuous variables
data$age <- scale(data$age, center = TRUE, scale = TRUE)
data$time <- scale(data$time, center = TRUE, scale = TRUE)
data$dist <- scale(data$dist, center = TRUE, scale = TRUE)

# Filter complete cases
data <- data %>%
  filter(!is.na(activity_dur) & !is.na(event)) %>%
  filter(!is.na(gender) & !is.na(education) & !is.na(employment) & !is.na(car_own) &
           !is.na(income) & !is.na(dest) & !is.na(dist) & !is.na(mode) & !is.na(age))

# Split data into train (80%) and test (20%)
set.seed(123) # For reproducibility
train_indices <- sample(seq_len(nrow(data)), size = 0.8 * nrow(data))
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]

# Ensure factor levels in train and test data match
categorical_vars <- c("gender", "education", "employment", "income", "car_own", "mode",
"dest")
for (var in categorical_vars) {
  levels(train_data[[var]]) <- levels(data[[var]])
  levels(test_data[[var]]) <- levels(data[[var]])
}
# Fit Cox model on the train set with x and y set to TRUE
cox_model <- coxph(
  Surv(activity_dur, event) ~ gender + education + employment + income +
    car_own + time + dist + mode + age,
  data = train_data,
  x = TRUE,
  y = TRUE
)
```

```r
# Predict linear predictors (log-hazard) for the test set
test_data$linear_predictor <- predict(cox_model, newdata = test_data, type = "lp")

# Compute survival probabilities for the Cox model
baseline_hazard <- basehaz(cox_model, centered = FALSE)  # Estimate baseline hazard
baseline_times <- baseline_hazard$time
test_data$survival_probabilities <- lapply(1:nrow(test_data), function(i) {
  exp(-cumsum(baseline_hazard$hazard) * exp(test_data$linear_predictor[i]))
})

# Recompute mean durations from survival probabilities
test_data$mean_duration <- sapply(1:nrow(test_data), function(i) {
  if (!is.null(test_data$survival_probabilities[[i]])) {
    trapz(baseline_times, test_data$survival_probabilities[[i]])
  } else {
    NA
  }
})

# Filter out rows with missing mean_duration
test_data <- test_data[!is.na(test_data$mean_duration), ]

# Evaluate model predictions on the test set
# Compute Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)
actual_durations <- test_data$activity_dur
predicted durations <- test data$mean duration
mae <- mean(abs(actual_durations - predicted_durations), na.rm = TRUE)
rmse <- sqrt(mean((actual_durations - predicted_durations)^2, na.rm = TRUE))

# Save MAE and RMSE results
sink("metrics_simplecox_others.txt")
cat("MAE and RMSE for Simple Cox Model:\n")
cat("Mean Absolute Error (MAE):", mae, "\n")
cat("Root Mean Square Error (RMSE):", rmse, "\n")
sink()

# Save survival probabilities to a CSV file
survival probs df     <-     do.call(rbind,     lapply(test data$survival probabilities,
function(x) as.numeric(x)))
colnames(survival_probs_df) <- paste0("time_", seq_len(ncol(survival_probs_df)))
write.csv(survival_probs_df,   "simple_cox_survival_probabilities.csv",   row.names   =
FALSE)

# Evaluate IBS for the simple Cox model
ibs_cox <- pec(
  object = list("Simple Cox" = cox_model),
  formula = Surv(activity_dur, event) ~ 1,
  data = test_data,
  times = seq(0, max(test_data$activity_dur), by = 1) # Specify time points
)

# Save IBS results
sink("ibs_simplecox_other.txt")
cat("Integrated Brier Score (IBS) for Simple Cox Model:\n")
print(ibs_cox)
sink()

# Plot IBS for visual comparison
plot(ibs_cox, xlab = "Time", ylab = "Integrated Brier Score")

# Print completion message
cat("MAE, RMSE, and IBS for the Simple Cox Model have been calculated and saved.\n")
```

Split train/test comparison cox with frailty

```
# Load required libraries
library(survival)
library(rstan)
library(loo)
library(bayesplot)
library(dplyr)
library(pracma)  # For numerical integration
library(pec)     # For IBS calculation

# Set file path and load data
file_path <- "D:/Downloads/work_data.csv"
data <- read.csv(file path)

# Apply mappings
gender_map <- c('0: feamale' = 0, '1: male' = 1)
education_map <- c('1: primary school' = 1, '2: high school' = 2, '3: bachelor' = 3,
'4: master or phd' = 4)
employment map <- c('1: inactive' = 1, '2: unemployed' = 2, '3: student' = 3, '4: active'
= 4)
car_own_map <- c('0: no' = 0, '1: yes' = 1)
income_map <- c('0: no income' = 0, '1: 750 or less' = 1, '2: 750-1500' = 2, '3: 1500-
2500' = 3, '4: 2500 or more' = 4)
modes_map <- c('1: car' = 1, '2: taxi' = 1, '3: bus' = 2, '4: train' = 3, '5: motorcycle'
= 4, '6: bicycle' = 5, '7: walk' = 6, '8: escooter' = 7)
zones_map <- c(
  "1" = 1, "2" = 1, "3" = 1, "4" = 1, "5" = 1, "6" = 1, "7" = 1, "13" = 1, "14" = 1,
"15" = 1, "16" = 1, "20" = 1, "26" = 1, # Central Athens
  "8" = 2, "9" = 2, "11" = 2, "17" = 2, "19" = 2, "32" = 2, # West Athens
  "10" = 3, "12" = 3, "36" = 3, # East Attica
  "18" = 4, "21" = 4, "23" = 4, "25" = 4, "28" = 4, # South Athens
  "22" = 5, "24" = 5, "29" = 5, "34" = 5, "35" = 5, # North Athens
  "27" = 6, "30" = 6, "31" = 6, "33" = 6  # Piraeus
)

# Apply mappings to the data
data$gender <- gender_map[data$gender]
data$education <- education_map[data$education]
data$employment <- employment_map[data$employment]
data$car_own <- car_own_map[data$car_own]
data$income <- income_map[data$income]
data$mode <- modes_map[as.character(data$mode)]
data$dest <- zones_map[as.character(data$dest)]

# Standardize continuous variables
data$age <- scale(data$age, center = TRUE, scale = TRUE)
data$time <- scale(data$time, center = TRUE, scale = TRUE)
data$dist <- scale(data$dist, center = TRUE, scale = TRUE)

# Filter complete cases
data <- data %>%
  filter(!is.na(activity_dur) & !is.na(event)) %>%
  filter(!is.na(gender) & !is.na(education) & !is.na(employment) & !is.na(car_own) &
           !is.na(income) & !is.na(dest) & !is.na(dist) & !is.na(mode) & !is.na(age))

# Split data into train (80%) and test (20%)
set.seed(123) # For reproducibility
train_indices <- sample(seq_len(nrow(data)), size = 0.8 * nrow(data))
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]

# Fit baseline Cox model
cox model <- coxph(Surv(activity dur, event) ~ gender + education + employment + income
+
                     car_own + dest + dist + mode + age, data = train_data, x = TRUE, y
= TRUE)

# Compute baseline hazard
baseline hazard <- basehaz(cox model, centered = FALSE)
baseline_times <- baseline_hazard$time

# Compile and fit the Stan model (Frailty)
stan_train_data <- list(
  N = nrow(train_data),
```

```
  K = ncol(model.matrix(~ gender + education + employment + income + car_own + mode +
                        dest + time + age + dist - 1, data = train_data)),
  time = train_data$activity_dur,
  status = train_data$event,
  X = model.matrix(~ gender + education + employment + income + car_own + mode +
                     dest + time + age + dist - 1, data = train_data)
)

stan_model_file <- "D:/Downloads/duration/latent_class_frailty_non_centered.stan"
sm <- stan_model(file = stan_model_file)
fit <- sampling(
  sm,
  data = stan_train_data,
  iter = 1000,
  chains = 2,
  seed = 123,
  control = list(adapt_delta = 0.9999, max_treedepth = 20),
  init = 0
)

# Extract frailty effects from Stan model
posterior_samples <- as.data.frame(as.matrix(fit))
frailty_indices <- grep("^frailty", colnames(posterior_samples))
frailty_means <- colMeans(posterior_samples[, frailty_indices])

# Combine baseline Cox and frailty effects
cox_lp <- predict(cox_model, newdata = test_data, type = "lp")
test_data$linear_predictor <- sapply(1:nrow(test_data), function(i) {
  cox_lp[i] + sample(frailty_means, 1)  # Add frailty effects to Cox linear predictors
})

# Compute survival probabilities for the merged model
test_data$merged_surv_prob <- lapply(1:nrow(test_data), function(i) {
  exp(-cumsum(baseline_hazard$hazard) * exp(test_data$linear_predictor[i]))
})

# Register the predictSurvProb method for the customModel class
predictSurvProb.customModel <- function(object, newdata, times, ...) {
  do.call(rbind, lapply(1:nrow(newdata), function(i) {
    approx(
      x = baseline_times,                                # Time points from baseline hazard
       y = test_data$merged_surv_prob[[i]],              # Survival probabilities for
individual
      xout = times,                                      # Requested time points
      method = "linear", rule = 2                        # Linear interpolation, extrapolate
if needed
    )$y
  }))
}

# Evaluate IBS using the custom model
ibs_merged <- pec(
  object = list("Merged Frailty Cox" = merged_model),  # Use the custom model
  formula = Surv(activity_dur, event) ~ 1,
  data = test_data,
  times = baseline_times,
  exact = FALSE
)

# Print and save IBS results
cat("Integrated Brier Score (IBS) for Merged Frailty Cox Model:\n")
print(ibs_merged)
sink("ibs_coxfrailty_work.txt")
cat("Integrated Brier Score (IBS) for Merged Frailty Cox Model:\n")
print(ibs_merged)
sink()

# Calculate mean durations from survival probabilities
test_data$merged_mean_duration <- sapply(1:nrow(test_data), function(i) {
  trapz(baseline_times, test_data$merged_surv_prob[[i]])
})

# Calculate MAE and RMSE for the merged model
mae_merged <- mean(abs(test_data$activity_dur - test_data$merged_mean_duration), na.rm
= TRUE)
rmse_merged <- sqrt(mean((test_data$activity_dur - test_data$merged_mean_duration)^2,
na.rm = TRUE))
```

```
# Save MAE and RMSE results
sink("metrics_coxfrailty_work.txt")
cat("MAE and RMSE for Merged Frailty Cox Model:\n")
cat("Mean Absolute Error (MAE):", mae_merged, "\n")
cat("Root Mean Square Error (RMSE):", rmse_merged, "\n")
sink()

# Save survival probabilities for the merged model
merged_surv_probs_df <- as.data.frame(do.call(rbind, test_data$merged_surv_prob))
write.csv(merged_surv_probs_df, "merged_survival_probabilities.csv", row.names = FALSE)

# Print completion message
cat("Integrated Brier Score, MAE, and RMSE for Merged Frailty Cox Model have been
calculated and saved.\n")
```

Επίσης παρακάτω θα παρουσιαστεί ο κώδικας σε Python (έκδοση 3.8) του AthensPop ο οποίος χρησιμοποιήθηκε για να εξαχθούν τα μοτίβα δραστηριοτήτων και η κυκλοφοριακή ανάλυση.

Population synthesis and travel activity patterns generation

```
"""
Creates a toy population for Athens by ingesting Travel Survey data into PAM
"""

# %% Import dependencies
from re import T
from athenspop import preprocessing
import os
from datetime import timedelta
import sys
from pathlib import Path
import pandas as pd
import geopandas as gp
import matplotlib.pyplot as plt

from pam import read, write
from pam.plot.stats import plot_activity_times, plot_leg_times
from pam.samplers.spatial import RandomPointSampler
from pam.samplers.population import sample as population_sampler
from pam.samplers.time import apply_jitter_to_plan

sys.path.insert(0, os.path.join(Path(__file__).parent.absolute(), '..'))

# %% User Input
path_survey = r"C:\Users\katsa\athenspop\tests\example_data"
path_outputs = r"C:\Users\katsa\athenspop\outputs"

total_population = 3.8 * 10**6  # total population of Attica
sample_perc = 0.001

# %% Ingest travel survey data
survey raw = preprocessing.read survey(
    os.path.join(path_survey, 'NEW_diaries_athens_final.csv')
)

# person attributes
person_attributes = preprocessing.get_person_attributes(survey_raw)

# trips dataset
trips = preprocessing.get_trips_table(survey_raw)

# %% create PAM population
population = read.load_travel_diary(
    trips=trips,
    persons_attributes=person_attributes
)

# %% resample to match totals target
```

```
scale_factor = total_population * sample_perc / len(population)
population = population_sampler(population, scale_factor)
print(population)

# %% apply some jitter
#       (so that not all activities start at xx:00:00)
for hid, pid, person in population.people():
    apply_jitter_to_plan(
        person.plan,
        jitter=timedelta(minutes=30),
        min_duration=timedelta(minutes=10)
    )
    # crop to 24-hours
    person.plan.crop()

# %% some plots
population.random_person().plot()  # plot a random-person diary
plot_activity_times(population)  # activity start times distribution by purpose
plt.savefig(os.path.join(path_outputs, 'activity_times.png'))
plot_leg_times(population)  # leg start times distribution by mode
plt.savefig(os.path.join(path_outputs, 'mode_times.png'))

# %% facility sampling
zones = preprocessing.get_zones(
    path=os.path.join(path_survey, 'shp_zones', 'zones_attica.shp')
)
zones.plot()

# random point-in-polygon sampling
sampler = RandomPointSampler(geoms=zones)
population.sample_locs(sampler)

# %% export
path_out = os.path.join(path_outputs, 'plans.xml')
write.write_matsim(
    population,
    plans_path=path_out,
    comment='Athens example pop'
)
population.to_csv(path_outputs, crs=2100)
print(f'Population exported to {path_out}')
```

Analysis

```
"""
Exploration of the diary data
"""

#%% import dependencies
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
from athenspop import preprocessing

# Check and set the interactive mode to off explicitly
plt.ioff()  # Ensure interactive mode is off to prevent automatic display of figures

# Define paths for data
path_survey = r'C:\Users\katsa\athenspop\tests\example_data'
path_outputs = r'C:\Users\katsa\athenspop\outputs'

# Load the survey data
survey_raw = preprocessing.read_survey(
    os.path.join(path_survey, 'NEW_diaries_athens_final(1).csv')
)
person_attributes = preprocessing.get_person_attributes(survey_raw)
print (person_attributes)
trips = preprocessing.get_trips_table(survey_raw)

# Merge person attributes with trip data
df = pd.merge(trips, person_attributes, on='pid')
```

```
# %% Function to plot number of trips distribution by different groups
def n_trips_distribution(df: pd.DataFrame, groupby: str) -> None:
    """
    Distribution of the number of trips by demographic group
    """
    fig, ax = plt.subplots()  # Use subplots to better manage figure lifecycle
    group_data = df.groupby(groupby).pid.value_counts()
                                    cumulative_data                                    =
group_data.groupby(level=0).value_counts(normalize=True).sort_index().groupby(level=0)
.cumsum()
    cumulative_data.unstack(level=0).plot(marker='o', ax=ax)
    ax.set_xlabel('Number of trips')
    ax.set_ylabel('Cumulative frequency')
    ax.set_ylim(0, 1.3)
    ax.set_xlim(1, 6.5)
    ax.set_title(f'Number of Trips Distribution by {groupby.capitalize()}')
    ax.grid(True)
    plt.show()
    plt.close(fig)  # Explicitly close the figure after displaying it

# Call the function with different groupings
n_trips_distribution(df, 'income')
n_trips_distribution(df, 'age_group')
n_trips_distribution(df, 'gender')

# %% Correlations between variables: Age vs Income
fig, ax = plt.subplots()
age_income_data                                                                        =
person_attributes.groupby('age_group').income.value_counts(normalize=True).unstack(lev
el='income')[['zero', 'low', 'medium', 'high']].fillna(0)
age income data.plot(kind='bar', stacked=True, ax=ax)
ax.set_title('Age vs Income Distribution')
ax.set_xlabel('Age Group')
ax.set_ylabel('Proportion')
plt.show()
plt.close(fig)

# %% Percentage of Trips by Hour of the Day
fig, ax = plt.subplots()

# Filter trips to include only those from the first day
df_first_day = df[df['time'] < 24]  # Ensure only times within 0-23 hours are included

# Percentage of Trips by Hour of the Day
fig, ax = plt.subplots()

# Count the trips by hour and normalize to get percentages
hourly_trip_distribution                                                               =
df_first_day['time'].value_counts(normalize=True).sort_index() * 100

# Plot the hourly distribution as a bar plot
hourly_trip_distribution.plot(kind='bar', ax=ax, color='skyblue')

# Customize the plot
ax.set_title('Percentage of Trips by Hour of the Day')
ax.set_xlabel('Hour of the Day')
ax.set_ylabel('Percentage of Trips (%)')
ax.grid(True)

# Show plot
plt.tight_layout()
plt.show()
plt.close(fig)


# %% Percentage of Trips by Hour of the Day
fig, ax = plt.subplots()

# Count the trips by hour and normalize to get percentages
hourly_trip_distribution                                                               =
df_first_day['time'].value_counts(normalize=True).sort_index() * 100

# Plot the hourly distribution as a bar plot
hourly_trip_distribution.plot(kind='bar', ax=ax, color='skyblue')

# Customize the plot
```

```
ax.set_title('Percentage of Trips by Hour of the Day')
ax.set_xlabel('Hour of the Day')
ax.set_ylabel('Percentage of Trips (%)')
ax.grid(True)

# Show plot
plt.tight_layout()
plt.show()
plt.close(fig)

# Filter trips to include only those within the first 24 hours
df_within_24_hours = df[df['time'] < 24]  # Ensure only times within 0-23 hours are
included

# %% Line plot of trips by hour for different income groups
fig, ax = plt.subplots()

for income_group in ['low', 'medium', 'high']:
        hourly_data     =     df_within_24_hours[df_within_24_hours['income']     ==
income_group]['time'] \
        .value_counts(normalize=True).sort_index()
    hourly_data.plot(ax=ax, label=income_group.capitalize())

# Customize the plot
ax.set_title('% of Trips by Hour by Income Group')
ax.set_xlabel('Hour of the Day')
ax.set_ylabel('Proportion')
ax.legend(title="Income Group")
ax.grid(True)

# Show the plot
plt.tight_layout()
plt.show()
plt.close(fig)

# %% Purpose by income group
fig, ax = plt.subplots()
purpose_data                                                                    =
df.groupby('gender').purp.value_counts(normalize=True).unstack(level='purp')
purpose_data.plot(kind='bar', stacked=True, ax=ax)
ax.set_title('Purpose of Trips by Gender')
ax.set_xlabel('Gender')
ax.set_ylabel('Proportion')
plt.show()
plt.close(fig)
```