ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΟΜΕΑΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ &
ΕΠΙΧΕΙΡΗΣΙΑΚΗΣ ΕΡΕΥΝΑΣ

Μελέτη Σκοπιμότητας και Διερεύνηση των Δυνατοτήτων του Διαδικτυακού Εργαλείου Google Trends για την Πρόγνωση Πωλήσεων Οχημάτων στην Ελλάδα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΖΟΥΛΦΟΣ ΛΟΥΚΑΣ

Επιβλέπων: Πόνης Σταύρος, Αναπληρωτής Καθηγητής ΕΜΠ

Αθήνα, Οκτώβριος 2021

--- κενή σελίδα ---

Έχω διαβάσει και κατανοήσει τους κανόνες για τη λογοκλοπή και τον τρόπο σωστής αναφοράς των πηγών που περιέχονται στον οδηγό συγγραφής Διπλωματικών Εργασιών. Δηλώνω ότι, από όσα γνωρίζω, το περιεχόμενο της παρούσας Διπλωματικής Εργασίας είναι προϊόν δικής μου εργασίας και υπάρχουν αναφορές σε όλες τις πηγές που χρησιμοποίησα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτή τη Διπλωματική εργασία είναι του συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις της Σχολής Μηχανολόγων Μηχανικών ή του Εθνικού Μετσόβιου Πολυτεχνείου.

Ζούλφος Λουκάς

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της διπλωματικής μου εργασίας κ. Σταύρο Πόνη, για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα θέμα που με ενδιέφερε πολύ αλλά και για την εμπιστοσύνη που μου έδειξε όσον αφορά το σχεδιασμό της όλης διαδικασίας εκπόνησης αυτής της εργασίας και τη συγγραφή του παρόντος τεύχους.

Επίσης θα ήθελα να ευχαριστήσω την οικογένειά μου η οποία με στήριξε καθ' όλη τη διάρκεια των σπουδών μου και ιδιαίτερα σε αυτήν τη δύσκολη περίοδο της πανδημίας κατά την οποία εκπονήθηκε και αυτή η εργασία.

Σύνοψη - Περίληψη

Η χρήση του διαδικτύου συνεχώς αυξάνεται, ολοένα και περισσότεροι άνθρωποι το χρησιμοποιούν καθημερινά σχεδόν σε κάθε δραστηριότητα της ζωής τους. Αφιερώνουν σημαντικά μεγάλο μέρος του χρόνου τους σε αυτό, καθιστώντας το σε κάτι παραπάνω από ένα απλό εργαλείο. Μια από τις κύριες χρήσεις του διαδικτύου αποτελεί η διενέργεια αναζητήσεων για πληροφορίες αγαθών και υπηρεσιών μέσω μηχανών αναζήτησης. Ο όγκος των δεδομένων των διαδικτυακών αναζητήσεων είναι τεράστιος.

Η παρούσα διπλωματική εργασία προσπαθεί να εξετάσει αν αυτά τα δεδομένα διαδικτυακών αναζητήσεων και συγκεκριμένα αυτά που προέρχονται από την πλατφόρμα Google Trends, μπορούν να αποτυπώσουν επιτυχώς το ενδιαφέρον των καταναλωτών για αγορά προϊόντων όπως είναι τα αυτοκίνητα. Στην ουσία μελετά σε ποιο βαθμό τα δεδομένα Google Trends έχουν προγνωστική ισχύ, όσον αφορά την πρόβλεψη πωλήσεων αυτοκινήτων στην Ελλάδα. Η πρόγνωση πωλήσεων αποτελεί βασική συνιστώσα ενός αποτελεσματικού σχεδιασμού παραγωγής και κατά συνέπεια της λήψης βέλτιστων αποφάσεων ιδιαίτερα στο κλάδο της αυτοκινητοβιομηχανίας.

Για το σκοπό αυτό δημιουργείται ένας αριθμός προγνωστικών μοντέλων βασισμένα σε ποικίλες τεχνικές πρόβλεψης εισάγοντας δεδομένα από το Google Trends αλλά και διάφορες άλλες οικονομικές μεταβλητές. Πιο συγκεκριμένα για κάθε μια από τις τέσσερις μάρκες αυτοκινήτων που μελετώνται στην παρούσα εργασία, διεξάγονται μηνιαίες προβλέψεις χρονικού ορίζοντα 12 μηνών για καθορισμένο χρονικό διάστημα. Η απόδοση κάθε μοντέλου αξιολογείται με τον υπολογισμό αποκλίσεων ανάμεσα στις προβλέψεις του και στις πραγματικές παρατηρήσεις για αυτό το διάστημα.

Κρίνοντας από τα αποτελέσματα της διαδικασίας διεξαγωγής προβλέψεων, στη παρούσα εργασία επετεύχθη η δημιουργία προγνωστικού μοντέλου με την εισαγωγή δεδομένων Google Trends το οποίο κατάφερε να ξεπεράσει σε απόδοση παραδοσιακές μεθόδους πρόβλεψης κατά 17%. Επίσης στις περισσότερες περιπτώσεις —πλην ορισμένων εξαιρέσεων— η εισαγωγή τέτοιων δεδομένων βελτίωσε έστω και σε μικρό βαθμό την ακρίβεια των προβλέψεων. Συνεπώς φαίνεται ότι τα δεδομένα Google Trends έχουν σημαντική προγνωστική ισχύ, σε κάποιες περιπτώσεις τουλάχιστον.

Τα αποτελέσματα της παρούσας εργασίας καταδεικνύουν τη χρησιμότητα των δεδομένων διαδικτυακών αναζητήσεων για τη δημιουργία προγνωστικών μοντέλων, τα οποία μπορούν να χρησιμοποιηθούν από τις αυτοκινητοβιομηχανίες σαν προσθήκη στις ήδη υπάρχουσες παραδοσιακές μεθόδους πρόβλεψης. Ακόμα επισημαίνεται η αξία της πληροφορίας που μπορεί να εξαχθεί από τέτοιου είδους δεδομένα και γίνονται προτάσεις για μελλοντική διερεύνηση τρόπων που ίσως καταστήσουν αποτελεσματικότερη την εκμετάλλευση της.

Abstract

The use of the internet is constantly increasing, more and more people use it every day in almost every activity of their life. They spend a considerable amount of time on it, making it more than just a tool. One of the main uses of the internet is to search for information on goods and services through search engines. The volume of internet search data is huge.

This dissertation seeks to examine whether these online search data, and in particular those derived from the Google Trends platform, can successfully capture the interest of consumers in purchasing products such as cars. Actually, it studies the extent to which Google Trends data have predictive power, in terms of forecasting car sales in Greece. Sales forecasting is a key component of effective production planning and consequently of optimal decision making especially in the automotive industry.

For this purpose, a number of forecasting models have been created based on various forecasting techniques by entering data from Google Trends and several other economic variables. More specifically, for each of the four car brands studied in the present thesis, monthly forecasts have been made with 12 month time horizon for a specified period of time. The performance of each model is evaluated by calculating the discrepancies between its forecasts and the actual observations for this period.

Judging by the results of the forecasting process, in the present dissertation a forecasting model was created using Google Trends data and managed to outperform traditional forecasting methods by 17%. Also in most cases —with a few exceptions— the introduction of such data has improved the accuracy of the predictions, albeit to a small extent. So it seems that Google Trends data have significant predictive power, at least in some cases.

The results of this thesis demonstrate the usefulness of online search data in construction of predictive models, which can be used by the automotive industry as an addition to the existing traditional forecasting methods. The value of the information that can be extracted from such data is also pointed out and suggestions have been made for future exploration of ways that may make its exploitation more effective.

Περιεχόμενα

Ευχαριστίες **4**

Σύνοψη - Περίληψη **5**

Abstract **6**

Περιεχόμενα **7**

Κατάλογος Σχημάτων **10**

Κατάλογος Πινάκων **13**

1. Εισαγωγή **14**

1.1 Αντικείμενο και Σκοπός **14**

1.2 Βασικές Βιβλιογραφικές Αναφορές **15**

1.3 Δομή Εργασίας **16**

2. Γενικά για την Πρόγνωση **18**

2.1 Γενικά **18**

2.1.1 Εισαγωγή 18

2.1.2 Πρόγνωση ζήτησης στην εφοδιαστική αλυσίδα 19

2.1.3 Κατηγορίες προβλέψεων 20

2.2 Τεχνικές προβλέψεων **22**

2.2.1 Αποσύνθεση Χρονοσειρών 23

2.2.2 Μέθοδοι Εκθετικής Εξομάλυνσης 26

2.2.3 Μέθοδοι Γραμμικής Παλινδρόμησης 29

2.2.4 Μοντέλα ARIMA 31

2.2.5 Παλινδρόμηση με Δένδρα Αποφάσεων 34

2.2.6 Τεχνητά Νευρωνικά Δίκτυα 36

2.2.7 Μηχανές Διανυσμάτων Υποστήριξης 39

2.3 Αξιολόγηση προβλέψεων – Σφάλματα **41**

3. Παρουσίαση Google Trends **44**

3.1 Γενικά **44**

3.2 Βασικά στοιχεία εφαρμογής **45**

3.2.1 Λειτουργίες εφαρμογής 45

3.2.2 Δεδομένα Google Trends ... 46
3.3 Περιβάλλον διεπαφής χρήστη ... 50
3.3.1 Αρχική σελίδα ... 51
3.3.2 Εξερεύνηση ... 53
3.3.3 Δημοφιλείς Αναζητήσεις ... 64
3.3.4 Οι αναζητήσεις τους έτους ... 67
3.4 Εφαρμογές της πλατφόρμας Google Trends ... 68
3.4.1 Χρήσεις των δεδομένων της ιστοσελίδας ... 68
3.4.2 Google Trends από ερευνητική σκοπιά ... 69
3.4.3 Μειονεκτήματα Google Trends ... 71
4. Εφοδιαστική Αλυσίδα της Αυτοκινητοβιομηχανίας στην Ελλάδα ... 73
4.1 Ιστορία της αυτοκινητοβιομηχανίας στην Ελλάδα ... 73
4.2 Αυτοκινητοβιομηχανία στην Ευρώπη ... 76
4.3 Προώθηση αυτοκινήτων στην ελληνική αγορά ... 76
5. Μεθοδολογία ... 80
5.1 Εργαλεία υλοποίησης ... 80
5.2 Συλλογή κι επεξεργασία δεδομένων ... 80
5.3 Διαχώριση δεδομένων – δεδομένα ελέγχου ... 84
5.4 Μοντέλα μέτρου σύγκρισης (benchmarks) ... 86
5.5 Επιλογή μεθόδων πρόβλεψης ... 87
5.6 Επιλογή ανεξάρτητων μεταβλητών ... 88
5.7 Τελικά προγνωστικά μοντέλα ... 88
5.8 Αξιολόγηση ... 90
6. Παρουσίαση Αποτελεσμάτων ... 91
7. Συμπεράσματα και Μελλοντικές Προεκτάσεις ... 97
7.1 Συμπεράσματα ... 97
7.2 Μελλοντικές προεκτάσεις ... 98
Κατάλογος Αναφορών ... 100
Παράρτημα Ι ... 103
Παράρτημα ΙΙ ... 108
Παράρτημα ΙΙΙ ... 124

Κατάλογος Σχημάτων

Σχήμα 2–1: *Κατηγορίες Μεθόδων Πρόβλεψης* *22*

Σχήμα 2–2: *Εβδομαδιαίος αριθμός επιβατών οικονομικής θέσης Μελβούρνη-Σύδνεϋ (https://otexts.com/fpp2/time-plots.html)* *23*

Σχήμα 2–3: *Παράδειγμα αποσύνθεσης χρονοσειράς (https://medium.com/better-programming/a-visual-guide-to-time-series-decomposition-analysis-a1472bb9c930)* *24*

Σχήμα 2–4: *Εφαρμογή κινητού μέσου όρου σε χρονοσειρά (https://otexts.com/fpp2)* *25*

Σχήμα 2–5: *Διάγραμμα διασποράς δεδομένων και ευθεία προσαρμογής (el.wikipedia.org)* *30*

Σχήμα 2–6: *Γραφική αναπαράσταση δένδρου αποφάσεων* *35*

Σχήμα 2–7: *Δομή νευρώνα* *37*

Σχήμα 2–8: *Παράδειγμα απλού τεχνητού νευρωνικού δικτύου* *37*

Σχήμα 2–9: *SVM (https://www.saedsayad.com/support_vector_machine_reg.htm)* *40*

Σχήμα 2–10: *SVM slack variables (https://www.saedsayad.com/support_vector_machine_reg.htm)* *41*

Σχήμα 3–1: *Δεδομένα αναζήτησης όρου "bmw" (trends.google.com)* *47*

Σχήμα 3–2: *Αναζήτηση όρου "bmw" για χρονικό εύρος έτος 2018 (trends.google.com)* *47*

Σχήμα 3–3: *Αναζήτηση όρου "bmw" για χρονικό εύρος 2018-2019 (trends.google.com)* *48*

Σχήμα 3–4: *Αναζήτηση όρου "bmw" ενδιαφέρον ανά χώρα (trends.google.com)* *49*

Σχήμα 3–5: *Όρος "bmw" ενδιαφέρον ανά χώρα τελευταίες 90 μέρες (trends.google.com)* *49*

Σχήμα 3–6: *Όρος "bmw" ενδιαφέρον ανά πολιτεία ΗΠΑ (trends.google.com)* *49*

Σχήμα 3–7: *Ενότητες Google Trends (trends.google.com)* *51*

Σχήμα 3–8: *Αρχική σελίδα εφαρμογής (trends.google.com)* *51*

Σχήμα 3–9: *Πρόσφατες ιστορίες αρχική σελίδα (trends.google.com)* *52*

Σχήμα 3–10: *Πρόσφατες ανερχόμενες αναζητήσεις αρχική σελίδα (trends.google.com)* *53*

Σχήμα 3–11: *Δημοφιλέστερες αναζητήσεις του έτους αρχική σελίδα (trends.google.com)* *53*

Σχήμα 3–12: *Ενότητα Εξερεύνηση (trends.google.com)* *54*

Σχήμα 3–13: *Επιλογή περιοχής αναζητήσεων (trends.google.com)* *54*

Σχήμα 3–14: *Επιλογή χρονικού εύρους (trends.google.com)* *55*

Σχήμα 3–15: *Αυτόματη επιλογή διευκρίνηση έννοιας όρου (trends.google.com)* *56*

Σχήμα 3–16: *Επιλογή κατηγορίας έννοιας όρου (trends.google.com)* *56*

Σχήμα 3–17: *Κατηγορίες και υποκατηγορίες εννοιών (trends.google.com)* *56*

Σχήμα 3–18: *Πλατφόρμα αναζήτησης (trends.google.com)* *57*

Σχήμα 3–19: *Γράφημα χρονοσειράς (trends.google.com)* *58*

Σχήμα 3–20: *Δεδομένα ενδιαφέροντος ανά χώρα (trends.google.com)* *59*

Σχήμα 3–21: *Δεδομένα ενδιαφέροντος ανά πόλη (trends.google.com)*..... *59*

Σχήμα 3–22: *Σχετικά θέματα κι ερωτήματα αναζητούμενου όρου (trends.google.com)* *60*

Σχήμα 3–23: *Αποτελέσματα σύγκρισης δύο όρων στο χρόνο (trends.google.com)*..... *61*

Σχήμα 3–24: *Αποτελέσματα σύγκρισης δύο όρων κατά περιοχή (trends.google.com)* *62*

Σχήμα 3–25: *Ενδιαφέρον ανά περιοχή και σχετικά ερωτήματα συγκρινόμενων όρων (trends.google.com)*..... *62*

Σχήμα 3–26: *Σύγκριση όρων σε διαφορετικές περιοχές (trends.google.com)* *63*

Σχήμα 3–27: *Αλλαγή φίλτρων σύγκρισης όρων (trends.google.com)* *63*

Σχήμα 3–28: *Ημερήσιες τάσεις αναζήτησης (trends.google.com)*..... *65*

Σχήμα 3–29: *Τάσεις αναζήτησης σε πραγματικό χρόνο (trends.google.com)* *66*

Σχήμα 3–30: *Ανάπτυξη ιστορίας (trends.google.com)*..... *66*

Σχήμα 3–31: *Αναζητήσεις του έτους (trends.google.com)*..... *68*

Σχήμα 3–32: *Σύγκριση όρων με παρεμφερείς σημασίες (trends.google.com)*..... *69*

Σχήμα 4–1: *Pony-Citroen (el.wikipedia.org)* *74*

Σχήμα 4–2: *MAVA-Renault Farma F (el.wikipedia.org)*..... *74*

Σχήμα 4–3: *Enfield Neorion E 8000 Bicini (el.wikipedia.org)*..... *75*

Σχήμα 4–4: *Χάρτης εργοστασίων παραγωγής αυτοκινήτων και κινητήρων στην Ευρώπη (www.acea.be)* *76*

Σχήμα 4–5: *Αυτοκινητοφόρα (en.wikipedia.org)* *77*

Σχήμα 4–6: *Φυσικά κανάλια διανομής αυτοκινήτων στην Ελλάδα*..... *78*

Σχήμα 5–1: *Εξαγωγή δεδομένων Google Trends με την πρώτη προσέγγιση (trends.google.com)*..... *82*

Σχήμα 5–2: *Εξαγωγή δεδομένων Google Trends με την δεύτερη προσέγγιση (trends.google.com)* *83*

Σχήμα 5–3: *Απλός διαχωρισμός δεδομένων σε δεδομένα εκπαίδευσης κι ελέγχου (https://otexts.com/fpp2/accuracy.html)*..... *85*

Σχήμα 5–4: *Διαχωρισμός δεδομένων με την προσέγγιση διερυνόμενου παραθύρου - expanding window (https://otexts.com/fpp2/accuracy.html)*..... *86*

Σχήμα 6–1: *Διάγραμμα τιμών προβλέψεων και πραγματικών παρατηρήσεων 09-2018 ως 07-2021 TOYOTA* *92*

Σχήμα 6–2: *Σύγκριση των δύο καλύτερων σε απόδοση μοντέλων TOYOTA* *92*

Σχήμα 6–3: *Γραφική αναπαράσταση τιμών σφαλμάτων TOYOTA*..... *93*

Σχήμα 6–4: *Διάγραμμα τιμών προβλέψεων και πραγματικών παρατηρήσεων 09-2018 ως 07-2021 AUDI* *93*

Σχήμα 6–5: *Σύγκριση των δύο καλύτερων σε απόδοση μοντέλων AUDI* *94*

Σχήμα 6–6: *Σύγκριση του καλύτερου μοντέλου με το καλύτερο benchmark AUDI 94*

Σχήμα 6–7: *Γραφική αναπαράσταση τιμών σφαλμάτων AUDI..... 94*

Σχήμα 6–8: *Διάγραμμα τιμών προβλέψεων και πραγματικών παρατηρήσεων 09-2018 ως 07-2021 OPEL 95*

Σχήμα 6–9: *Γραφική αναπαράσταση τιμών σφαλμάτων OPEL..... 95*

Σχήμα 6–10: *Διάγραμμα τιμών προβλέψεων και πραγματικών παρατηρήσεων 09-2018 ως 07-2021 FORD 96*

Σχήμα 6–11: *Γραφική αναπαράσταση τιμών σφαλμάτων FORD..... 96*

Κατάλογος Πινάκων

Πίνακας 5-1: *Συγκεντρωτική παράθεση μεταβλητών που συλλέχθηκαν..... 84*

Πίνακας 5-2: *Προγνωστικά μοντέλα 89*

Πίνακας 6-1: *Συγκεντρωτικός πίνακας σφαλμάτων..... 91*

1. Εισαγωγή

1.1 Αντικείμενο και Σκοπός

Το διαδίκτυο έχει αναδειχθεί σε ένα σημαντικότατο εργαλείο και πλέον αποτελεί αναπόσπαστο κομμάτι της καθημερινότητας ολοένα και περισσότερων ανθρώπων. Οι χρήσεις του είναι αναρίθμητες· ξεκινώντας από την αναζήτηση πληροφοριών και δεδομένων, την επικοινωνία, την ψυχαγωγία, τα κοινωνικά δίκτυα και φτάνοντας μέχρι την αγοραπωλησία αγαθών και τη μίσθωση υπηρεσιών. Από όλες αυτές τις χρήσεις οι διαδικτυακές αναζητήσεις αποτελούν ένα μεγάλο μέρος της δραστηριότητας των χρηστών στο διαδίκτυο.

Είναι γνωστό λοιπόν πως ένας τεράστιος όγκος πληροφοριών και δεδομένων δημιουργείται και μεταφέρεται κάθε στιγμή. Αυτός ο όγκος δεδομένων αποτελείται φυσικά από το περιεχόμενο των ιστοσελίδων του διαδικτύου αλλά και από κάτι όχι τόσο φανερό και ξεκάθαρο το οποίο είναι οι πληροφορίες που προέρχονται από τη δραστηριότητα των χρηστών κατά την περιήγησή τους σε αυτό.

Η παρούσα διπλωματική εργασία λοιπόν ασχολείται με τη διερεύνηση για το αν είναι δυνατόν τέτοιου είδους δεδομένα —όπως αυτά που σχετίζονται με τη δραστηριότητα των χρηστών στο διαδίκτυο— να αποτελέσουν ικανή πηγή πληροφορίας που θα μπορέσει να βελτιώσει την ακρίβεια μοντέλων πρόγνωσης πωλήσεων αγαθών. Για να διαπιστωθεί αυτό πρέπει να δημιουργηθούν προγνωστικά μοντέλα υποστηριζόμενα από τέτοια δεδομένα και να διεξαχθούν προβλέψεις των πωλήσεων των αγαθών και τα αποτελέσματα να αξιολογηθούν για την ακρίβειά τους.

Πιο συγκεκριμένα τα δεδομένα διαδικτυακής δραστηριότητας των χρηστών που θα εξετάσει αυτή η εργασία είναι δεδομένα διαδικτυακών αναζητήσεων τα οποία έχουν εξαχθεί από την πλατφόρμα Google Trends της εταιρείας Google. Θα μελετηθεί αν τα δεδομένα αυτά μπορούν να παράσχουν χρήσιμες πληροφορίες σχετικά με το ενδιαφέρον των χρηστών για αγορά προϊόντων στη συγκεκριμένη περίπτωση αυτοκινήτων. Εν συντομία αντικείμενο αυτής της διπλωματικής εργασίας είναι η δημιουργία προγνωστικών μοντέλων υποστηριζόμενα από δεδομένα διαδικτυακών αναζητήσεων Google Trends, τα οποία μπορούν να προβλέψουν τις πωλήσεις επιβατικών οχημάτων στην Ελλάδα.

Η πρόγνωση πωλήσεων συνεπαγόμενη της πρόγνωσης ζήτησης αποτελεί μια ιδιαίτερα σημαντική συνιστώσα της διαδικασίας του προγραμματισμού κι ελέγχου παραγωγής (production planning). Μπορεί να χρησιμεύσει στη λήψη αποφάσεων που αφορούν τις μεταφορές, τη διαθέσιμη χωρητικότητα, το απαιτούμενο εργατικό δυναμικό, τον απαραίτητο εξοπλισμό και τον στρατηγικό σχεδιασμό όπως την τροποποίηση του δικτύου διανομής κ.τ.λ..

Η πλατφόρμα Google Trends αποτελεί μια εφαρμογή που λάνσαρε η εταιρεία Google και παρέχει δεδομένα σχετικά με τη δημοτικότητα αναζητούμενων όρων στη μηχανή αναζήτησής της. Τα δεδομένα αυτά βρίσκονται υπό τη μορφή χρονοσειρών και δε

φανερώνουν το απόλυτο αριθμό αναζητήσεων αλλά τη σχετική δημοτικότητα του αναζητούμενου όρου συγκριτικά με τη δημοτικότητά του μια άλλη χρονική στιγμή.

Η βιομηχανική δραστηριότητα στην Ελλάδα όσον αφορά την κατασκευή επιβατικών οχημάτων ειδικά αυτοκινήτων είναι σχεδόν μηδαμινή. Τα αυτοκίνητα που προορίζονται για την ελληνική αγορά εισάγονται από το εξωτερικό και κυρίως από την Ευρώπη. Αυτό καθιστά ίσως πιο δύσκολή την αποτελεσματική διαχείριση της εφοδιαστικής αλυσίδας κάτι που μπορεί να διευκολύνει αρκετά ένα αποδοτικό σύστημα διεξαγωγής προβλέψεων. Λόγω της φύσης ενός προϊόντος όπως είναι τα αυτοκίνητα (τιμή, χρόνος κατοχής κ.τ.λ.) ίσως είναι ευκολότερο τα δεδομένα των διαδικτυακών αναζητήσεων να αποτυπώσουν καλύτερα το ενδιαφέρον των καταναλωτών.

Σκοπός της παρούσας διπλωματικής είναι να συλλέξει δεδομένα σχετικά με τις αναζητήσεις των χρηστών στην Ελλάδα που αφορούν την αγορά αυτοκινήτων και στη συνέχεια να δημιουργήσει μοντέλα πρόγνωσης χρησιμοποιώντας διάφορες τεχνικές και εισάγοντας δεδομένα από την πλατφόρμα Google Trends μαζί με κάποια άλλα δεδομένα οικονομικών δεικτών. Έπειτα αφού διεξαχθούν προβλέψεις για κάποιο χρονικό διάστημα ελέγχου, γίνεται αξιολόγηση μεταξύ των μοντέλων υπολογίζοντας κάποιες μετρικές σφάλματος που μετρούν την απόκλιση των προβλέψεων από τις πραγματικές τιμές. Τέλος εξάγονται συμπεράσματα βασιζόμενα στις προαναφερθείσες μετρικές σφάλματος σχετικά με το αν είναι δυνατό να δημιουργηθούν μοντέλα πρόβλεψης με την υποστήριξη δεδομένων Google Trends, αν αυτά το μοντέλα έχουν ικανοποιητική ακρίβεια στις προβλέψεις τους και κατά πόσο η απόδοσή τους ξεπερνά παραδοσιακές μεθόδους πρόβλεψης.

Η συλλογή των απαιτούμενων δεδομένων, η διεξαγωγή των προβλέψεων αλλά και η αξιολόγηση των μοντέλων πραγματοποιήθηκε με τη βοήθεια της γλώσσας προγραμματισμού Python και των διαθέσιμων βιβλιοθηκών της, οι οποίες είναι κατάλληλες για τέτοιου είδους έργα και διευκόλυναν κατά πολύ την όλη προγραμματιστική εργασία.

1.2 Βασικές Βιβλιογραφικές Αναφορές

Στην παρούσα ενότητα θα γίνει αναφορά σε μελέτες που επικεντρώνονται μόνο στην πρόγνωση πωλήσεων αυτοκινήτων με τη βοήθεια δεδομένων Google Trends. Σε επόμενο κεφάλαιο της παρούσας εργασίας γίνεται επίσης αναφορά σε δημοσίευση μελετών σχετικά με τη χρήση των δεδομένων Google Trends για πρόγνωση σε διάφορους τομείς από τον τουρισμό και την επιδημιολογία ως τις τιμές χρηματιστηρίου και τα αποτελέσματα εκλογών.

Από τους πρώτους που εξέτασαν, αν το εργαλείο του Google Trends έχει προγνωστική ισχύ ήταν οι Choi και Varian (Choi & Varian, 2012) οι οποίοι εισήγαγαν δεδομένα από την εφαρμογή σε παλινδρομικά μοντέλα προβλέψεων και κατάφεραν να επιτύχουν βελτίωση στην ακρίβεια της πρόγνωσης πωλήσεων σε πολλούς κλάδους συμπεριλαμβανομένων αυτών της αυτοκινητοβιομηχανίας και της αγοράς αυτοκινήτων.

Σε αυτήν τη μελέτη βρέθηκε ότι τα δεδομένα Google Trends βελτιώνουν την απόδοση προγνωστικών μοντέλων με μικρό χρονικό ορίζοντα.

Άλλες σχετικές δημοσιεύσεις που μελετούν την πρόγνωση πωλήσεων αυτοκινήτων με υποστήριξη δεδομένων Google Trends είναι αυτή των Fantazzini και Toktamysova (Fantazzini & Toktamysova, 2015) η οποία προτείνει νέα μοντέλα πολλαπλών μεταβλητών για να προβλέψει τις μηνιαίες πωλήσεις αυτοκινήτων στη Γερμανία χρησιμοποιώντας οικονομικές μεταβλητές και δεδομένα διαδικτυακών αναζητήσεων της Google. Σε αυτήν τη μελέτη διεξήχθησαν προβλέψεις για δέκα μάρκες αυτοκινήτων για διάφορους χρονικούς ορίζοντες πρόβλεψης ως και 2 χρόνια μπροστά. Τα μοντέλα που περιείχαν δεδομένα αναζήτησης Google ξεπέρασαν σε απόδοση τα ανταγωνιστικά μοντέλα για τις περισσότερες μάρκες και χρονικούς ορίζοντες.

Άλλη ενδιαφέρουσα μελέτη είναι αυτή των Winjhoven και Plant (Wijnhoven & Plant, 2017) όπου διεξήγαγαν προβλέψεις πωλήσεων για 11 μοντέλα αυτοκινήτων χρησιμοποιώντας τη μέθοδο της παλινδρόμησης με δένδρα αποφάσεων. Στο προγνωστικό μοντέλο τους εισήγαγαν δεδομένα Google Trends, δεδομένα σχετικά με τον αριθμό αναφορών σε κοινωνικά δίκτυα και το μοντέλο αυτοκινήτου. Ο χρονικός ορίζοντας των προβλέψεών τους ήταν 4 μήνες και συμπέραναν ότι τα παραπάνω δεδομένα είχαν ικανή προγνωστική ισχύ σε κάποιες περιπτώσεις.

Στο άρθρο τους οι Woo και Owen (Woo & Owen, 2019) εξέτασαν τη προγνωστική σχέση που είχαν δεδομένα Google Trends (σχετιζόμενα με ειδήσεις και κατανάλωση) με αλλαγές στην κατανάλωση στις ΗΠΑ. Τα μοντέλα τους, που περιείχαν και άλλες μακροοικονομικές μεταβλητές, βελτιώθηκαν με την εισαγωγή δεδομένων Google Trends.

Το 2019 οι Kim et al. (Kim, Woo, Shin, Lee, & Kim, 2019) διερεύνησαν αν δεδομένα μηχανών διαδικτυακής αναζήτησης μπορούν να βελτιώσουν την ακρίβεια πρόγνωσης ζήτησης αυτοκινήτων. Πρότειναν ένα μοντέλο διάχυσης προϊόντος βασισμένο στο μοντέλο διάχυσης Bass, το οποίο στην ουσία είναι μια διαφορική εξίσωση. Στο μοντέλο αυτό ενσωμάτωσαν δεδομένα διαδικτυακών αναζητήσεων και βρήκαν ότι τα συγκεκριμένα δεδομένα έχουν σημαντική προγνωστική ισχύ και η απόδοση των μοντέλων που τα χρησιμοποιούσε ήταν ανώτερη των υπολοίπων τόσο στις μακροπρόθεσμες όσο και στις βραχυπρόθεσμες προβλέψεις.

Τέλος οι Wachter, Widmer και Klein (Wachter, Widmer, & Klein, 2019) πρότειναν μια τεχνική πρόγνωσης πωλήσεων αυτοκινήτων που ενσωματώνει τόσο δεδομένα διαδικτυακών αναζητήσεων όσο και οικονομικές μεταβλητές όπως η ανεργία, χρηματιστηριακοί δείκτες, τιμή βενζίνης κ.τ.λ. Το μοντέλο τους βασίστηκε στη μέθοδο γραμμικής παλινδρόμησης και είχε μια ιδιαιτερότητα στον τρόπο συλλογής δεδομένων από το Google Trends. Ο χρονικός ορίζοντας των προβλέψεών τους ήταν οι 18 μήνες. Το μοντέλο τους πέτυχε αύξηση της ακρίβειας πρόβλεψης ως και 27% σε σχέση με το μοντέλο σύγκρισης.

1.3 Δομή Εργασίας

Η δομή της παρούσας διπλωματικής εργασίας οργανώνεται ως εξής:

Το **παρόν κεφάλαιο** αποτελεί την εισαγωγή στο αντικείμενο και τους σκοπούς της εργασίας. Επίσης κάνει μια σύντομη ανασκόπηση στη βιβλιογραφία σχετικά με τις προβλέψεις πωλήσεων αυτοκινήτων με τη βοήθεια Google Trends και παραθέτει τη δομή της εργασίας.

Το **δεύτερο κεφάλαιο** αναφέρεται στη θεωρία των προβλέψεων κι αποτελεί το θεωρητικό υπόβαθρο στο οποίο στηρίζεται η εργασία. Αρχικά γίνεται μια εισαγωγή στις προβλέψεις, αναφέρεται η χρησιμότητά τους στο πεδίο της διοίκησης εφοδιαστικής αλυσίδας και περιγράφονται οι κατηγορίες και τα είδη των μεθόδων πρόγνωσης. Ύστερα παρουσιάζονται αναλυτικά εφτά τεχνικές προβλέψεων κάθε μία από τις από τις οποίες έχει διαφορετική προσέγγιση. Τέλος γίνεται αναφορά στον τρόπο αξιολόγησης των προβλέψεων και πιο συγκεκριμένα στα μέτρα σφάλματος όπου παρουσιάζονται ορισμένα από αυτά.

Στο **τρίτο κεφάλαιο** παρουσιάζεται λεπτομερώς η πλατφόρμα του Google Trends. Αρχικά παρατίθενται κάποια στοιχεία σχετικά με την ολοένα και αυξανόμενη χρήση του διαδικτύου και των διαδικτυακών αναζητήσεων. Έπειτα γίνεται αναφορά στις βασικές λειτουργίες της εφαρμογής και στους τύπους των δεδομένων που προσφέρονται από αυτή. Επιπρόσθετα, περιγράφεται με λίγα λόγια ο τρόπος που προκύπτουν αυτά τα δεδομένα. Στη συνέχεια παρουσιάζεται αναλυτικά το περιβάλλον διεπαφής χρήστη της εφαρμογής και όλα τα μέρη από τα οποία συντίθεται, μαζί με τις επιλογές που προσφέρει σε κάθε μία από τις ενότητές της. Τέλος γίνεται λόγος σε πεδία που τα δεδομένα της πλατφόρμας μπορεί να αποδειχθούν χρήσιμα, σε απόπειρες έρευνας της χρησιμότητας της εφαρμογής αλλά και στα μειονεκτήματα που μπορεί να ενέχει η χρήση της.

Το **τέταρτο κεφάλαιο** αναφέρεται με συντομία στην Εφοδιαστική αλυσίδα της αυτοκινητοβιομηχανίας στην Ελλάδα. Αρχικά γίνεται μια ιστορική ανασκόπηση της βιομηχανικής δραστηριότητας στη Ελλάδα όσον αφορά τα οχήματα. Στην πορεία περιγράφεται η εικόνα της Ευρώπης σχετικά με τον κλάδο παραγωγής αυτοκινήτων αλλά και η διαδικασία προώθησης αυτοκινήτων στην ελληνική αγορά.

Στο **πέμπτο κεφάλαιο** περιγράφεται βήμα-βήμα (από τη συλλογή δεδομένων μέχρι την αξιολόγηση των προβλέψεων με μέτρα σφάλματος) η διαδικασία που ακολουθήθηκε για τους σκοπούς αυτής της εργασίας και τη διεξαγωγή των προβλέψεων. Παρουσιάζεται αναλυτικά η μεθοδολογία και τα εργαλεία υλοποίησης του συστήματος προβλέψεων.

Στο **έκτο κεφάλαιο** γίνεται παρουσίαση των αποτελεσμάτων των προβλέψεων που διεξάχθηκαν για όλα τα μοντέλα κάθε μάρκας, συνοδευόμενα από τον απαραίτητο σχολιασμό.

Στο **έβδομο κεφάλαιο** παρατίθενται τα συμπεράσματα που εξήχθησαν κατά την εκπόνηση της παρούσας εργασίας καθώς και προτάσεις για μελλοντική έρευνα.

2. Γενικά για την Πρόγνωση

2.1 Γενικά

2.1.1 Εισαγωγή

Συχνά υπάρχει μια χρονική καθυστέρηση ανάμεσα στην επίγνωση ενός επικείμενου γεγονότος και του συμβάντος αυτού. Αυτό το χρονικό διάστημα είναι και ο κύριος λόγος του προγραμματισμού και της πρόβλεψης ή πρόγνωσης όπως ορθότερα αποκαλείται (Makridakis, Wheelwright, & Hyndman, 1998). Ως πρόγνωση μπορεί να θεωρηθεί η εκτίμηση μιας πραγματικής τιμής ή κατάστασης αναφερόμενη σε μελλοντική χρονική περίοδο. Η διεξαγωγή επιτυχών προβλέψεων παίζει σημαντικό ρόλο στο σχεδιασμό και στην σωστή λήψη αποφάσεων είτε αυτό αφορά μια επιχείρηση είτε ένα δημόσιο οργανισμό (Armstrong, 2002).

Η πρόγνωση βρίσκει πολλές εφαρμογές σε διάφορους τομείς. Ενδεικτικά αναφέρονται: η πρόγνωση τιμών μετοχών και χρηματιστηριακών δεικτών στο χρηματοοικονομικό τομέα, η πρόγνωση ενεργειακής ζήτησης και συγκεκριμένα η ημερήσια ή ωριαία πρόγνωση ζήτησης φορτίου, η πρόγνωση παραγωγής ενέργειας από ανανεώσιμες πηγές κ.τ.λ. Ακόμα όσον αφορά το περιβάλλον μπορεί να γίνει πρόγνωση της ποσότητας των υδάτινων αποθεμάτων, των ατμοσφαιρικών ρύπων, της ηχορύπανσης αλλά και στο πεδίο της μετεωρολογίας γίνεται πρόγνωση του καιρού π.χ. εμφάνιση βροχοπτώσεων ή έκτακτων καιρικών φαινομένων. Επιπροσθέτως για το κοινωνικό περιβάλλον με τη χρήση γεωδημογραφικών δεδομένων, βρίσκουν εφαρμογή οι προβλέψεις εγκληματικότητας ή επιδημιών, ενώ στον οικονομικό τομέα είναι εφικτό να προβλεφθούν το ποσοστό ανεργίας, η τιμή του ακαθάριστου εθνικού προϊόντος και άλλοι οικονομικοί δείκτες. Για την αγορά ακινήτων μπορεί να γίνει εκτίμηση των πραγματικών ή αντικειμενικών αξιών τους, στο κλάδο του τουρισμού χρήσιμες είναι οι προβλέψεις των αφίξεων των τουριστών ενώ στο μεταφορικό τομέα γίνονται προβλέψεις του κυκλοφοριακού φόρτου ή του αριθμού ατυχημάτων (FSU NTUA, 2019). Τέλος από την σκοπιά μιας εφοδιαστικής αλυσίδας όπου επικεντρώνεται και η παρούσα εργασία πολύ σημαντική είναι η πρόγνωση ζήτησης ή και πωλήσεων των αγαθών της.

Πρέπει ακόμα να σημειωθεί ότι οι προβλέψεις δεν αποτελούν κάποιο υποκατάστατο προφητείας συνεπώς πάντα θα είναι σε κάποιο ποσοστό ανακριβείς και θα έχουν σφάλματα. Η δυνατότητα πρόγνωσης κάποιων γεγονότων εξαρτάται από διάφορους παράγοντες κάποιοι από τους οποίους είναι (Hyndman & Athanasopoulos, 2020) :

- Πόσο καλά γίνονται κατανοητοί οι παράγοντες που συμβάλλουν στο γεγονός
- Η ποσότητα και η ποιότητα των δεδομένων που είναι διαθέσιμη
- Η πιθανότητα οι ίδιες οι προβλέψεις να επηρεάζουν το γεγονός που επιχειρείται να προβλεφθεί

Για παράδειγμα σε πεδία όπως το χρηματιστήριο, η δημοσίευση προβλέψεων μπορεί να οδηγήσει τον κόσμο προς τη μαζική εφαρμογή της υπόδειξης της πρόβλεψης κι έτσι να

υπάρξει μια αυτοεκπλήρωση της. Αντίθετα αν λόγου χάρη δημοσιευθεί μια πρόβλεψη για την εξάπλωση μιας ασθένειας τότε μπορεί ο κόσμος να πάρει περισσότερα μέτρα προστασίας έτσι ώστε να περιορισθεί η εξάπλωση της και συνεπώς η πρόβλεψη να αυτοακυρωθεί.

Τα βασικά βήματα σε μια διαδικασία πρόγνωσης είναι τα εξής:

1. **Καθορισμός προβλήματος:** Πρέπει να απαντηθούν ερωτήματα όπως ποιος και πως θα χρησιμοποιήσει τις προβλέψεις, ποιος θα συλλέξει τα στοιχεία, ποιος θα συντηρήσει τις βάσεις δεδομένων κλπ.
2. **Συλλογή πληροφοριών:** Πρέπει να συλλεχθούν πριν ξεκινήσει η διαδικασία πρόγνωσης, απαιτούνται στατιστικά δεδομένα αλλά και εμπειρία.
3. **Προκαταρτική ή διερευνητική ανάλυση:** Απεικόνιση δεδομένων, διαχείριση κενών και μηδενικών τιμών, ημερολογιακές προσαρμογές, στατιστική ανάλυση.
4. **Επιλογή και προσαρμογή μοντέλου:** Επιλογή της μεθόδου πρόγνωσης και καθορισμός των παραμέτρων.
5. **Χρήση και αξιολόγηση του προγνωστικού μοντέλου:** Χρησιμοποιείται το επιλεγμένο μοντέλο και υπολογίζεται η ακρίβειά του, δηλαδή πόσο οι παραγόμενες προβλέψεις πλησιάζουν την πραγματικότητα (FSU NTUA, 2019).

2.1.2 Πρόγνωση ζήτησης στην εφοδιαστική αλυσίδα

Η πρόγνωση ζήτησης αποτελεί τη βάση όλου του σχεδιασμού της εφοδιαστικής αλυσίδας. Ο προγραμματισμός πολλών δραστηριοτήτων από την παραγωγή, τις μεταφορές, τη διαθέσιμη χωρητικότητα, το επίπεδο των αποθεμάτων αλλά και το απαιτούμενο μέγεθος του εργατικού δυναμικού ή τον απαραίτητο εξοπλισμό εξαρτάται κατά κύριο λόγο από την πρόγνωση ζήτησης. Επίσης αποφάσεις για τον στρατηγικό σχεδιασμό όπως η τροποποίηση του δικτύου διανομής, η κατασκευή επιπλέον εγκαταστάσεων κ.λπ. λαμβάνονται με γνώμονα την αποτελεσματική πρόγνωση ζήτησης. Έπειτα αναλύονται κάποια χαρακτηριστικά των προβλέψεων και πως συνδέονται με αποφάσεις στην εφοδιαστική αλυσίδα.

Πολλές φορές η πρόγνωση ζήτησης ταυτίζεται με την πρόγνωση πωλήσεων, αυτό όμως δεν ισχύει πάντα καθώς η επιχείρηση μπορεί να μη βρίσκεται σε θέση να καλύψει την απαιτούμενη ζήτηση κι έτσι ο αριθμός πωλήσεων να είναι διαφορετικός από αυτόν της ζήτησης του προϊόντος από τους πελάτες. Συνεπώς η διαφορά ανάμεσα σε αυτά τα δύο μεγέθη πρέπει να εξετάζεται σε κάθε περίπτωση κατά τη διεξαγωγή προβλέψεων.

Όπως έχει ήδη αναφερθεί οι προβλέψεις είναι πάντα ανακριβείς και περιλαμβάνουν στην τιμή τους κάποιο μέτρο σφάλματος. Πρέπει να γίνει κατανοητή η σημασία του σφάλματος και η τιμή του δεν πρέπει να αγνοηθεί αλλά να χρησιμοποιηθεί σε ενέργειες που θα ελαχιστοποιούν το βαθμό της αβεβαιότητας σε μια εφοδιαστική αλυσίδα όπως για παράδειγμα να γίνει ορισμός κατάλληλου αποθέματος ασφαλείας βάσει του σφάλματος πρόβλεψης.

Οι μακροπρόθεσμες προβλέψεις είναι συνήθως λιγότερο ακριβείς από τις βραχυπρόθεσμες, δηλαδή έχουν μεγαλύτερη τυπική απόκλιση σφάλματος και αυτό

γενικά τονίζει τη σημασία που έχει ο χρόνος απόκρισης σε κάθε τμήμα της εφοδιαστικής αλυσίδας. Για παράδειγμα αν το χρονικό διάστημα παράδοσης μιας παραγγελίας είναι σύντομο τότε ο ανεφοδιασμός μπορεί να γίνει βάσει βραχυπρόθεσμων προβλέψεων οι οποίες είναι πιο ακριβείς. Ακόμα έχει παρατηρηθεί ότι και οι συγκεντρωτικές προβλέψεις είναι πιο ακριβείς από τις επιμέρους και αυτό καθορίζει τις αποφάσεις που αφορούν την προμήθεια υλικών που μπορούν να γίνουν είτε σε επίπεδο καταστήματος παραδείγματος χάριν είτε κεντρικού σημείου διανομής και μετά να μοιραστούν στα επιμέρους καταστήματα.

Επίσης όσο πιο μακριά από τον καταναλωτή βρίσκεται μια επιχείρηση στην εφοδιαστική αλυσίδα τόσο μεγαλύτερο είναι το σφάλμα πρόβλεψης. Αυτό οφείλεται στο λεγόμενο φαινόμενο του μαστίγιου (bullwhip effect) το οποίο έχει να κάνει με τη διαστρέβλωση των πληροφοριών συνεπώς και των δεδομένων πρόβλεψης από το ένα τμήμα της εφοδιαστικής αλυσίδας στο άλλο. Άρα κάθε στάδιο έχει μια διαφορετική εκτίμηση της ζήτησης κάτι που επηρεάζει την απόδοση της εφοδιαστικής αλυσίδας. Η συνεργατική πρόβλεψη που βασίζεται στις πωλήσεις στον τελικό πελάτη βοηθά να μειωθεί το σφάλμα πρόβλεψης.

Συνοψίζοντας η πρόγνωση ζήτησης σχετίζεται με πολλούς παράγοντες τους οποίους κάθε επιχείρηση πρέπει να κατανοήσει προτού επιλέξει κατάλληλη μεθοδολογία πρόβλεψης (Chopra & Meindl, 2015), μερικοί από αυτούς είναι οι εξής:

- Παρελθούσα ζήτηση
- Χρόνος ανεφοδιασμού του προϊόντος
- Προγραμματισμένη διαφήμιση ή μάρκετινγκ
- Προγραμματισμένες εκπτώσεις
- Κατάσταση της οικονομίας
- Ενέργειες των ανταγωνιστών

2.1.3 Κατηγορίες προβλέψεων

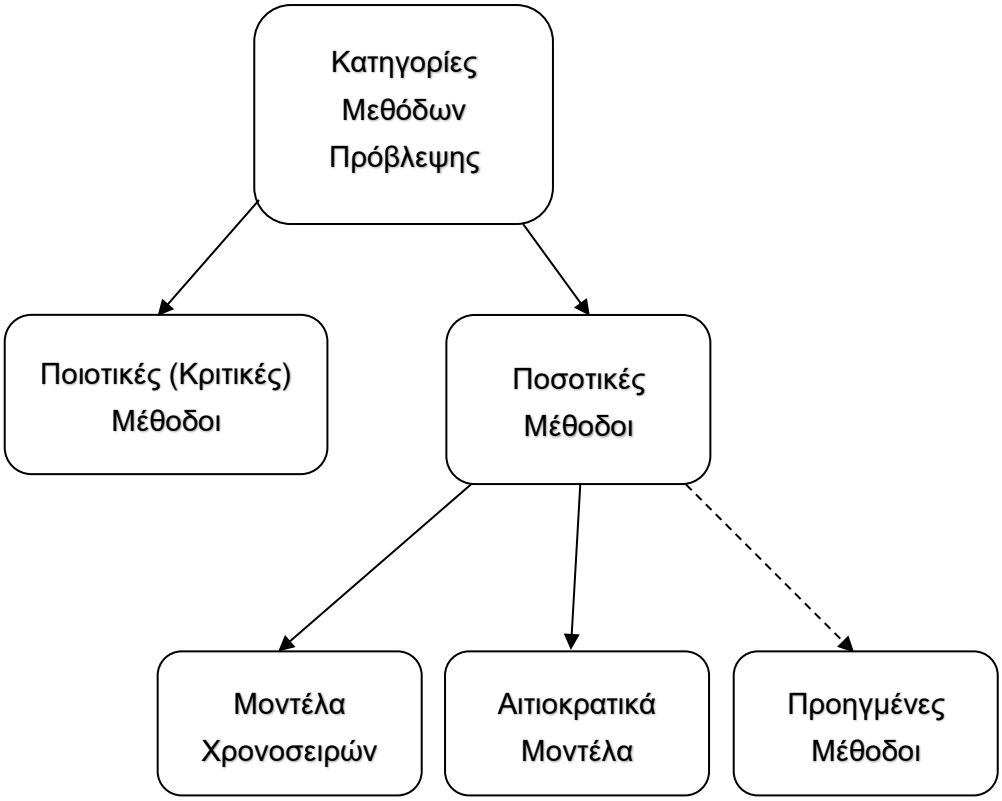
Τα προβλήματα πρόγνωσης ποικίλουν όσον αφορά το χρονικό ορίζοντα, την επάρκεια ή το μοτίβο των δεδομένων τους αλλά και διάφορες άλλες μεταβλητές τους. Υπάρχει λοιπόν πληθώρα μεθόδων προβλέψεων καθεμία κατάλληλη για κάθε περίπτωση. Μια συνήθης κατηγοριοποίηση αυτών των μεθόδων βασίζεται στη διαθεσιμότητα των δεδομένων του προβλήματος. Επομένως γίνεται ο διαχωρισμός στις ακόλουθες κατηγορίες:

- **Ποιοτικές μέθοδοι πρόβλεψης:** Οι ποιοτικές ή αλλιώς κριτικές μέθοδοι είναι κυρίως υποκειμενικές και βασίζονται στην ανθρώπινη κρίση. Ενδείκνυνται όταν είναι διαθέσιμα ελάχιστα ιστορικά δεδομένα αλλά υπάρχει επαρκής γνώση του προβλήματος από τους ειδικούς. Αυτές οι μέθοδοι δεν είναι απλές εικασίες αλλά καλά διαρθρωμένες προσεγγίσεις οι οποίες δίνουν καλές προβλέψεις σε προβλήματα χωρίς να χρησιμοποιήσουν παρελθοντικά στοιχεία. Εκτός από περιπτώσεις που δεν

υπάρχουν ιστορικά δεδομένα τέτοιου είδους προβλέψεις μπορούν να χρησιμοποιηθούν όταν λανσάρεται ένα καινούργιο προϊόν, όταν μπαίνει νέος ανταγωνιστής στην αγορά, για να προβλεφθεί μακροχρόνια ζήτηση σε ένα νέο κλάδο, στις τεχνολογικές προβλέψεις κλπ. Ακόμα οι ποιοτικές μέθοδοι χρησιμοποιούνται σε συνδυασμό με ποσοτικές επιδιώκοντας έτσι μεγαλύτερη ακρίβεια για ορισμένες περιπτώσεις. Κάποιες γνωστές μέθοδοι είναι η μέθοδος Delphi, η πρόγνωση κατά αναλογία, η έρευνα αγοράς (market research) κ.λπ. Για τη χρήση τέτοιων μεθόδων είναι απαραίτητη εμπειρία, γνώση και κριτική ικανότητα.

- **Ποσοτικές μέθοδοι πρόβλεψης:** Οι συγκεκριμένες μέθοδοι είναι και αυτές στις οποίες επικεντρώνεται η παρούσα εργασία. Οι ποσοτικές μέθοδοι μπορούν να εφαρμοστούν όταν υπάρχει διαθέσιμη πληροφορία για το παρελθόν, αυτή η πληροφορία μπορεί να ποσοτικοποιηθεί υπό την μορφή αριθμητικών δεδομένων (συνήθως χρονοσειρών) καθώς και όταν υποτεθεί ότι κάποια χαρακτηριστικά παρελθοντικών μοτίβων αυτών θα εξακολουθήσουν να υφίστανται και στο μέλλον. Υπάρχει πληθώρα διαθέσιμων ποσοτικών μεθόδων και κάθε μέθοδος είναι κατάλληλη για συγκεκριμένη περίσταση ανάλογα με το κόστος, την επιδιωκόμενη ακρίβεια και τις ιδιότητες κάθε προβλήματος. Με τη σειρά τους οι ποσοτικές μέθοδοι μπορούν να χωρισθούν στις ακόλουθες υποκατηγορίες:
    - **Μέθοδοι χρήσης μοντέλων χρονοσειρών:** Αυτές οι μέθοδοι χρησιμοποιούν ιστορικά δεδομένα για να κάνουν μια πρόβλεψη. Βασίζονται στην υπόθεση ότι η προβλεπόμενη τιμή εξαρτάται από τις προηγούμενες από αυτή τιμές. Είναι σχετικά απλούστερες μέθοδοι όσον αφορά την εφαρμογή τους και αποτελούν ένα καλό σημείο εκκίνησης για τη διεξαγωγή προβλέψεων.
    - **Μέθοδοι αιτιοκρατικών μοντέλων:** Οι αιτιοκρατικές ή αλλιώς επεξηγηματικές μέθοδοι υποθέτουν ότι η τιμή της πρόβλεψης έχει υψηλό βαθμό συσχέτισης με μια ή περισσότερες ανεξάρτητες μεταβλητές. Σύμφωνα με αυτό το μοντέλο οποιαδήποτε αλλαγή στις ανεξάρτητες μεταβλητές (εισόδου) θα επηρεάσει το αποτέλεσμα πρόβλεψης του συστήματος (έξοδος) θεωρώντας ότι η μεταξύ τους επεξηγηματική σχέση θα παραμείνει ίδια.
    - **Προηγμένες μέθοδοι:** Πρόκειται για μεθόδους που έχουν αναπτυχθεί με βάση τα παραπάνω μοντέλα, θα μπορούσαν να υπαχθούν σε αυτές τις κατηγορίες αλλά μελετώνται ξεχωριστά καθώς χρησιμοποιούν πιο εξελιγμένες και περίπλοκες τεχνικές λαμβανόμενες από τα πεδία της εξόρυξης δεδομένων, της τεχνητής νοημοσύνης και της μηχανικής μάθησης.

Πιο πάνω παρουσιάστηκε μια κατηγοριοποίηση των μεθόδων προβλέψεων χωρίς αυτό να σημαίνει ότι η συγκεκριμένη είναι και η μοναδική μιας και θα μπορούσαν να προστεθούν μέθοδοι προσομοίωσης που αφορούν για παράδειγμα την πρόβλεψη ζήτησης ή να γίνει αναφορά σε συνδυασμούς μοντέλων όπως αυτός του μοντέλου χρονοσειρών με το αιτιοκρατικό. Η κατηγοριοποίηση αυτή παριστάνεται σχηματικά στο **Σχήμα 2–1** και κάποιες από τις τεχνικές προβλέψεων αυτών των κατηγοριών περιγράφονται αναλυτικά σε επόμενο κεφάλαιο της εργασίας.



Σχήμα 2–1: Κατηγορίες Μεθόδων Πρόβλεψης

Κάθε μέθοδος επιλέγεται με βάση τον ορίζοντα πρόβλεψης δηλαδή τον δείκτη που δείχνει πόσες τιμές μελλοντικών χρονικών σημείων είναι αναγκαίο να εκτιμηθούν, έτσι παρακάτω χωρίζονται οι προβλέψεις όσον αφορά το χρονικό ορίζοντα στις ακόλουθες κατηγορίες (Hyndman & Athanasopoulos, 2020):

- **Βραχυπρόθεσμες** οι οποίες χρειάζονται για τον προγραμματισμό του προσωπικού, της παραγωγής και της μεταφοράς. Αυτού του είδους οι προβλέψεις επωφελούνται από το γεγονός ότι συνήθως δεν υπάρχουν αλλαγές σε μικρό χρονικό διάστημα και για αυτές χρησιμοποιούνται ποσοτικές στατιστικές μέθοδοι. Ο ορίζοντας πρόβλεψης συνήθως είναι μικρότερος των 3 περιόδων.
- **Μεσοπρόθεσμες** οι οποίες είναι απαραίτητες για το καθορισμό μελλοντικών απαιτήσεων σε πόρους, την πρόσληψη προσωπικού και την αγορά εξοπλισμού. Ο χρονικός ορίζοντας πρόβλεψης τις περισσότερες φορές κυμαίνεται ανάμεσα σε 12 με 15 μήνες δηλαδή περίπου ένα οικονομικό έτος.
- **Μακροπρόθεσμες** οι οποίες χρησιμοποιούνται για τον στρατηγικό σχεδιασμό, απόφαση μελλοντικών επεκτάσεων, λανσάρισμα νέων προϊόντων κλπ. Για τέτοιες προβλέψεις χρησιμοποιούνται συνήθως ποιοτικές μέθοδοι και ο ορίζοντας πρόβλεψης είναι μεγαλύτερος των τριών ετών (FSU NTUA, 2019).

2.2 Τεχνικές προβλέψεων

Σε αυτήν την ενότητα θα παρουσιασθούν αναλυτικά κάποιες τεχνικές προβλέψεων που ανήκουν στις ποσοτικές μεθόδους πρόγνωσης και εφαρμόζονται στις χρονοσειρές.

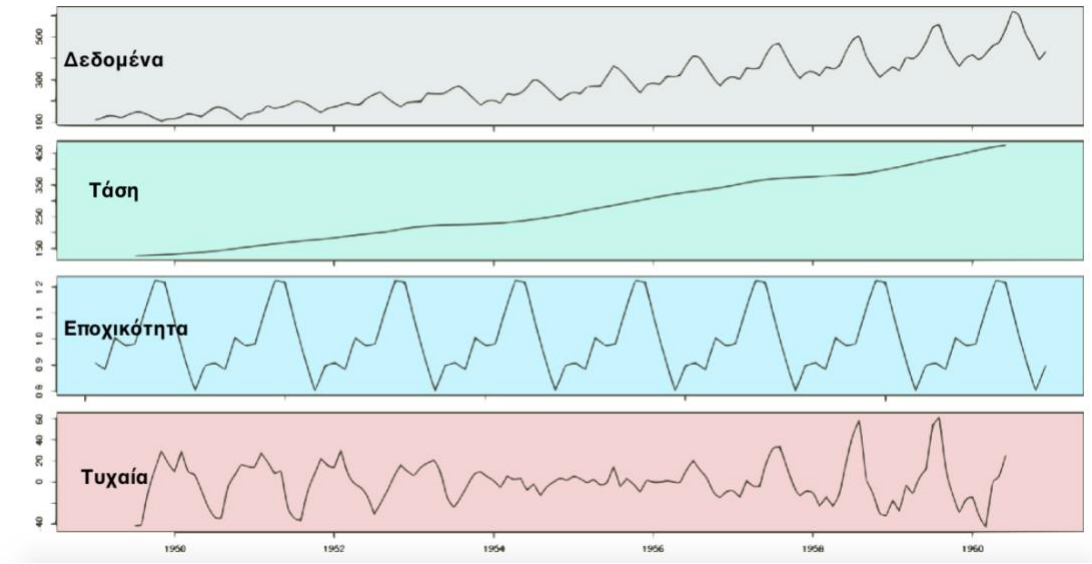
2.2.1 Αποσύνθεση Χρονοσειρών

Μια **χρονοσειρά** ή **χρονολογική σειρά** είναι μια ακολουθία παρατηρήσεων που λαμβάνονται διαδοχικά στο χρόνο (Box, Jenkins, & Reinsel, 1994). Πολλές συλλογές δεδομένων εμφανίζονται υπό την μορφή χρονοσειράς όπως οι ημερήσιες τιμές κάποιας μετοχής στο χρηματιστήριο, οι εβδομαδιαίες τιμές του αριθμού τροχαίων ατυχημάτων, η μηνιαία ακολουθία ποσοτήτων αγαθών που παράγονται ή πρέπει να παραχθούν από ένα εργοστάσιο κλπ. Γενικά πρόκειται για ακολουθία δεδομένων διακριτού χρόνου και τα χρονικά διαστήματα κατά τα οποία λαμβάνονται οι παρατηρήσεις συνήθως ισαπέχουν μεταξύ τους. Πολλές φορές κατά την ανάλυση μιας χρονοσειράς οι παρατηρήσεις της αναπαρίστανται γραφικά σε σχέση με το χρόνο σε διάγραμμα για καλύτερη μελέτης της (**Σχήμα 2–2**).

Σχήμα 2–2: Εβδομαδιαίος αριθμός επιβατών οικονομικής θέσης Μελβούρνη-Σύδνεϋ (https://otexts.com/fpp2/time-plots.html)

Μια χρονοσειρά απαρτίζεται από μια **συστηματική** συνιστώσα και μια **τυχαία**. Κατά τη μελέτη μιας χρονοσειράς μπορούν να παρατηρηθούν διάφορα μοτίβα τα οποία χωρίζονται σε κατηγορίες και αποτελούν συστατικά στοιχεία της συστηματικής συνιστώσας της χρονοσειράς. Αν βρεθούν οι τιμές των στοιχείων της συστηματικής συνιστώσας και εξαλειφθεί ο παράγοντας της τυχαίας τότε μπορεί να γίνει προβολή των μοτίβων της χρονοσειράς στο μέλλον κάτι που μπορεί να χρησιμοποιηθεί ως πρόβλεψη. Η συστηματική συνιστώσα μπορεί να αναλυθεί στα επιμέρους στοιχεία της **εποχικότητας** (seasonality), δηλαδή τις προβλέψιμες εποχικές διακυμάνσεις των παρατηρήσεων, την **τάση-κύκλο** (trend-cycle), που αντιπροσωπεύει τις μακροπρόθεσμες αλλαγές στο επίπεδο της σειράς, κάποιες φορές η τάση-κύκλος χωρίζεται στα στοιχεία της τάσης που αναπαριστά το ρυθμό αύξησης ή μείωσης των

τιμών της χρονοσειράς και του κύκλου. Παρακάτω φαίνεται γραφικά μια αποσύνθεση χρονοσειράς (**Σχήμα 2–3**).



Σχήμα 2–3: Παράδειγμα αποσύνθεσης χρονοσειράς (https://medium.com/better-programming/a-visual-guide-to-time-series-decomposition-analysis-a1472bb9c930)

Σύμφωνα με τα παραπάνω η αποσύνθεση μιας χρονοσειράς μπορεί να διατυπωθεί μαθηματικά με το ακόλουθο μοντέλο:

$$Y_t = f(S_t, T_t, C_t, R_t)$$

όπου Y_t είναι η παρατήρηση της χρονοσειράς κατά τη χρονική περίοδο t

S_t είναι η συνιστώσα εποχικότητας κατά την περίοδο t

T_t είναι η συνιστώσα τάσης κατά την περίοδο t

C_t είναι η συνιστώσα κύκλου κατά την περίοδο t

R_t είναι η τυχαία συνιστώσα κατά την περίοδο t

Το παραπάνω μοντέλο μπορεί να εκφρασθεί σε διάφορες μορφές όπως αυτή του προσθετικού (additive) τύπου:

$$Y_t = S_t + T_t + C_t + R_t$$

και του πολλαπλασιαστικού τύπου (multiplicative):

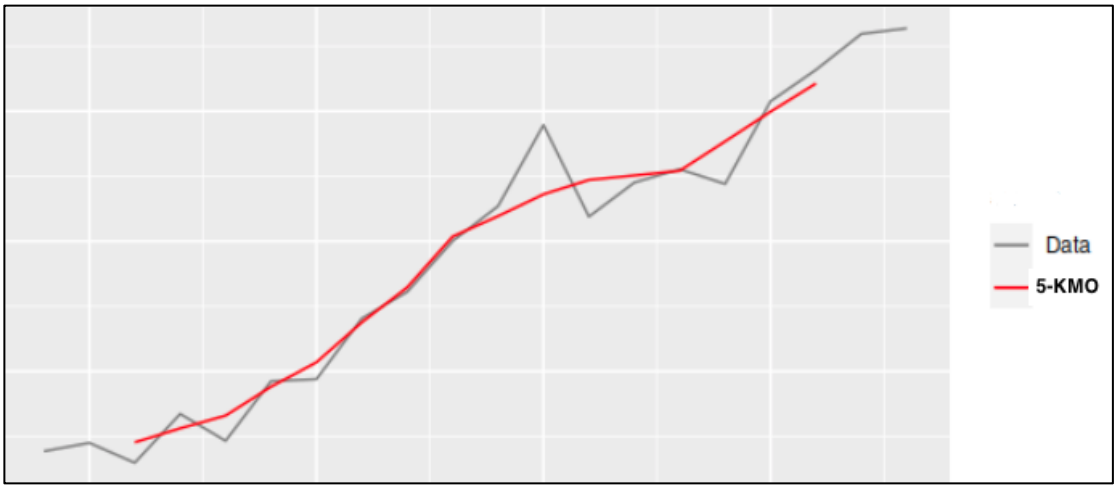
$$Y_t = S_t \times T_t \times C_t \times R_t$$

Υπάρχουν διάφορες μέθοδοι για την προσέγγιση των στοιχείων μιας χρονοσειράς και κατά συνέπεια της αποσύνθεσής της. Παραδείγματος χάριν ένας τρόπος για την εκτίμηση της συνιστώσας τάσης-κύκλου είναι η χρήση της μεθόδου του **κινητού μέσου όρου**. Αυτή η μέθοδος μπορεί να εξομαλύνει τα δεδομένα της χρονοσειράς κι έτσι να μειώσει τις τυχαίες παρεκκλίσεις. Μεγάλο πλήθος τέτοιων μεθόδων εξομάλυνσης είναι διαθέσιμο αλλά πρώτα θα γίνει εκτενής αναφορά σε αυτή του κινητού μέσου όρου μιας και είναι η απλούστερη αλλά και η παλιότερη μέθοδος ενώ αποτελεί βάση για σύγχρονες και πιο εξελιγμένες τεχνικές τόσο ανάλυσης όσο και πρόβλεψης χρονοσειρών.

Ένας **κινητός μέσος όρος** τάξης **m** μπορεί να γραφτεί ως:

$$\hat{T}_t = \frac{1}{m}\sum_{j=-k}^{k} Y_{t+j}$$

όπου $m = 2k + 1$ και $\hat{T}_t$ η εκτίμηση του στοιχείου της τάσης-κύκλου της χρονοσειράς. Δηλαδή η εκτίμηση της τάσης-κύκλου την στιγμή t υπολογίζεται από τον μέσο όρο των k τιμών της χρονοσειράς εκατέρωθέν του σημείου t. Η παραπάνω σχέση αναπαριστά έναν απλό κινητό μέσο όρο που ορίζεται για κάθε περιττό αριθμό m κι εξομαλύνει τη χρονοσειρά εξουδετερώνοντας έτσι σε κάποιο βαθμό την τυχαία συνιστώσα. Στο παρακάτω **Σχήμα 2–4** παρουσιάζεται σε γράφημα η εφαρμογή κινητού μέσου όρου τάξης 5 σε κάποια χρονοσειρά. Όσο πιο μεγάλη η τάξη m τόσο πιο μεγάλος ο βαθμός της εξομάλυνσης. Υπάρχουν πολλές παραλλαγές του κινητού μέσου όρου —οι οποίες δεν θα αναλυθούν στην παρούσα εργασία— όπως αυτήν για κινητό μέσο όρο με τάξη άρτιου αριθμού, ο διπλός κινητός μέσος όρος, ο σταθμισμένος κινητός μέσος όρος κλπ.



Σχήμα 2–4: Εφαρμογή κινητού μέσου όρου σε χρονοσειρά (https://otexts.com/fpp2)

Εκτός από μια μέθοδο εξομάλυνσης για την εκτίμηση λόγου χάριν της συνιστώσας της τάσης-κύκλου μιας χρονοσειράς, μέθοδοι μέσου όρου μπορούν να χρησιμοποιηθούν και ως μέθοδοι πρόβλεψης μιας χρονοσειράς. Πιο συγκεκριμένα μια πολύ απλή μέθοδος πρόβλεψης είναι αυτή του απλού μέσου όρου η οποία προβλέπει την επόμενη τιμή της χρονοσειράς υπολογίζοντας τη μέση τιμή όλων των παρατηρήσεων πριν την πρόβλεψη και διατυπώνεται μαθηματικά με τον παρακάτω τύπο:

$$F_{t+1} = \frac{1}{t}\sum_{i=1}^{t} Y_i$$

όπου F_{t+1} συμβολίζεται η πρόβλεψη για τη χρονική στιγμή $t + 1$ και Y_i η παρατήρηση της χρονοσειράς την χρονική στιγμή i.

Επίσης η μέθοδος κινητού μέσου όρου όπως παρουσιάστηκε προηγουμένως μπορεί να πραγματοποιήσει προβλέψεις για τη χρονοσειρά περιγραφόμενη από τον παρακάτω τύπο:

$$F_{t+1} = \frac{1}{k} \sum_{i=t-k+1}^{t} Y_i$$

όπου F_{t+1} συμβολίζεται η πρόβλεψη για τη χρονική στιγμή $t+1$ και Y_i η παρατήρηση της χρονοσειράς την χρονική στιγμή i και k συμβολίζεται η τάξη του κινητού μέσου όρου παρόμοια με παραπάνω.

Από στατιστική σκοπιά υπάρχουν πλήθος θεωρητικών αδυναμιών στην αποσυνθετική προσέγγιση ανάλυσης μιας χρονοσειράς. Οι επαγγελματίες ωστόσο σε μεγάλο μέρος αγνοούν αυτές τις αδυναμίες κι έχουν χρησιμοποιήσει αυτή τη προσέγγιση με αξιοσημείωτη επιτυχία. Για την αποσύνθεση χρονοσειρών υπάρχει πληθώρα μεθόδων, ενδεικτικά παρατίθενται ονόματα μερικών από αυτών όπως η κλασσική αποσύνθεση, η αποσύνθεση X11, η αποσύνθεση SEATS και η STL (Hyndman & Athanasopoulos, 2020).

Αν και είναι δυνατόν να διεξαχθούν προβλέψεις με τη χρήση της αποσύνθεσης χρονοσειρών πολλές απόπειρες έχουν καταλήξει σε όχι τόσο αξιόλογα αποτελέσματα. Η κύρια δυσκολία είναι να προβλεφθούν αποτελεσματικά οι τιμές των συνιστωσών της χρονοσειράς και ειδικά αυτές της τάσης-κύκλου. Ωστόσο είναι προτιμότερο να γίνει χρήση των μεθόδων αποσύνθεσης χρονοσειρών ως ένα εργαλείο για την καλύτερη κατανόηση των χαρακτηριστικών της χρονοσειράς, δηλαδή η αποσύνθεση να αποτελέσει ένα προκαταρτικό βήμα πριν την επιλογή και εφαρμογή μεθόδου πρόβλεψης. Το ίδιο μη ικανοποιητικά αποτελέσματα προκύπτουν και από τη χρήση μεθόδων μέσων όρων —ειδικά αν συγκριθούν με πιο σύγχρονες— αν και αυτές αποτελούν θεμέλιο για την ανάπτυξη πιο εξελιγμένων και σύγχρονων τεχνικών πρόβλεψης με εξομάλυνση. Για αυτό το λόγο κι έγινε μια συνοπτική περιγραφή αυτών των τεχνικών στην παρούσα εργασία.

2.2.2 Μέθοδοι Εκθετικής Εξομάλυνσης

Στην προηγούμενη ενότητα έγινε λόγος για τη χρήση μεθόδων εξομάλυνσης και συγκεκριμένα μεθόδων μέσου όρου για τη διεξαγωγή προβλέψεων. Οι μέθοδοι αυτές παράγουν προβλέψεις υπολογίζοντας το μέσο όρο των προηγούμενων παρατηρήσεων και σταθμίζοντας ισομερώς τις προηγούμενες τιμές ή εισάγοντάς τους προκαθορισμένα βάρη. Οι μέθοδοι που θα παρουσιασθούν σε αυτήν την ενότητα λειτουργούν με το ίδιο σκεπτικό απλά τα βάρη που εφαρμόζουν στα παρελθοντικά δεδομένα δεν είναι ίσα αλλά φθίνουν με εκθετικό τρόπο όσο οι παρατηρήσεις γίνονται παλιότερες (Makridakis, Wheelwright, & Hyndman, 1998). Δηλαδή με τη συγκεκριμένη μέθοδο η πρόβλεψη επηρεάζεται περισσότερο από τις πρόσφατες παρατηρήσεις της χρονοσειράς και ολοένα και λιγότερο από τις πιο μακρινές.

Η εκθετική εξομάλυνση προτάθηκε στα τέλη του 1950 και έδωσε ώθηση σε κάποιες από τις πιο επιτυχημένες μεθόδους προβλέψεων. Υπάρχουν πολλές τεχνικές προβλέψεων με εκθετική εξομάλυνση κι εδώ θα γίνει αναφορά σε μερικές από τις πιο σημαντικές αλλά και στην εφαρμογή τους για πρόγνωση χρονοσειρών διαφόρων χαρακτηριστικών.

Απλή εκθετική εξομάλυνση

Είναι η απλούστερη από τις μεθόδους εκθετικής εξομάλυνσης και είναι κατάλληλη για πρόγνωση δεδομένων χωρίς ξεκάθαρα στοιχεία τάσης και εποχικότητας. Οι προβλέψεις υπολογίζονται από τον ακόλουθο τύπο:

$$F_{t+1} = aY_t + (1-a)F_t$$

όπου F_{t+1} η πρόβλεψη στο χρόνο $t+1$, Y_t και F_t η παρατήρηση και η πρόβλεψη αντίστοιχα την στιγμή t και a μια σταθερά που παίρνει τιμές από 0 ως 1.

Ο παραπάνω τύπος αν γραφτεί σε μη αναδρομική μορφή αντικαθιστώντας το F_t σύμφωνα με τα παραπάνω και έπειτα το F_{t-1} και ούτω καθεξής προκύπτει η εξής διατύπωση:

$$F_{t+1} = aY_t + a(1-a)Y_{t-1} + a(1-a)^2Y_{t-2} + \cdots + a(1-a)^{t-1}Y_1 + (1-a)^tF_1$$

στην οποία φαίνεται πως η πρόβλεψη F_{t+1} αναπαριστά ένα σταθμισμένο μέσο όρο όλων των προηγουμένων παρατηρήσεων με βάρη που μειώνονται εκθετικά όσο πιο μακρινή είναι η παρατήρηση.

Για την εκκίνηση της διαδικασίας πρόβλεψης με απλή εκθετική εξομάλυνση όπως φαίνεται από τα παραπάνω είναι απαραίτητος ο ορισμός της τιμής της πρώτης πρόβλεψης F_1 η οποία μπορεί να αρχικοποιηθεί ως $F_1 = Y_1$ ή με άλλες μεθόδους αρχικοποίησης. Επίσης σημαντική είναι και η επιλογή της σταθεράς α η οποία μπορεί να γίνει έχοντας ορίσει μια σειρά δεδομένων από παρατηρήσεις και τρέχοντας τη διαδικασία πρόβλεψης σε αυτά, έτσι για κάθε τιμή της σταθεράς α μπορεί να μετρηθεί πόσο απέχουν οι εκάστοτε προβλέψεις από τις πραγματικές παρατηρήσεις με τη χρήση κάποιου μεγέθους σφάλματος και να επιλεγεί η σταθερά για την οποία ελαχιστοποιείται σφάλμα. Για την έννοια του σφάλματος λεπτομερής αναφορά σε επόμενη ενότητα του παρόντος κεφαλαίου. Η τιμή του σφάλματος μπορεί να ελαχιστοποιηθεί είτε δοκιμάζοντας διάφορες τιμές εμπειρικά είτε με χρήση κατάλληλου αλγορίθμου μη γραμμικής βελτιστοποίησης.

Στο συγκεκριμένο μοντέλο αφού επιλεγεί η σταθερά α ή αλλιώς συντελεστής εξομάλυνσης η τιμής της παραμένει ίδια κατά τη διεξαγωγή της διαδικασίας πρόβλεψης σε όλα τα δεδομένα. Μια παραλλαγή του μοντέλου αυτού είναι η τιμή της σταθεράς να αλλάζει με κατάλληλο τρόπο κι έτσι να προσαρμόζεται στις όποιες αλλαγές προκύπτουν στα μοτίβα των δεδομένων. Αυτή η παραλλαγή του μοντέλου φαίνεται ιδανική όταν απαιτηθεί να γίνουν προβλέψεις για εκατοντάδες ή και χιλιάδες αντικείμενα.

Εκθετική εξομάλυνση προσαρμοσμένη ως προς στην τάση (Υπόδειγμα Holt)

Για πρόγνωση με δεδομένα στα οποία παρατηρείται κάποιο μοτίβο τάσης ο Holt (1957) πρότεινε το μοντέλο γραμμικής εκθετικής εξομάλυνσης το οποίο αποτελεί μια επέκταση αυτού της απλής. Το μοντέλο αυτό χρησιμοποιεί δύο σταθερές εξομάλυνσης την α και

την β οι οποίες παίρνουν τιμές ανάμεσα στο 0 και το 1 και διατυπώνεται μαθηματικά από τις ακόλουθες εξισώσεις:

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} - b_{t-1})$$

$$b_t = \beta(L_t - L_{t-1)} + (1-\beta)b_{t-1}$$

$$F_{t+m} = L_t + b_t m$$

όπου το L_t δηλώνει μια εκτίμηση του επιπέδου της χρονοσειράς την στιγμή t και το b_t δηλώνει μια εκτίμηση της κλίσης της χρονοσειράς την στιγμή t ακόμα m είναι ο αριθμός των περιόδων μπροστά από το παρόν. F και Y είναι η πρόβλεψη και η παρατήρηση της χρονοσειράς κατά τα γνωστά.

Όπως και στην μέθοδο της απλής εκθετικής εξομάλυνσης έτσι κι εδώ είναι αναγκαία η αρχικοποίηση τιμών, συγκεκριμένα πρέπει να εκτιμηθούν η τιμή επιπέδου L_1και η τιμής κλίσης b_1. Μια λύση για αυτό το πρόβλημα είναι να ορισθεί $L_1 = Y_1$και $b_1 = Y_2 - Y_1$ ή $b_1 = (Y_4 - Y_1)/3$ αλλιώς υπάρχουν και άλλες προτάσεις πιο περίπλοκες για την αρχικοποίηση που δεν θα αναλυθούν εδώ.

Επίσης και σε αυτό το μοντέλο πολύ σημαντική είναι και η επιλογή των συντελεστών εξομάλυνσης α και β η οποία πρέπει να γίνει με κριτήριο την ελαχιστοποίηση του μεγέθους του σφάλματος δηλαδή την απόκλιση των πραγματικών τιμών της χρονοσειράς από αυτών των προβλέψεων που διεξήχθησαν για τον ίδιο χρόνο. Για αυτόν το σκοπό μπορεί να χρησιμοποιηθεί και κάποιος αλγόριθμος μη γραμμικής βελτιστοποίησης.

Εκθετική εξομάλυνση προσαρμοσμένη ως προς την τάση και την εποχικότητα (Υπόδειγμα Holt-Winters')

Ως τώρα τα μοντέλα εκθετικής εξομάλυνσης που παρουσιάστηκαν ήταν κατάλληλα για προβλέψεις σε δεδομένα τα οποία δεν εμφάνιζαν μοτίβα εποχικότητας. Αν οι χρονοσειρές παρουσιάζουν στοιχεία εποχικότητας μια πρόταση για τη διεξαγωγή προβλέψεων με τα παραπάνω μοντέλα θα ήταν να γίνει εποχική προσαρμογή των δεδομένων με μια μέθοδο αποσύνθεσης. Ωστόσο η παραπάνω μέθοδος του Holt επεκτάθηκε από τον Winters για μπορεί να κάνει προβλέψεις με δεδομένα τα οποία παρουσιάζουν μεν τάση αλλά εμφανίζουν και μοτίβο εποχικότητας, έτσι προέκυψε η μέθοδος Holt-Winters' η οποία μπορεί να διατυπωθεί είτε σε πολλαπλασιαστική μορφή είτε σε προσθετική.

Στην παρούσα εργασία παρουσιάζεται το πολλαπλασιαστικό μοντέλο το οποίο είναι και πιο κοινό και οι εξισώσεις που το εκφράζουν έχουν ως εξής:

$$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1})$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma)S_{t-s}$$

$$F_{t+m} = (L_t + b_t m) S_{t-s+m}$$

όπου το s δηλώνει το μήκος της εποχικότητας (πχ. αριθμός μηνών στο χρόνο), το Lt αναπαριστά το επίπεδο της χρονοσειράς, το b_t την κλίση ή τάση, το S_t το στοιχείο της εποχικότητας και το F_{t+m} την πρόβλεψη για m περιόδους μπροστά στο μέλλον.

Όπως είναι ήδη γνωστό από την περιγραφή των παραπάνω μοντέλων κι εδώ είναι απαραίτητο να ορισθούν οι αρχικές τιμές του επιπέδου L_t, της τάσης b_t και των δεικτών εποχικότητας S_t. Μια πρόταση αρχικοποίησης είναι η ακόλουθη και χρειάζεται δεδομένα τουλάχιστον μια ολόκληρης περιόδου για να γίνει. Πιο αναλυτικά η τιμή του επιπέδου ορίζεται ως:

$$L_s = \frac{1}{s}(Y_1 + Y_2 + \cdots Y_s)$$

για την αρχικοποίηση της τάσης b_s είναι βολικό να χρησιμοποιηθούν παρατηρήσεις δύο περιόδων ως εξής:

$$b_s = \frac{1}{s}\left[\frac{Y_{s+1} - Y_1}{s} + \frac{Y_{s+2} - Y_2}{s} + \cdots + \frac{Y_{s+s} - Y_s}{s}\right]$$

τέλος οι δείκτες εποχικότητας αρχικοποιούνται με τον παρακάτω τρόπο:

$$S_1 = \frac{Y_1}{L_s}, S_2 = \frac{Y_2}{L_s}, S_3 = \frac{Y_3}{L_s}$$

2.2.3 Μέθοδοι Γραμμικής Παλινδρόμησης

Ως τώρα οι μέθοδοι που έχουν παρουσιαστεί συγκαταλέγονται στην κατηγορία των μοντέλων χρονοσειρών δηλαδή είναι μέθοδοι που παράγουν προβλέψεις χρησιμοποιώντας ιστορικά δεδομένα και υποθέτοντας ότι η προβλεπόμενη τιμή εξαρτάται από τις προηγούμενες από αυτήν τιμές. Αντίθετα οι μέθοδοι παλινδρόμησης ανήκουν στην κατηγορία των αιτιοκρατικών μεθόδων πρόβλεψης και υποθέτουν ότι η τιμή που πρέπει να προβλεφθεί εξαρτάται από ένα πλήθος άλλων παραγόντων οι οποίοι αναπαρίστανται συνήθως ως σειρές δεδομένων με κάθε σημείο τους να αντιστοιχεί χρονικά σε ένα σημείο της χρονοσειράς που χρειάζεται να γίνει πρόβλεψη.

Με άλλα λόγια κάνοντας χρήση αυτών των μεθόδων μια πρόβλεψη μπορεί να εκφραστεί ως μια συνάρτηση ενός ορισμένου αριθμού παραγόντων που επηρεάζουν την έκβασή της. Έτσι αν θεωρηθεί ότι πρέπει να γίνει πρόβλεψη της μεταβλητής Y που από εδώ και πέρα θα ονομάζεται εξαρτημένη μεταβλητή έχοντας διαθέσιμο ένα πλήθος γνωστών μεταβλητών $X_1, X_2 \ldots X_k$(ανεξάρτητες μεταβλητές) τότε η πρόβλεψη θα εκφράζεται γενικά ως:

$$Y = f(X_1, X_2 \ldots X_k, e)$$

όπου Y η εξαρτημένη μεταβλητή, X οι ανεξάρτητες μεταβλητές και e το σφάλμα αφού τα Y και X συσχετίζονται προσεγγιστικά.

Σκοπός της πρόβλεψης είναι να προσδιοριστεί η συνάρτηση που συσχετίζει την εξαρτημένη μεταβλητή με τις ανεξάρτητες. Στην παρούσα εργασία οι σχέσεις μεταξύ των

μεταβλητών που θα περιγραφούν είναι γραμμικές. Υπάρχουν όμως και άλλα μοντέλα μη γραμμικά όπως αυτά της πολυωνυμικής παλινδρόμησης που υποθέτουν ότι η σχέση μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών είναι πολυώνυμα κάποιου βαθμού. Στη συνέχεια παρουσιάζονται δύο βασικά μοντέλα γραμμικής παλινδρόμησης.

Απλή γραμμική παλινδρόμηση

Στην απλούστερη περίπτωση παλινδρόμησης θεωρείται μια γραμμική σχέση της εξαρτημένης μεταβλητής Y και μιας μόνο ανεξάρτητης X για κάθε στιγμή, διατυπωμένη ως εξής:

$$Y_i = a + bX_i + e_i$$

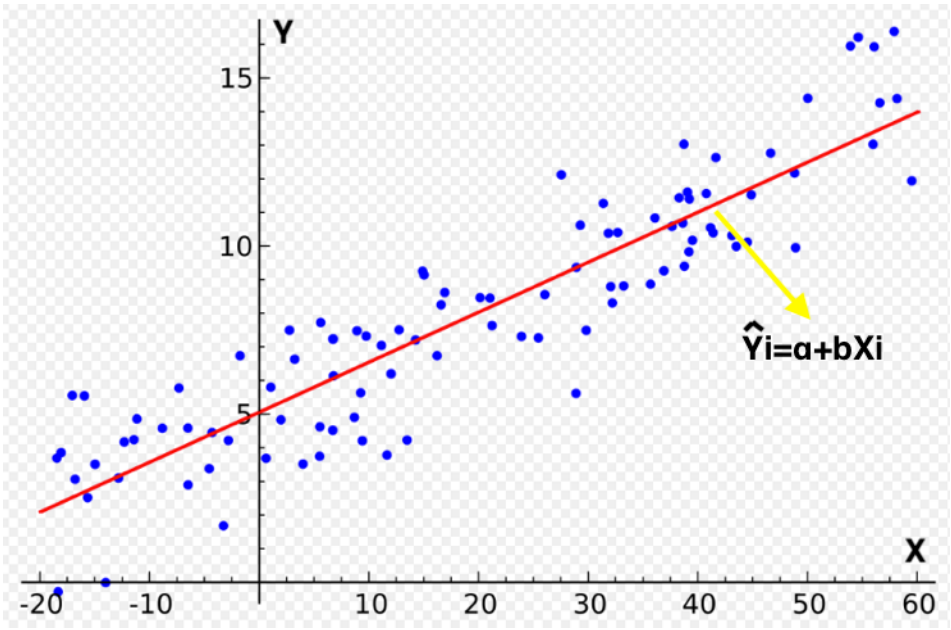
Συνεπώς αν είναι διαθέσιμες δύο σειρές n σημείων X και Y που αναπαρίστανται γραφικά στο διάγραμμα διασποράς (**Σχήμα 2–5**) τότε πρέπει να βρεθεί η ευθεία η οποία προσαρμόζεται στα δεδομένα με το βέλτιστο τρόπο. Η ευθεία έχει τη μορφή

$$\hat{Y}_i = a + bX_i$$

έτσι ώστε η παραπάνω εξίσωση να μετασχηματίζεται ως

$$Y_i = \hat{Y}_i + e_i$$

όπου



Σχήμα 2–5: Διάγραμμα διασποράς δεδομένων και ευθεία προσαρμογής (el.wikipedia.org)

Επομένως για την εύρεση της γραμμικής σχέσης των δύο μεταβλητών πρέπει να προσδιορισθούν οι σταθερές a και b. Ο βασικότερος τρόπος για να γίνει αυτό είναι η μέθοδος των ελάχιστων τετραγώνων, η οποία βρίσκει τις τιμές των a και b οι οποίες ελαχιστοποιούν το άθροισμα των τετραγώνων των σφαλμάτων e. Πιο αναλυτικά έστω ότι ορισθεί ένα μέγεθος SSE που μετρά το πόσο καλά προσαρμόζεται η ευθεία $\hat{Y}_i = a + bX_i$ στα δεδομένα, το μέγεθος αυτό ορίζεται με την ακόλουθη σχέση ως το άθροισμα των τετραγώνων των σφαλμάτων e.

$$SSE = e_1^2 + e_2^2 \dots e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

Με τη χρήση διαφορικού λογισμού για την ελαχιστοποίηση του SSE τα a και b υπολογίζονται από τις παρακάτω σχέσεις

$$b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

όπου $\bar{X}$ και $\bar{Y}$ οι μέσες τιμές των σημείων $X_1, X_2 \dots X_k$ και $Y_1, Y_2 \dots Y_k$ αντίστοιχα.

Πολλαπλή γραμμική παλινδρόμηση

Σε αυτήν την περίπτωση η εξαρτημένη μεταβλητή Y εξαρτάται από ένα πλήθος ανεξάρτητων μεταβλητών $X_1, X_2 \dots X_n$ ακολουθώντας την παρακάτω γραμμική σχέση:

$$Y_i = b_0 + b_1 X_{1,i} + \cdots b_n X_{n,i} + e_i$$

Στο συγκεκριμένο μοντέλο όπως και σε αυτό της απλής γραμμικής παλινδρόμησης πρέπει να υπολογιστούν οι τιμές των $b_0, b_1, b_2 \dots b_n$ αυτό επιτυγχάνεται πάλι με τη χρήση της μεθόδου των ελάχιστων τετραγώνων. Το σχήμα όμως της συνάρτησης που σχετίζει την εξαρτημένη μεταβλητή με τις ανεξάρτητες είναι δύσκολο να περιγραφεί σε αντίθεση με αυτό της ευθείας του προηγούμενου μοντέλου έτσι αν οι ανεξάρτητες μεταβλητές είναι δύο τότε το σχήμα της υποκείμενης συνάρτησης είναι ένα επίπεδο ενώ αν το πλήθος των ανεξάρτητων μεταβλητών είναι μεγαλύτερο τότε η συνάρτηση αναπαριστά ένα υπέρεπίπεδο δηλαδή μια ανώτερη διαστατική επιφάνεια.

Οι σταθερές $b_0, b_1, b_2 \dots b_n$ προσδιορίζονται λοιπόν ελαχιστοποιώντας το μέγεθος του αθροίσματος των τετραγώνων των σφαλμάτων S που προκύπτει ως εξής:

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1,i} - b_2 X_{2,i} - \cdots - b_n X_{n,i})^2$$

χρησιμοποιώντας διαφορικό λογισμό και λαμβάνοντας τις μερικές παραγώγους του S ως προς κάθε άγνωστο συντελεστή $b_0, b_1, b_2 \dots b_n$ υπολογίζονται έτσι οι τιμές των $b_0, b_1, b_2 \dots b_n$.

Τέλος αξίζει να σημειωθεί ότι μια εξαρτημένη μεταβλητή Y που παρατηρείται στον χρόνο t μπορεί να εξαρτάται από ανεξάρτητες μεταβλητές που παρατηρούνται σε χρόνο προγενέστερο $t - m$ ή και μεταγενέστερο $t + m$ δηλαδή να ισχύει η ακόλουθη σχέση:

$$Y_t = f(X_{1,t+l}, X_{2,t-1}, \dots X_{n,t-m}, e)$$

Σε αυτό το σημείο εισάγεται και η έννοια της καθυστέρησης (lag) δηλαδή ότι οι παρατηρήσεις των μεταβλητών σε συγκεκριμένο στιγμή επηρεάζουν μια παρατήρηση της εξαρτημένης σε διαφορετικό χρόνο.

2.2.4 Μοντέλα ARIMA

Τα ολοκληρωμένα αυτοπαλινδρομικά μοντέλα κινητών μέσων όρων ARIMA (Auto Regressive Integrated Moving Average) αποτελούν μια διαφορετική προσέγγιση στην πρόβλεψη χρονοσειρών. Ανήκουν στις ποσοτικές μεθόδους πρόβλεψης και συγκεκριμένα στα μοντέλα χρονοσειρών μιας και χρησιμοποιούν ιστορικά δεδομένα για να κάνουν μια πρόβλεψη. Η διαφορά τους με άλλες μεθόδους που ανήκουν σε αυτήν την κατηγορία όπως εκείνη της εκθετικής εξομάλυνσης είναι ότι αναλύουν τις χρονοσειρές ακολουθώντας μια στοχαστική προσέγγιση.

Πιο συγκεκριμένα τα μοντέλα ARIMA μπορούν να εκφράσουν κάθε τιμή της χρονοσειράς ως ένα γραμμικό συνδυασμό παλαιότερων παρατηρήσεων και προηγούμενων τιμών σφαλμάτων πρόβλεψης. Για την αναλυτικότερη περιγραφή των μοντέλων όμως πρέπει πρώτα να γίνει αναφορά στην έννοια της στασιμότητας χρονοσειράς η οποία παίζει σημαντικό ρόλο στη χρήση των αναφερόμενων μοντέλων.

Όσον αφορά τη **στασιμότητα**, **στάσιμη** θεωρείται μια χρονοσειρά όταν οι διακυμάνσεις των τιμών της δε διαφοροποιούνται με το χρόνο. Με άλλα λόγια για να είναι στάσιμη μια χρονοσειρά πρέπει χονδρικά τα δεδομένα της να είναι οριζόντια κατά μήκος του άξονα του χρόνου, δηλαδή να διακυμαίνονται γύρω από ένα σταθερό μέσο —ανεξάρτητο του χρόνου— και η διακύμανση της ταλάντευσης των δεδομένων να παραμένει ουσιαστικά σταθερή με το χρόνο. Στον πραγματικό κόσμο λίγες χρονοσειρές είναι στάσιμες για αυτό επινοήθηκαν διάφορες τεχνικές που κάνουν τις μη στάσιμες χρονοσειρές στάσιμες. Μία από αυτές είναι η **διαφόριση** η οποία ορίζεται ως η διαδικασία υπολογισμού της διαφοράς μια τιμής της χρονοσειράς από την προηγούμενή της. Για μια χρονοσειρά Y η διαφόρισή της εκφράζεται ως εξής:

$$Y'_t = Y_t - Y_{t-1}$$

Όπως ήδη αναφέρθηκε ένα μοντέλο ARIMA αποτελεί ένα γραμμικό συνδυασμό προηγούμενων παρατηρήσεων και τιμών σφαλμάτων της χρονοσειράς. Για χρονοσειρές που είναι στάσιμες διατυπώνεται ως ένα μοντέλο ARIMA $(p, 0, q)$ με την ακόλουθη μορφή:

$$Y_t = c + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$$

όπου η Y_t παρατήρηση την στιγμή t, $\varphi_1 \dots \varphi_p$ και $\theta_1 \dots \theta_q$ σταθεροί συντελεστές, c σταθερά, e_t το σφάλμα της πρόβλεψης για την στιγμή t .

Το μοντέλο αυτό λέγεται και ARMA (p, q) το οποίο είναι κατάλληλο για στάσιμες χρονοσειρές. Για μη στάσιμες χρονοσειρές χρησιμοποιείται το μοντέλο ARIMA(p,d,q) το οποίο εκφράζεται από την παρακάτω εξίσωση:

$$Y'_t = c + \varphi_1 Y'_{t-1} + \cdots + \varphi_p Y'_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$$

όπου Y'_t η παρατήρηση της διαφορισμένης χρονοσειράς την στιγμή t, $\varphi_1 \dots \varphi_p$ και $\theta_1 \dots \theta_q$ σταθεροί συντελεστές, c σταθερά, e_t το σφάλμα της πρόβλεψης για την στιγμή t.

Τα p και q συμβολίζουν τους αριθμούς των περιόδων πίσω για τις προηγούμενες παρατηρήσεις και τα προηγούμενα σφάλματα αντίστοιχα. Ενώ ο αριθμός d το βαθμό της

διαφόρισης δηλαδή πόσες φορές έχουν διαφοριστεί τα δεδομένα της χρονοσειράς. Όλα μαζί αποτελούν την τάξη του μοντέλου.

Στόχος της πρόγνωσης είναι η ανάλυση της χρονοσειράς με τα παραπάνω μοντέλα και η προέκτασή της βραχυπρόθεσμα στο μέλλον για την διεξαγωγή προβλέψεων. Για να επιτευχθεί αυτό μεγάλη σημασία έχει η πορεία επιλογής κατάλληλου μοντέλου που αποτελείται από τα παρακάτω βήματα:

- **Αναγνώριση του μοντέλου**, δηλαδή προσδιορισμός των τιμών των p, d, q. Αφού τα δεδομένα γίνουν στάσιμα συνήθως με διαφόριση αν είναι ανάγκη, υπάρχουν διάφοροι τρόποι ώστε να αναγνωρίσει κανείς ποια μοντέλα ενδέχεται να είναι κατάλληλα για την προέκταση της. Ένας τρόπος είναι η αναγνώριση με τη χρήση διαγραμματικών τεχνικών όπου αναπαρίστανται γραφικά οι συναρτήσεις της αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (PACF), τα μοτίβα που εμφανίζονται μπορούν να αποτελέσουν ενδείξεις για την αναγνώριση κατάλληλου μοντέλου. Αυτός ο τρόπος προϋποθέτει ότι η χρονοσειρά είναι στάσιμη κι έτσι αν χρειαστεί να διαφορισθεί κάποιες φορές η παράμετρος d παίρνει την αντίστοιχη τιμή. Ένα άλλος τρόπος αναγνώρισης είναι η χρήση στατιστικών μεθόδων, τεστ υποθέσεων που εφαρμόζουν μια σειρά από ελέγχους και κριτήρια για να συμπεράνουν αν η χρονοσειρά είναι στάσιμη ή όχι. Ανάλογα με τα αποτελέσματα γίνονται οι κατάλληλες ενέργειες για την επιλογή των παραμέτρων συνήθως με δοκιμές για τη βέλτιστη επιλογή. Το πιο διαδεδομένο τεστ ελέγχου στασιμότητας είναι το Augmented Dickey-Fuller (ADF).
- **Εκτίμηση συντελεστών.** Αφού γίνει η αναγνώριση του μοντέλου και ο προσδιορισμός των παραμέτρων p, d, q, πρέπει να εκτιμηθούν και οι συντελεστές $c, \varphi_1, \dots, \varphi_p, \theta_1, \dots \theta_q$. Για μοντέλα αυτοπαλινδρομικά που είναι συνδυασμός μόνο προηγούμενων παρατηρήσεων δηλαδή έχουν τη μορφή ARIMA$(p, d, 0)$, η εκτίμηση των συντελεστών θα μπορούσε να γίνει με τη μέθοδο των ελάχιστων τετραγώνων όπως και στην περίπτωση της γραμμικής παλινδρόμησης που αναφέρθηκε στην προηγούμενη ενότητα. Ωστόσο για τα μοντέλα που περιέχουν και όρους σφάλματος δεν είναι δυνατή η εκτίμηση των συντελεστών με απλούς τύπους όπως στην περίπτωση της παλινδρόμησης. Αντί αυτού πρέπει να χρησιμοποιηθούν επαναληπτικές μέθοδοι με τη βοήθεια υπολογιστών ελαχιστοποιώντας κάποιο μέγεθος σφάλματος. Ακόμα μια μέθοδος που συχνά χρησιμοποιείται είναι η μεγιστοποίηση της προσδοκώμενης πιθανοφάνειας L η οποία προτιμάται από τους στατιστικολόγους. Η πιθανοφάνεια εκφράζεται ως εξής:

$$L = \prod_{t=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(Y_t - F_t)^2}{2\sigma^2}}$$

όπου L η προσδοκώμενη πιθανοφάνεια, F_tη προβλεπόμενη από το μοντέλο τιμή την περίοδο t, n ο αριθμός των ιστορικών δεδομένων, e_t το σφάλμα πρόβλεψης, σ^2 η διακύμανση των σφαλμάτων του μοντέλου. Η εξίσωση ισχύει για μοντέλα ARMA δηλαδή δεν λήφθηκαν υπόψιν διαφορίσεις.

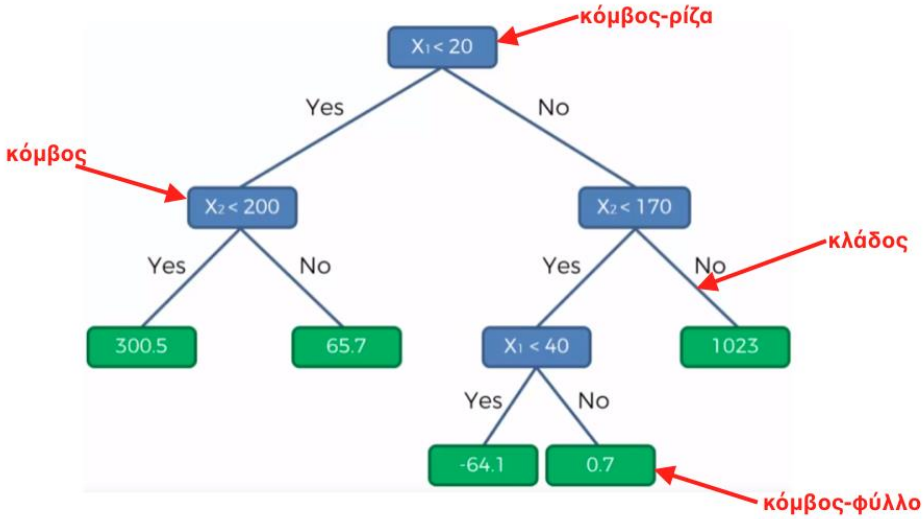
- **Διάγνωση**. Σε αυτό βήμα διεξάγονται στατιστικοί έλεγχοι προκειμένου να εξακριβωθεί αν τα μοντέλα που αναγνωρίστηκαν κι εκτιμήθηκαν είναι προβλεπτικά άρτια. Ο διαγνωστικός έλεγχος γίνεται κυρίως μελετώντας την κατανομή των σφαλμάτων πρόβλεψης e_t των υποψήφιων μοντέλων είτε με στατιστικές μεθόδους είτε ακόμα και οπτικά με διαγραμματικές. Επίσης γίνεται χρήση κάποιων κριτηρίων όπως το Akaike's Information Criterion (AIC) και το Bayesian Information Criterion (BIC).

Για τον υπολογισμό μέσω ενός μοντέλου ARIMA της τιμής Yt της χρονοσειράς, απαιτείται γνώση των τιμών $Y_{t-1}, Y_{t-2} \dots Y_{t-p}$ ή και των $e_{t-1} \dots e_{t-q}$ έτσι και για τον υπολογισμό της πρόβλεψης $\hat{Y}_{t+T}$ απαιτείται η γνώση των αντίστοιχων $Y_{t+T-1}, Y_{t+T-2}, Y_{t+T-p}$ ή και των $e_{t+T-1} \dots e_{t+T-q}$ όπου T ο ορίζοντας πρόβλεψης. Ενώ για την πρώτη πρόβλεψη Y_{t+1} οι τιμές Y_t και e_t είναι διαθέσιμες για την επόμενη πρόβλεψη Y_{t+2} οι Y_{t+1} και e_{t+1} δεν είναι διαθέσιμες κοκ. Σε αυτήν την περίπτωση θεωρείται ότι $Y_{t+1} = \hat{Y}_{t+1}$ και $e_{t+1} = 0$ συνεχίζοντας με ανάλογο τρόπο για επόμενες προβλέψεις του χρονικού ορίζοντα. Έτσι για μεγάλο χρονικό ορίζοντα οι προβλέψεις εξαρτώνται μόνο από τις προβλέψεις που έχει κάνει ήδη το μοντέλο και όχι από τίς παρατηρήσεις της χρονοσειράς αυξάνοντας έτσι το σφάλμα. Αυτός είναι και ο λόγος που αυτά τα μοντέλα χρησιμοποιούνται κυρίως για βραχυπρόθεσμες προβλέψεις.

2.2.5 Παλινδρόμηση με Δένδρα Αποφάσεων

Όπως αναφέρθηκε πρωτύτερα κατά την περιγραφή μεθόδων γραμμικής παλινδρόμησης μπορεί να γίνει πρόβλεψη μιας εξαρτημένης μεταβλητής Y με βάση ένα πλήθος ανεξάρτητων μεταβλητών$X_1 \dots X_k$ αν είναι δυνατόν να προσδιορισθεί η συνάρτηση με την οποία συσχετίζονται αυτά τα μεγέθη. Αυτές οι μέθοδοι προσέγγιζαν το πρόβλημα της παλινδρόμησης και κατ' επέκταση της πρόβλεψης από στατιστική σκοπιά, η μέθοδος των δένδρων αποφάσεων που θα παρουσιασθεί σε αυτή την ενότητα προσεγγίζει το πρόβλημα από την πλευρά της λογικής. Δηλαδή αντί μαθηματικών πράξεων και λειτουργιών γίνεται χρήση λογικών εκφράσεων εφαρμόζοντας λογικούς ή συγκριτικούς τελεστές στις τιμές των ανεξάρτητων μεταβλητών ή αλλιώς γνωρισμάτων.

Η χρήση των δένδρων αποφάσεων στην παλινδρόμηση προέρχεται από τα πεδία της τεχνητής νοημοσύνης και της εξόρυξης δεδομένων. Κυρίως χρησιμοποιούνται για προβλήματα ταξινόμησης αλλά με κατάλληλες τροποποιήσεις μπορούν να αντιμετωπίσουν και προβλήματα παλινδρόμησης. Η συγκεκριμένη μέθοδος χρησιμοποιεί το μοντέλο ενός δένδρου αποφάσεων το οποίο απαρτίζεται από ένα πλήθος κόμβων αποφάσεων οι οποίοι συνδέονται μεταξύ τους με κλάδους κι επεκτείνεται προς τα κάτω ξεκινώντας από το κόμβο-ρίζα (root node) μέχρι να τερματίσει στους κόμβους-φύλλα (leaf-nodes) (Larose & Larose, 2019) όπως φαίνεται στο **Σχήμα 2–6**.



Σχήμα 2–6: Γραφική αναπαράσταση δένδρου αποφάσεων

Κατά τη διαδικασία της παλινδρόμησης ξεκινώντας από τη ρίζα-κόμβο η οποία κατά σύμβαση τοποθετείται στην κορυφή του διαγράμματος του δένδρου αποφάσεων και σε κάθε κόμβο εξετάζεται για συγκεκριμένη μεταβλητή σε ποιο διάστημα βρίσκεται η τιμή της. Ανάλογα με το διάστημα υπάρχουν κλάδοι που οδηγούν στον επόμενο κόμβο του παρακάτω επιπέδου όπου γίνεται ξανά έλεγχος για την τιμή διαφορετικής μεταβλητής ωσότου η διαδικασία να τερματίσει φτάνοντας σε ένα κόμβο φύλλο. Για παράδειγμα αν ένα δένδρο έχει δημιουργηθεί σύμφωνα με ένα σύνολο γνωστών δεδομένων το οποίο απαρτίζεται από την εξαρτημένη μεταβλητή Y και τις ανεξάρτητες X_1 και X_2 όπως στο **Σχήμα 2–6**, για μια είσοδο με γνωστές τις τιμές $X_{1,k}$ και $X_{2,k}$ σύμφωνα με την παραπάνω διάταξη και αφού γίνουν οι έλεγχοι στους αντίστοιχους κόμβους η τιμή Y_k θα εκτιμηθεί ως μια από τις τιμές των κόμβων-φύλλων.

Έτσι όπως περιγράφηκε η παραπάνω η χρήση ενός δένδρου αποφάσεων για παλινδρόμηση είναι μια πολύ απλή και ευθεία διαδικασία. Η δυσκολία όμως αυτής της μεθόδου έγκειται στη βέλτιστη κατασκευή του δένδρου από κάποιον αλγόριθμο με βάση μια σειρά διαθέσιμων τιμών των μεταβλητών $Y, X_1 \dots X_k$ δηλαδή δεδομένων εκπαίδευσης (train set). Γενικά πρόκειται για ένα πρόβλημα δύσκολο υπολογιστικά οπότε στην πράξη χρησιμοποιούνται ευρετικές μέθοδοι για την κατασκευή του που προσπαθούν να καταλήξουν στο βέλτιστο μοντέλο.

Υπάρχουν διάφοροι αλγόριθμοι που εκτελούν τη διαδικασία της κατασκευής του δένδρου όπως ο CART (classification and regression tree), ο ID3 και οι διάδοχες εκδόσεις του C4.5 και C5.0 που αναπτύχθηκαν από τον Ross Quinlan. Σε γενική μορφή με διαθέσιμο ένα σύνολο δεδομένων εκπαίδευσης (train set) ένας τέτοιος αλγόριθμος καλείται να αποφασίσει τη διαδοχή των γνωρισμάτων για τα οποία θα ελεγχθεί κάθε σημείο, αλλά και την τιμή του ορίου με την οποία θα συγκριθεί κάθε σημείο των δεδομένων ώστε να ακολουθήσει τη βέλτιστη διαδρομή από κόμβο σε κόμβο και να φτάσει στο κόμβο-φύλλο που το χαρακτηρίζει. Υπάρχουν διάφοροι τρόποι για να ποσοτικοποιηθεί η καταλληλόλητα κάθε διαχωρισμού όπως η βελτιστοποίηση της συναρτήσης κέρδους

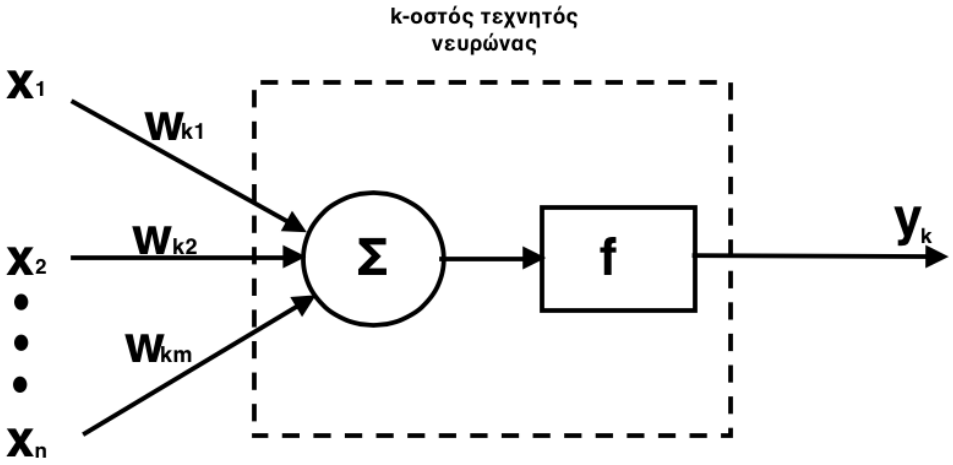
πληροφορίας (information gain) ή της συνάρτησης Gini. Η διαδικασία σταματά όταν μετά το διαχωρισμό κάθε κόμβος-φύλλο έχει σημεία ίδιας τιμής ή το δένδρο έχει φθάσει σε μια προκαθορισμένη τιμή βάθους, δηλαδή έχει αποκτήσει ένα προκαθορισμένο μέγιστο αριθμό επιπέδων. Όπως είναι φυσικό κάθε αλγόριθμος έχει τις δικές του ιδιαιτερότητες με τα πλεονεκτήματα και τα μειονεκτήματά τους.

Εν κατακλείδι τα δένδρα αποφάσεων αποτελούν μια σχετικά απλή κι ευανάγνωστη μέθοδο για παλινδρόμηση και η παραγωγή τους με τη χρήση των σχετικών αλγορίθμων είναι μια γρήγορη διαδικασία όμως ανάλογα με τη φύση και τα χαρακτηριστικά του προβλήματος μπορούν να γίνουν υπερβολικά πολύπλοκα και συχνά τείνουν προς την υπερπροσαρμογή (overfitting) στα δεδομένα εκπαίδευσης. Επίσης αποτελούν δομικό στοιχείο για άλλες τεχνικές όπως τα τυχαία δάση (Random Forest). Τα τυχαία δάση αποτελούν μια συστοιχία δένδρων αποφάσεων, το καθένα εκπαιδευμένο σε διαφορετικό τμήμα του συνόλου δεδομένων εκπαίδευσης (train set). Στην περίπτωση της παλινδρόμησης η πρόβλεψη που παράγουν αποτελεί το μέσο όρο των προβλέψεων των επιμέρους δένδρων, αντιμετωπίζοντας έτσι το φαινόμενο της υπερπροσαρμογής (overfitting).

2.2.6 Τεχνητά Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks) αποτελούν μια κατηγορία αφηρημένων μαθηματικών μοντέλων εμπνευσμένα από αντίστοιχα βιολογικά μοντέλα δηλαδή προσπαθούν να μιμηθούν τη λειτουργία των νευρώνων του ανθρώπινου εγκεφάλου. Προέρχονται από το πεδία της τεχνητής νοημοσύνης και της μηχανικής μάθησης κι ενώ η ανάπτυξή τους σαν θεωρία έχει ξεκινήσει πριν από μισό αιώνα περίπου, σήμερα η χρήση τους είναι ιδιαίτερα δημοφιλής και σε αυτό έχει συντελέσει και η μεγάλη πρόοδος στην κατασκευή των ηλεκτρονικών υπολογιστών και η αύξηση της διαθέσιμης υπολογιστικής ισχύος. Τα τεχνητά νευρωνικά δίκτυα χρησιμοποιούνται σε διάφορα προβλήματα όπως στην αναγνώριση εικόνας, στην επεξεργασία φωνής, στο χρονοπρογραμματισμό (Scheduling), στην εξόρυξη πληροφορίας και βρίσκουν εφαρμογή σε πολλούς τομείς όπως τη Ρομποτική, την Άμυνα, την Ιατρική Διαγνωστική, τον τραπεζικό τομέα κτλ. Επίσης αποτελούν ένα πολύ ισχυρό εργαλείο για την πρόβλεψη χρονοσειρών για αυτό το λόγο παρουσιάζονται και στην παρούσα εργασία.

Όπως αναφέρθηκε ένα τεχνητό νευρωνικό δίκτυο αποτελείται από έναν αριθμό τεχνητών νευρώνων οι οποίοι προσπαθούν να μιμηθούν τη λειτουργία των βιολογικών. Ένας τέτοιος απλός νευρώνας μπορεί να αναπαρασταθεί γραφικά από το παρακάτω **Σχήμα 2–7** στο οποίο διακρίνονται τρία βασικά στοιχεία: ένα σύνολο συνδέσμων οι οποίοι χαρακτηρίζονται από ένα βάρος w_{ki} , ο δείκτης k αναφέρεται στο νευρώνα προς εξέταση και ο i στην είσοδο του συνδέσμου με αντίστοιχο βάρος, ένας αθροιστής Σ στον οποίο αθροίζονται όλα τα σήματα εισόδου πολλαπλασιαζόμενα το καθένα με το αντίστοιχο βάρος και μια συνάρτηση ενεργοποίησης f η οποία περιορίζει το εύρος της εξόδου y του νευρώνα. Η συνάρτηση ενεργοποίησης f μπορεί να επιλεχθεί από ένα μεγάλο εύρος συναρτήσεων όπως γραμμικές, βηματικές, σιγμοειδείς και άλλες συναρτήσεις.

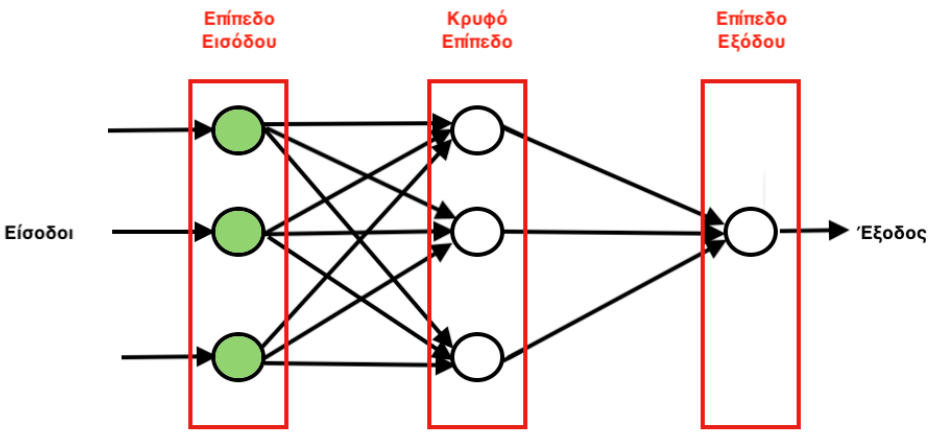


Σχήμα 2–7: Δομή νευρώνα

Η λειτουργία του τεχνητού νευρώνα διατυπώνεται από τον ακόλουθο τύπο:

$$y_k = f\left(\sum_{i=1}^{n} w_{ki}x_i\right)$$

Πολλοί νευρώνες μαζί συνθέτουν ένα τεχνητό νευρωνικό δίκτυο. Ένα απλό παράδειγμα νευρωνικού δικτύου παρουσιάζεται στο παρακάτω **Σχήμα 2–8**.



Σχήμα 2–8: Παράδειγμα απλού τεχνητού νευρωνικού δικτύου

Οι νευρώνες από τους οποίους αποτελείται ένα νευρωνικό δίκτυο συνήθως οργανώνονται σε επίπεδα ή αλλιώς στρώματα όπως φαίνεται στο παραπάνω **Σχήμα 2–8** και αναπαρίστανται με κόμβους. Υπάρχει το επίπεδο εισόδου (input layer) στο οποίο βρίσκονται οι νευρώνες οι οποίοι δέχονται τις εισόδους (inputs) του συστήματος, σε αυτούς τους νευρώνες δεν ενσωματώνεται κάποια συνάρτηση ενεργοποίησης. Ακόμα υπάρχει το επίπεδο εξόδου (output layer) το οποίο παράγει το τελικό αποτέλεσμα του

μοντέλου. Ανάμεσα στο επίπεδο εισόδου και στο επίπεδο εξόδου υπάρχουν ένα ή περισσότερα κρυφά επίπεδα (hidden layers) τα οποία εκτελούν τους υπολογισμούς και συνδέουν το στρώμα εισόδου με το στρώμα εξόδου. Κάθε κόμβος ενός επιπέδου συνδέεται συνήθως με όλους τους κόμβους του προηγούμενου επιπέδου, σε κάθε σύνδεση αντιστοιχεί ένα παραμετροποιήσιμο βάρος w κι έτσι αυτό που δέχεται κάθε κόμβος είναι ένα σταθμισμένο άθροισμα των τιμών των προηγούμενων κόμβων ανάλογα με εκάστοτε βάρη. Αυτή η περιγραφή αποτελεί μια απλή και συνηθισμένη μορφή νευρωνικού δικτύου όπου υφίσταται μόνο σύνδεση των νευρώνων διαδοχικών επιπέδων και η ροή των δεδομένων γίνεται από την είσοδο με κατεύθυνση προς την έξοδο, τα συγκεκριμένα δίκτυα λέγονται πρόσθιας τροφοδότησης (feed forward). Αντίθετα δίκτυα που μπορεί η ροή δεδομένων να γίνεται και με κυκλική διαδρομή από την έξοδο στην είσοδο, με τη χρήση βρόχων κτλ. λέγονται οπίσθιας τροφοδότησης ή ανατροφοδοτούμενα (recurrent).

Γενικά υπάρχει πληθώρα πιθανών διατάξεων ενός νευρωνικού δικτύου του οποίου η αρχιτεκτονική εξαρτάται από τον καθορισμό βασικών στοιχείων του όπως (Γεωργούλη, 2015):

- Ο αριθμός των ενδιάμεσων κρυφών επιπέδων
- Ο αριθμός των κόμβων ανά επίπεδο
- Ο τρόπος σύνδεσης των μονάδων μεταξύ τους
- Η μορφή της συνάρτησης ενεργοποίησης
- Οι τιμές των αρχικών βαρών μεταξύ των κόμβων
- Οι αλγόριθμοι (κανόνες εκπαίδευσης) που χρησιμοποιούνται, για να ενισχυθούν οι σύνδεσμοι μεταξύ των κόμβων κατά τη διαδικασία της εκπαίδευσης.

Για την εργασία της πρόβλεψης ένα νευρωνικό δίκτυο προσπαθεί να αντιληφθεί τις συσχετίσεις και τα μοτίβα που υπάρχουν στα διαθέσιμα δεδομένα κατά τη διάρκεια της εκπαίδευσης και να τροποποιήσει τις παραμέτρους του σύμφωνα με αυτά έτσι ώστε το μοντέλο να προσαρμοσθεί σε αυτές τις σχέσεις και όταν γίνει εισαγωγή νέων δεδομένων να μπορεί να διεξαγάγει προβλέψεις βασισμένες στην εκπαίδευσή του. Η τροποποίηση των παραμέτρων του νευρωνικού δικτύου αφορά την αλλαγή των τιμών των βαρών του ώστε το μοντέλο με δεδομένη μια είσοδο να παρέχει μία επιθυμητή έξοδο.

Η αλλαγή των συνδεσμικών βαρών του νευρωνικού δικτύου γίνεται κατά τη διάρκεια της εκπαίδευσης του μοντέλου με τη βοήθεια ενός αλγορίθμου εκπαίδευσης (training algorithm) που υλοποιείται αποκλειστικά από το δίκτυο χωρίς κάποια εξωτερική παρέμβαση. Κάποιο από τους αλγόριθμους εκπαίδευσης είναι αυτός της ανταγωνιστικής μάθησης (competitive learning), της τυχαίας μάθησης (random learning) και ο αλγόριθμος οπισθοδιάδοσης του λάθους (backpropagation) ο οποίος χρησιμοποιείται σε νευρωνικά δίκτυα με πολλαπλά κρυμμένα επίπεδα.

Ένας αλγόριθμος εκπαίδευσης χρησιμοποιεί κανόνες εκμάθησης με τον πιο συνηθισμένο να είναι αυτός της εκμάθησης σφάλματος-διόρθωσης. Πιο αναλυτικά για

ένα δίκτυο με πολλαπλά επίπεδα, έστω ότι είναι διαθέσιμη μια σειρά δεδομένων η οποία περιέχει ένα πλήθος εισόδων x και αντίστοιχες γνωστές εξόδους d για κάθε είσοδο. Αν για κάθε είσοδο που εισαχθεί στο νευρωνικό δίκτυο μετά από τους κατάλληλους υπολογισμούς σε κάθε νευρώνα προκύψει μια απόκριση y, τότε το σφάλμα εξόδου του νευρώνα j για τη n επανάληψη ορίζεται ως:

$$e_j(n) = d_j(n) - y_j(n)$$

Ορίζεται συνολική ενέργεια σφάλματος αθροίζοντας τις τιμές σφαλμάτων των νευρώνων στο επίπεδο εξόδου ως εξής:

$$E(n) = \frac{1}{2}\sum_{j \in C} e_j^2(n)$$

όπου C ο αριθμός των νευρώνων στο επίπεδο εξόδου. Αν N ο αριθμός των ζευγαριών των διανυσμάτων εισόδου-εξόδου στην αρχική διαθέσιμη σειρά δεδομένων, τότε ορίζεται η μέση τετραγωνική ενέργεια σφάλματος:

$$E_{av} = \frac{1}{N}\sum_{n=1}^{N} E(n)$$

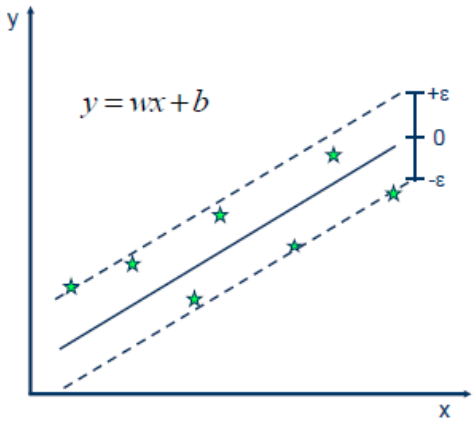
Η παραπάνω σχέση αποτελεί συνάρτηση όλων των ελεύθερων παραμέτρων του νευρωνικού δικτύου, στόχος της εκπαίδευσης του δικτύου είναι να επιλεχθούν οι ελεύθερες παράμετροι όπως βάρη κλπ. που θα ελαχιστοποιούν την τιμής της συνάρτησης η οποία μπορεί να ονομασθεί και αντικειμενική συνάρτηση (objective function) ή συνάρτηση απωλειών (loss function). Η εύρεση ελαχίστου μπορεί να γίνει με διάφορες μεθόδους βελτιστοποίησης αν και όσο πιο πολλοί οι νευρώνες και αντίστοιχα τα βάρη τόσο πιο περίπλοκο γίνεται το πρόβλημα της βελτιστοποίησης. Αφού βρεθεί το ελάχιστο της συνάρτησης η ρύθμιση των βαρών γίνεται με τον αλγόριθμο της οπίσθιας διάδοσης τους σφάλματος (error backpropagation) που στην ουσία είναι εφαρμογή του κανόνα της αλυσίδας στην παραγώγιση και δεν θα αναλυθεί περαιτέρω στην παρούσα εργασία.

2.2.7 Μηχανές Διανυσμάτων Υποστήριξης

Η μέθοδος των μηχανών διανυσμάτων υποστήριξης (Support Vector Machines) άρχισε να αναπτύσσεται τη δεκαετία του εξήντα και σήμερα αποτελεί ένα ισχυρό εργαλείο σε εφαρμογές μηχανικής μάθησης και τεχνητής νοημοσύνης. Ο αλγόριθμος αρχικά δημιουργήθηκε για την αντιμετώπιση προβλημάτων ταξινόμησης (Support Vector Classification) αλλά λίγο αργότερα αναπτύχθηκε ώστε να μπορεί να ανταπεξέλθει και σε προβλήματα παλινδρόμησης (Support Vector Regression). Στην παρούσα εργασία που το ενδιαφέρον επικεντρώνεται στις προβλέψεις, θα γίνει αναφορά στη χρήση της μεθόδου στο πρόβλημα της παλινδρόμησης.

Η βασική ιδέα της μεθόδου έχει ως εξής, έστω ότι είναι διαθέσιμο σετ δεδομένων για εκπαίδευση $(x_1, y_1) \dots (x_\lambda, y_\lambda)$. Στόχος του αλγορίθμου είναι να βρει μια συνάρτηση $f(x)$ που έχει απόκλιση το πολύ ε από τις εξαρτημένες μεταβλητές y_i για όλα τα δεδομένα

εκπαίδευσης και ταυτόχρονα να είναι όσο πιο επίπεδη γίνεται (**Σχήμα 2–9**). Με άλλα λόγια λαμβάνονται υπόψιν μόνο σφάλματα μεγαλύτερα από την τιμή ε.



Σχήμα 2–9: SVM (https://www.saedsayad.com/support_vector_machine_reg.htm)

Στην απλούστερη περίπτωση γραμμικής συνάρτησης f η οποία έχει τη μορφή

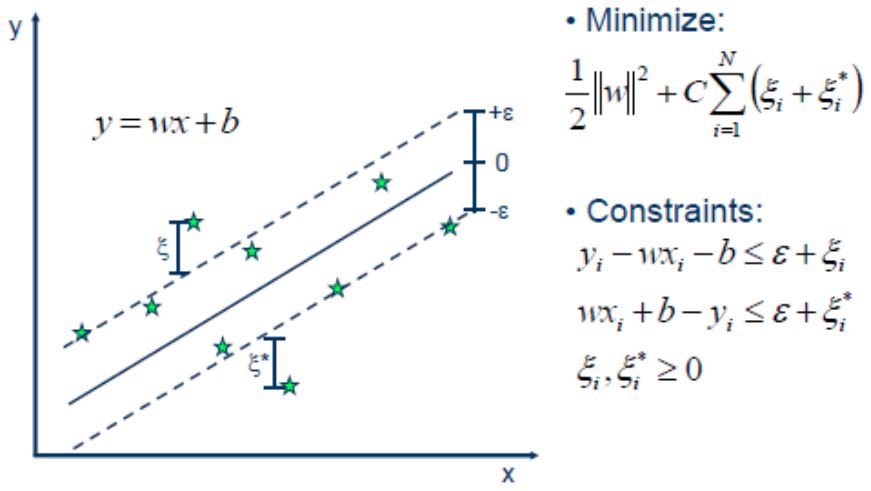
$$f(x) = \langle w, x \rangle + b$$

όπου $\langle\ \rangle$ δηλώνει το εσωτερικό γινόμενο στο χώρο X. Τα μεγέθη w και b είναι άγνωστα. Στόχος είναι να γίνει η συνάρτηση f όσο το δυνατόν πιο επίπεδη και αυτό επιτυγχάνεται ελαχιστοποιώντας το μέτρο $||w||$. Συνεπώς η εκπαίδευση του μοντέλου εκφράζεται ως ένα πρόβλημα κυρτής βελτιστοποίησης που διατυπώνεται ως εξής:

– Ελαχιστοποίηση $\frac{1}{2}\|w\|^2$

– Περιορισμοί $\begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases}$

Η υπόθεση που έγινε στους περιορισμούς είναι ότι υπάρχει η συνάρτηση f και προσεγγίζει όλα τα ζεύγη (x_i, y_i) με ακρίβεια ε, με άλλα λόγια το πρόβλημα κυρτής βελτιστοποίησης είναι επιλύσιμο. Σε περιπτώσεις που πρέπει να επιτραπεί η ύπαρξη κάποιων σφαλμάτων —όταν δεν υπάρχει στο πρόβλημα εφικτή λύση— εισάγονται κάποιες μεταβλητές χαλάρωσης (slack variables) ξ_i, ξ_i^* και το πρόβλημα αλλάζει μορφή (**Σχήμα 2–10**).



Σχήμα 2–10: SVM slack variables
(https://www.saedsayad.com/support_vector_machine_reg.htm)

Το πρόβλημα μπορεί να απλοποιηθεί αν κατασκευαστεί μια συνάρτηση Lagrange από την αντικειμενική συνάρτηση και τους περιορισμούς. Προκύπτει λοιπόν ότι το w εκφράζεται ως γραμμικός συνδυασμός των x_i έτσι δε χρειάζεται να υπολογιστεί ρητά αλλά εκτιμηθούν μόνο τα εσωτερικά γινόμενα μεταξύ δεδομένων, αυτό βοηθάει και στη μη γραμμική επέκταση της μεθόδου.

Σε περιπτώσεις μη γραμμικότητας των δεδομένων επειδή —όπως αναφέρθηκε— ο αλγόριθμος εξαρτάται μόνο από τα εσωτερικά γινόμενα μεταξύ των εισόδων των δεδομένων εκπαίδευσης, μπορούν να χρησιμοποιηθούν συναρτήσεις πυρήνα (kernel functions) που μετασχηματίζουν τα δεδομένα σε χώρο χαρακτηριστικών υψηλότερων διαστάσεων ώστε να μπορεί να γίνει γραμμικός διαχωρισμός.

2.3 Αξιολόγηση προβλέψεων – Σφάλματα

Μεγάλη σημασία κατά τη διαδικασία των προβλέψεων έχει η αξιολόγηση της ακρίβειας μιας πρόβλεψης. Σύμφωνα με την παραπάνω περιγραφή διάφορων τεχνικών προβλέψεων, η αξιολόγηση του μοντέλου με χρήση κάποιων μετρητών των λεγόμενων σφαλμάτων επιδρά τόσο στην επιλογή του κατάλληλου μοντέλου για το εκάστοτε πρόβλημα όσο και στον καθορισμό διάφορων παραμέτρων του μοντέλου λόγου χάρη τον προσδιορισμό του κατάλληλου συντελεστή εξομάλυνσης, την εφαρμογή της μεθόδου ελάχιστων τετραγώνων στις τεχνικές παλινδρόμησης, την επιλογή των βαρών σε ένα νευρωνικό δίκτυο, την εκτίμηση του μέρους κινητού μέσου όρου (MA) των σφαλμάτων ενός μοντέλου ARIMA κ.λπ.

Κατά την επιλογή μοντέλων, συνηθίζεται να χωρίζονται τα διαθέσιμα δεδομένα σε δύο τμήματα, τα δεδομένα εκπαίδευσης (train data) και τα δεδομένα δοκιμής (test data). Τα δεδομένα εκπαίδευσης χρησιμοποιούνται για τον καθορισμό των παραμέτρων της μεθόδου πρόβλεψης ενώ τα δεδομένα δοκιμής για την αξιολόγηση της ακρίβειάς της. Επειδή τα δεδομένα δοκιμής δεν συμμετέχουν στον καθορισμό των προβλέψεων

δηλαδή στην εκπαίδευση του μοντέλου, για αυτό το λόγο ίσως παρέχουν μια αξιόπιστη ένδειξη για το πόσο καλά θα μπορούσε το μοντέλο να προβλέψει πάνω σε νέα δεδομένα (εισόδους) και αν χρειάζεται κάποια βελτίωση ή αντικατάσταση με κάποιο άλλο (Hyndman & Athanasopoulos, 2020).

Όπως αναφέρθηκε προηγουμένως, κάθε χρονοσειρά περιλαμβάνει και μια τυχαία συνιστώσα. Μια καλή μέθοδος πρόβλεψης θα πρέπει να συλλαμβάνει τη συστηματική συνιστώσα, όχι όμως την τυχαία. Η τυχαία συνιστώσα εκδηλώνεται με τη μορφή ενός σφάλματος πρόβλεψης. Το σφάλμα πρόβλεψης (forecast error) δεν εκφράζει κάποιο λάθος αλλά αναπαριστά το μη προβλέψιμο μέρος μιας παρατήρησης κι έτσι μπορεί να ορισθεί για την περίοδο t ως η διαφορά e μεταξύ της πραγματικής παρατήρησης y και της πρόβλεψής της $\hat{y}$ ως εξής:

$$e = y_t - \hat{y_t}$$

Τα σφάλματα πρόβλεψης περιέχουν πολύτιμες πληροφορίες και πρέπει να αναλύονται προσεκτικά για τους ακόλουθους λόγους:

- Με την ανάλυση σφάλματος ελέγχεται αν η τρέχουσα μέθοδος πρόβλεψης προβλέπει με ακρίβεια τη συστηματική συνιστώσα.
- Τα σχέδια έκτακτης ανάγκης (contingency plans) καταστρώνονται με βάση τις προβλέψεις και τα μεγέθη των σφαλμάτων τους. Για παράδειγμα το απόθεμα ασφαλείας σε μια αποθήκη μπορεί να καθορισθεί με την εκτίμηση του σφάλματος μιας πρόβλεψης.

Παρακάτω παρουσιάζονται κάποια μέτρα σφάλματος τα οποία μετρούν την ακρίβεια ενός μοντέλου πρόβλεψης λαμβάνοντας υπόψιν τις αποκλίσεις e_t πολλών περιόδων με διάφορους τρόπους.

Το πιο απλό μέτρο σφάλματος που αναφέρεται στην παρούσα εργασία είναι το **μέσο σφάλμα ME (mean error)** το οποίο ορίζεται ως εξής:

$$ME = \frac{1}{n}\sum_{t=1}^{n} e_t$$

όπου n ο αριθμός των περιόδων για τις οποίες έχει γίνει πρόβλεψη. Συνήθως το συγκεκριμένο μέτρο σφάλματος έχει μικρές τιμές και αυτό γίνεται διότι τα σφάλματα με αρνητικές τιμές τείνουν να αντισταθμίσουν αυτά με θετικές. Συνεπώς η πληροφορία που λαμβάνεται από το συγκεκριμένο μέτρο αναφέρεται στον αν υπάρχει μεροληψία στην πρόβλεψη και όχι στο μέγεθος των χαρακτηριστικών αποκλίσεων για αυτό το λόγο ονομάζεται και μεροληψία (bias).

Για να εξαλειφθεί το παραπάνω πρόβλημα της αντιστάθμισης των αρνητικών τιμών από τις θετικές προτείνονται άλλα μέτρα σφάλματος, ένα από αυτά είναι το **μέσο απόλυτο σφάλμα MAE (mean absolute error)** το οποίο υπολογίζει το μέσο όρο των απόλυτων τιμών των αποκλίσεων e_t και ορίζεται από τον παρακάτω τύπο:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t|$$

Το προηγούμενο σφάλμα είναι πιο εύκολα ερμηνεύσιμο κι έτσι μπορεί να κατανοηθεί καλύτερα από τους μη ειδικούς. Όμως είναι σχετικά δύσκολο να χειρισθεί μαθηματικά και για αυτό το λόγο προτείνεται το **μέσο τετραγωνικό σφάλμα MSE (mean squared error)** το οποίο είναι καταλληλότερο για χρήση στους αλγόριθμους βελτιστοποίησης και ορίζεται από την παρακάτω σχέση:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}e_t^2$$

Παρ' όλα αυτά το MSE έχει μονάδες διαφορετικές από αυτές του μεγέθους για το οποίο γίνεται η πρόβλεψη έτσι για να αποφευχθεί αυτό μπορεί να ορισθεί η **ρίζα του μέσου τετραγωνικού σφάλματος RMSE (root-mean-square error)** ως εξής:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}e_t^2}$$

Η ακρίβεια των προβλέψεων που μετριέται από τα παραπάνω μεγέθη σφάλματος εξαρτάται από την κλίμακα των δεδομένων. Έτσι δεν μπορεί να γίνει σύγκριση μεταξύ διαφορετικών χρονοσειρών και διαφορετικών χρονικών διαστημάτων. Για αυτό το λόγο προτείνονται σχετικά ή ποσοστιαία μέτρα σφάλματος, ένα τέτοιο είναι το **μέσο απόλυτο ποσοστιαίο σφάλμα MAPE (mean absolute percentage error)** το οποίο ορίζεται από τον παρακάτω τύπο:

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right|$$

Το συγκεκριμένο σφάλμα δίνει το ποσοστό της απόκλισης του συνόλου των προβλέψεων από τις πραγματικές παρατηρήσεις. Ιδιαίτερα όταν γίνεται πρόγνωση σε χρονοσειρές για τις οποίες δεν είναι γνωστή η τάξη μεγέθους των δεδομένων είναι πιο εύκολα αντιληπτή η αναφορά σε μια ποσοστιαία τιμή σφάλματος παρά σε μια απόλυτη. Παραδείγματος χάριν μεγαλύτερη σημασία έχει η πληροφορία ότι η πρόγνωση έχει αποκλίνει από τις πραγματικές παρατηρήσεις κατά 7% παρά ότι υπάρχει μέσο σφάλμα στις προβλέψεις με τιμή 178. Βέβαια σε χρονοσειρές με μηδενικές τιμές αυτό το μέτρο δεν είναι κατάλληλο μιας και δεν είναι δυνατόν να υπολογιστεί το ποσοστό σύμφωνα με τη σχέση.

Γενικά υπάρχει πληθώρα μέτρων σφαλμάτων τα οποία είναι καταλληλότερα για κάθε διαφορετική περίπτωση και χρήση, σε αυτήν την εργασία έγινε αναφορά σε κάποια βασικά τα οποία όμως έχουν ευρεία εφαρμογή στην αξιολόγηση προβλέψεων.

3. Παρουσίαση Google Trends

3.1 Γενικά

Η χρήση του διαδικτύου συνεχώς αυξάνεται και πλέον έχει γίνει ένα αναπόσπαστο κομμάτι της καθημερινότητας πολλών ανθρώπων ιδιαίτερα στον ανεπτυγμένο κόσμο. Πιο αναλυτικά οι χρήστες του διαδικτύου στις μέρες μας ανέρχονται σε 4,54 δισεκατομμύρια παγκοσμίως (statista, 2020). Είναι προφανές λοιπόν ότι αυτή η ευρεία χρήση του διαδικτύου μπορεί να αποτελέσει μια πλούσια πηγή πληροφοριών σχετικά με τις συνήθειες και τις δραστηριότητες του κοινού σε αυτό. Κάθε προσωπική επιλογή που γίνεται καταγράφεται, ο τεράστιος αυτός όγκος δεδομένων που δημιουργείται είναι πολύ πιο εύκολο να αποθηκευτεί και να προσπελασθεί καθώς εξελίσσεται η τεχνολογία. Όσο ο όγκος αυτός μεγαλώνει τόσο μικραίνει το ποσοστό αυτού για το οποίο οι άνθρωποι μπορούν να βγάλουν κάποιο νόημα. Με χρήση κατάλληλων μεθόδων είναι εφικτό να αναγνωρισθούν κάποια μοτίβα από αυτά τα δεδομένα ώστε να εξαχθούν γνώσεις και συμπεράσματα χρήσιμα σε τομείς όπως η βιομηχανία, το εμπόριο κ.λπ. (Witten & Frank, 2005).

Σύμφωνα με έρευνα της Ελληνικής Στατιστικής Αρχής για το πρώτο τρίμηνο του 2020 (ELSTAT, 2020), ο κυρίαρχος λόγος χρήσης του διαδικτύου για ποσοστό 88,9% των ερωτηθέντων είναι η αναζήτηση πληροφοριών σχετικών με προϊόντα και υπηρεσίες. Η χρήση του διαδικτύου είναι συνυφασμένη με τη χρήση μηχανών αναζήτησης, καθώς ο μέσος χρήστης περιηγείται στον ιστό μέσω αυτών. Ακόμα και για να εισέλθει σε γνωστούς του ιστοτόπους αντί να πληκτρολογήσει τη διεύθυνση της ιστοσελίδας στο κατάλληλο πεδίο του προγράμματος φυλλομετρητή, πληκτρολογεί λέξεις κλειδιά στη μηχανή αναζήτησης και εισέρχεται στην ιστοσελίδα που επιθυμεί από το αντίστοιχο σύνδεσμο στα αποτελέσματα. Δηλαδή οι μηχανές αναζήτησης μπορούν να χαρακτηριστούν ως παρακάμψεις για την πρόσβαση σε ιστοσελίδες.

Ανάμεσα στις μηχανές αναζήτησης το μεγαλύτερο μερίδιο της αγοράς αδιαμφισβήτητα κατέχει αυτή της Google με ποσοστό που πλησιάζει το 92,05% παγκοσμίως και το 97,96% στην Ελλάδα (statcounter, 2020). Σαφώς υπάρχουν ορισμένες εξαιρέσεις όπου η μηχανή της Google δεν έχει την πρωτοκαθεδρία στην αγορά όπως για παράδειγμα αυτές της Κίνας, της Άπω Ανατολής και της Ρωσίας εξαιτίας γλωσσικών διαφορών αλλά και πολιτικών λόγων (icrossing, 2015). Υπολογίζεται ότι ο αριθμός αναζητήσεων που δέχεται η μηχανή της Google ξεπερνά τα 2 τρισεκατομμύρια το χρόνο και τα 6 δισεκατομμύρια τη μέρα (internet live stats, 2020), (Wikipedia, 2020).

Η Google έχει δημιουργήσει και παρέχει στο κοινό μια σειρά από διάφορα εργαλεία μερικά από τα οποία αφορούν τη μηχανή αναζήτησης, ένα από αυτά είναι το Google Trends. Το Google Trends είναι μια εφαρμογή που λάνσαρε η εταιρεία Google στις 11 Μαΐου του 2006 η οποία παρέχει δεδομένα σχετικά με τη δημοτικότητα αναζητούμενων όρων στη μηχανή αναζήτησής της. Η εφαρμογή είναι προσβάσιμη από την ιστοσελίδα με τη διεύθυνση https://trends.google.com/. Τον Αύγουστο του 2008 η εταιρεία διέθεσε στο κοινό και το Google Insights for Search μια πιο εκλεπτυσμένη και προηγμένη

υπηρεσία που απεικόνιζε τάσεις των δεδομένων αναζήτησης αλλά το Σεπτέμβρη του 2012 η εταιρεία συγχώνευσε τις δυο εφαρμογές σε αυτή του Google Trends (Wikipedia, 2020).

3.2 Βασικά στοιχεία εφαρμογής

Το Google Trends δεν δείχνει τον απόλυτο αριθμό αναζητήσεων για κάθε όρο αλλά έναν αριθμό ο οποίος δηλώνει τη σχετική δημοτικότητα του όρου συγκριτικά με τη δημοτικότητα του μια άλλη χρονική στιγμή ή ανά υποπεριοχή κομμάτι της ευρύτερης περιοχής που έχει ορισθεί κατά την αναζήτησή του στην εφαρμογή. Ο τρόπος που προκύπτουν αυτοί οι αριθμοί θα αναλυθεί στη συνέχεια αυτής της εργασίας.

3.2.1 Λειτουργίες εφαρμογής

Οι βασικές δυνατότητες που προσφέρει το Google Trends μόλις εισαχθεί κάποιος όρος αναζήτησης είναι οι ακόλουθες:

- **Απεικόνιση ενδιαφέροντος του όρου με την πάροδο του χρόνου.** Δηλαδή τα δεδομένα που προκύπτουν αναπαριστούν τη δημοτικότητα του όρου συγκριτικά με άλλες χρονικές περιόδους που ορίζονται από την επιλογή χρονικού παραθύρου κατά την εισαγωγή του όρου.
- **Απεικόνιση ενδιαφέροντος ανά περιοχή.** Πιο αναλυτικά παρουσιάζεται σε ποιες περιοχές είναι δημοφιλέστερος ο όρος για τον οποίο έγινε αναζήτηση. Οι περιοχές αυτές ορίζονται ως τμήματα μιας ευρύτερης περιοχής που επιλέγεται κατά την εισαγωγή του όρου.
- **Εμφάνιση σχετικών θεμάτων με τον όρο αναζήτησης.** Δηλαδή παρουσιάζονται θέματα που σύμφωνα με τη Google άλλοι χρήστες αναζήτησαν μαζί με τον αναζητούμενο όρο. Μπορεί να γίνει επιλογή μεταξύ κορυφαίων και ανερχόμενων θεμάτων. Τα κορυφαία είναι τα δημοφιλέστερα θέματα ενώ τα ανερχόμενα αυτά με την μεγαλύτερη αύξηση σε συχνότητα αναζήτησης από την τελευταία χρονική περίοδο.
- **Εμφάνιση σχετικών ερωτημάτων με τον όρο αναζήτησης.** Δηλαδή παρουσιάζονται ερωτήματα που σύμφωνα με τη Google άλλοι χρήστες αναζήτησαν μαζί με τον αναζητούμενο όρο. Μπορεί να γίνει επιλογή μεταξύ κορυφαίων και ανερχόμενων ερωτημάτων. Τα κορυφαία είναι τα δημοφιλέστερα ερωτήματα ενώ τα ανερχόμενα αυτά με την μεγαλύτερη αύξηση σε συχνότητα αναζήτησης από την τελευταία χρονική περίοδο.
- **Σύγκριση δημοτικότητας** δύο ως πέντε όρων τόσο ανά χρονική διάρκεια όσο και ανά περιοχή.

Διευκρινίζεται ότι τα «θέματα» ορίζονται από την εφαρμογή ως ομάδες όρων που μοιράζονται την ίδια έννοια σε κάποια γλώσσα. Δηλαδή η εφαρμογή ομαδοποιεί αυτόματα όρους όπως «London», «capital of UK» και «Λονδίνο» που έχουν ίδια σημασία. Επίσης άλλες δυνατότητες της εφαρμογής είναι η προβολή των

δημοφιλέστερων αναζητήσεων ανά έτος, η παροχή πληροφοριών σχετικά με τις ημερήσιες τάσεις αναζήτησης και η παρουσίαση διάφορων ιστοριών από την ομάδα News Lab της Google που παρέχουν πρόσθετες πληροφορίες οι οποίες περιλαμβάνονται στα δεδομένα. Οι βασικές δυνατότητες της εφαρμογής θα περιγραφούν εκτενέστερα κατά την παρουσίαση του περιβάλλοντος διεπαφής της ιστοσελίδας όπου θα γίνει λόγος και για επιπλέον παραμέτρους που μπορούν να ορισθούν κατά την αναζήτηση συγκεκριμένων όρων.

3.2.2 Δεδομένα Google Trends

Όπως αναφέρθηκε και προηγουμένως το Google Trends δεν παρέχει τον απόλυτο αριθμό αναζητήσεων ενός όρου αλλά τη σχετική δημοτικότητά του με την πάροδο του χρόνου, ανά περιοχή κτλ. Αυτή η σχετική δημοτικότητα υποδηλώνεται με μια βαθμολογία με τιμές από το 0 ως το 100 η οποία ορίζεται από το Google Trends με τον ακόλουθο τρόπο. Αρχικά γίνεται πρόσβαση σε ένα δείγμα πραγματικών αιτημάτων αναζήτησης που υποβάλλονται στο Google, το οποίο σε μεγάλο βαθμό δεν έχει φιλτραριστεί. Είναι ανώνυμο (δεν γίνεται ταυτοποίηση χρηστών), κατηγοριοποιημένο (καθορίζοντας το θέμα ενός ερωτήματος αναζήτησης) και συγκεντρωτικό (ομαδοποιημένο). Αυτό επιτρέπει να προβληθεί το ενδιαφέρον για ένα συγκεκριμένο θέμα από ολόκληρο τον κόσμο ή από μικρότερες γεωγραφικές περιοχές, έως και επιπέδου πόλης. Το δείγμα αυτό αν και δεν είναι γνωστή η ακριβής μεθοδολογία με την οποία προκύπτει, θεωρείται αντιπροσωπευτικό από τη Google καθώς στη συγκεκριμένη πλατφόρμα διενεργούνται δισεκατομμύρια αναζητήσεις κάθε ημέρα. Επίσης η παροχή πρόσβασης σε ολόκληρο το σύνολο των δεδομένων θα καθυστερούσε λόγω του τεράστιου όγκου του. Υπάρχουν δύο δείγματα δεδομένων του Google Trends στα οποία παρέχεται πρόσβαση:

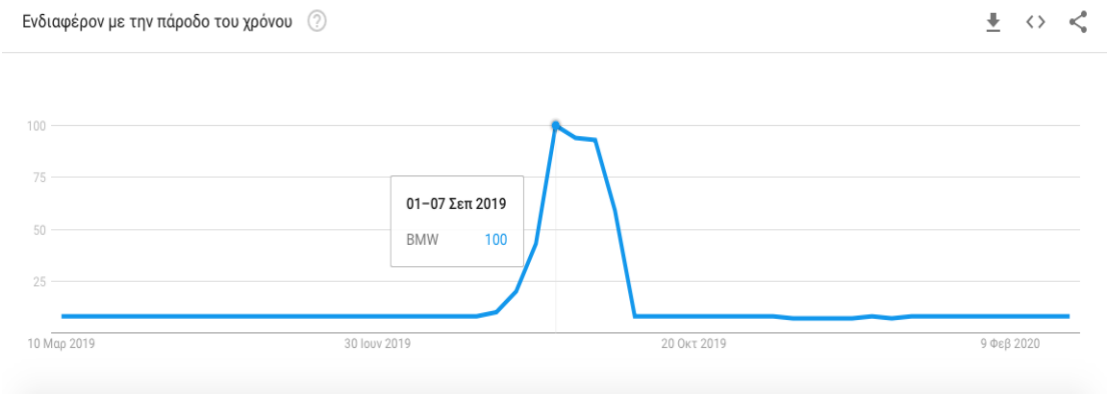
- Τα δεδομένα σε πραγματικό χρόνο είναι ένα δείγμα που καλύπτει τις τελευταίες επτά ημέρες.
- Τα δεδομένα μη πραγματικού χρόνου είναι ένα ξεχωριστό δείγμα από δεδομένα σε πραγματικό χρόνο και καλύπτουν την χρονική περίοδο από το 2004 έως και 36 ώρες πριν από την αναζήτησή.

Έπειτα για κάθε σημείο των δεδομένων υπολογίζονται οι αναζητήσεις του όρου προς τις συνολικές αναζητήσεις του δείγματος στη γεωγραφική περιοχή και στο χρονικό εύρος που αντιπροσωπεύει, προκειμένου να συγκριθεί η σχετική δημοτικότητά του. Διαφορετικά οι περιοχές με τον μεγαλύτερο όγκο αναζητήσεων θα είχαν πάντα υψηλότερη κατάταξη. Οι αριθμοί που υπολογίστηκαν κανονικοποιούνται σε μια κλίμακα από το 0 ως το 100. Η τιμή 100 αντιστοιχεί στη υψηλότερη δημοτικότητα για τον όρο ενώ η τιμή 50 σημαίνει ότι ο όρος στο συγκεκριμένο σημείο έχει τη μισή δημοτικότητα από την υψηλότερη. Η βαθμολογία 0 σημαίνει ότι δεν υπήρχαν αρκετά δεδομένα για το συγκεκριμένο σημείο.

Συνεπώς σύμφωνα με την παραπάνω διαδικασία τα δεδομένα που προκύπτουν σχετικά με τη δημοτικότητα του όρου κατά την πάροδο του χρόνου συνθέτουν μια χρονοσειρά που έχει ως υψηλότερη τιμή το 100 και αντιστοιχεί στη χρονική στιγμή με την υψηλότερη

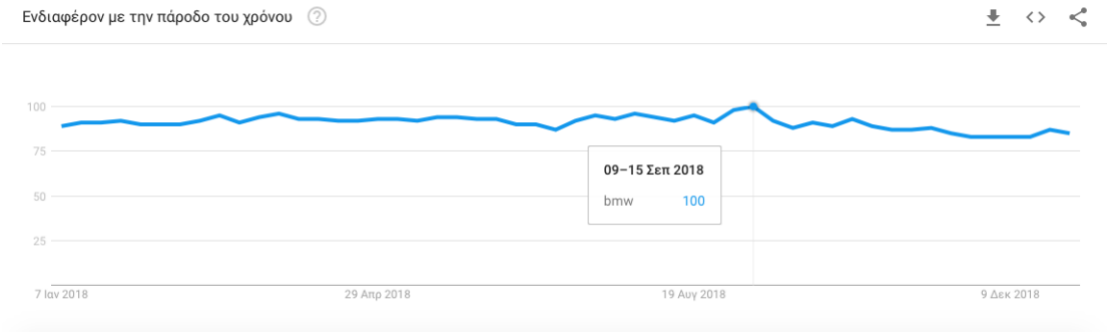
δημοτικότητα του όρου και χαμηλότερη το 0 που αντιστοιχεί σε χρονικές στιγμές όπου τα δεδομένα δεν είναι επαρκή. Δηλαδή αν ορισθεί διαφορετικό χρονικό εύρος τότε θα αλλάξουν και οι τιμές της χρονοσειράς με το 100 να αντιστοιχεί στο σημείο με το υψηλότερο αριθμό αναζητήσεων ανάμεσα στα άλλα του εκάστοτε ορισμένου χρονικού διαστήματος.

Για παράδειγμα αν γίνει αναζήτηση για τον όρο «bmw» παγκοσμίως και οριστεί το χρονικό εύρος ως το έτος 2019 προκύπτει η ακόλουθη χρονοσειρά της οποίας η γραφική παράσταση παρατίθεται στο **Σχήμα 3–1** όπως είναι εμφανές η μέγιστη τιμή βρίσκεται στο σημείο της εβδομάδας 01-07 Σεπ 2019.



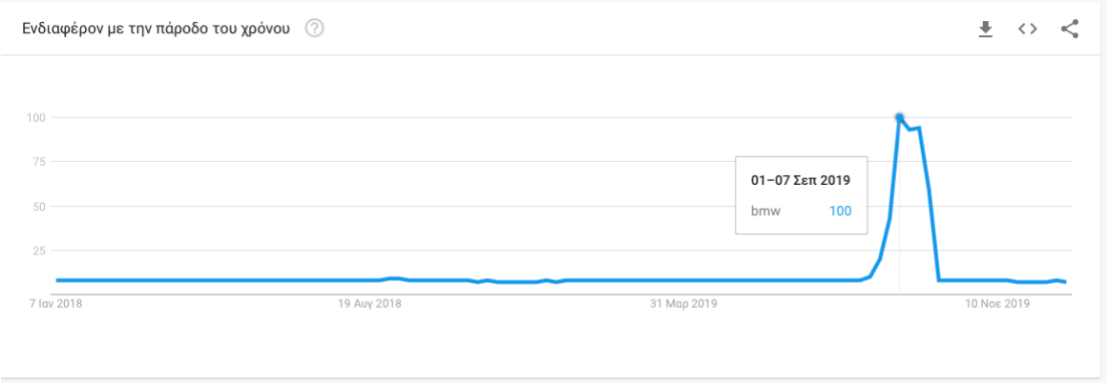
Σχήμα 3–1: Δεδομένα αναζήτησης όρου "bmw" (trends.google.com)

Αν τώρα ορισθεί κάποιο διαφορετικό χρονικό εύρος για τον ίδιο όρο όπως το έτος 2018 τότε προκύπτει η ακόλουθη χρονοσειρά (**Σχήμα 3–2**) όπου φαίνεται πως η μέγιστη τιμή 100 αντιστοιχεί στην εβδομάδα 09-15 Σεπ 2018.



Σχήμα 3–2: Αναζήτηση όρου "bmw" για χρονικό εύρος έτος 2018 (trends.google.com)

Τώρα αν ορισθεί χρονικό εύρος από την αρχή του έτους 2018 ως το τέλος του έτους 2019 προκύπτει η ακόλουθη χρονοσειρά (**Σχήμα 3–3**) με μέγιστη τιμή 100 πάλι την χρονική στιγμή 01-07 Σεπ 2019 σε εβδομαδιαία δεδομένα. Προφανώς το μέγιστο της δεύτερης χρονοσειράς δεν υπάρχει πλέον μιας και το μέγιστο της πρώτης συγκριτικά το ξεπερνά κατά πολύ σε δημοτικότητα.

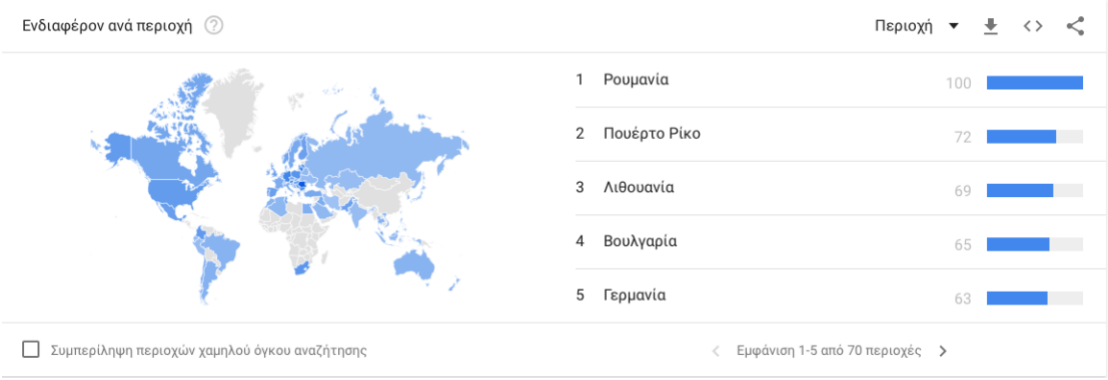


Σχήμα 3–3: Αναζήτηση όρου "bmw" για χρονικό εύρος 2018-2019 (trends.google.com)

Εν κατακλείδι πρέπει ξανά να επισημανθεί ότι τα δεδομένα αναπαριστούν το σχετικό ενδιαφέρον. Όταν εμφανιστεί απότομη αύξηση στις τιμές της χρονοσειράς κατά συνέπεια αιχμή (spike) στο διάγραμμα όπως στο **Σχήμα 3–3** σημαίνει ότι τη δεδομένη χρονική στιγμή διεξήχθησαν περισσότερες αναζητήσεις για το συγκεκριμένο όρο συγκριτικά με αυτές προηγούμενων στιγμών.

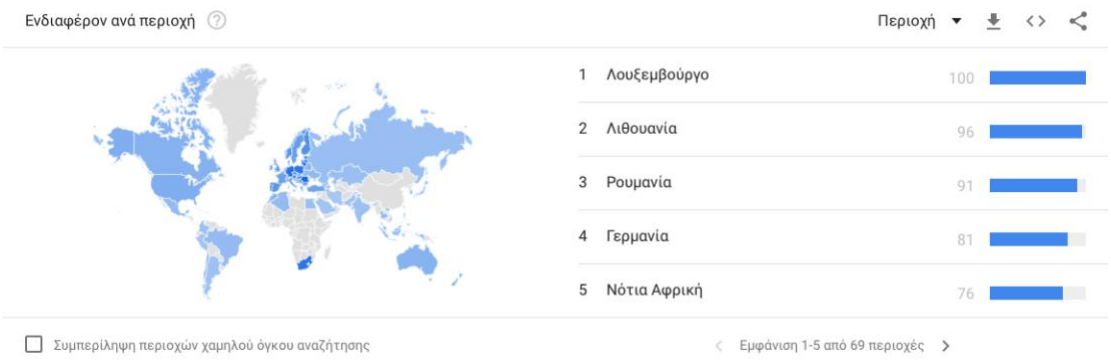
Όσον αφορά το ενδιαφέρον ανά περιοχή προκύπτουν δεδομένα που δείχνουν σε ποιες τοποθεσίες ήταν πιο δημοφιλής ο όρος που αναζητήθηκε κατά την επιλεγμένη χρονική περίοδο. Ακολουθώντας ανάλογη διαδικασία με την παραπάνω οι τιμές υπολογίζονται σε μια κλίμακα από το 0 έως το 100, όπου η τιμή 100 αντιστοιχεί στην τοποθεσία με την περισσότερη δημοτικότητα ως μέρος του συνόλου των αναζητήσεων σε αυτήν την τοποθεσία, η τιμή 50 υποδεικνύει μια τοποθεσία με τη μισή δημοτικότητα. Η τιμή 0 υποδεικνύει μια τοποθεσία όπου δεν υπήρχαν αρκετά δεδομένα για τον συγκεκριμένο όρο. Σημειώνεται ότι οι υψηλότερες τιμές αντιστοιχούν σε υψηλότερο ποσοστό ερωτημάτων, όχι σε υψηλότερο απόλυτο αριθμό. Επομένως, μια πολύ μικρή χώρα όπου το 80% των ερωτημάτων έχουν να κάνουν με «μπανάνες» θα έχει τη διπλάσια βαθμολογία από μια πολύ μεγάλη χώρα, όπου μόνο το 40% των ερωτημάτων έχουν να κάνουν με τον όρο «μπανάνες».

Πάλι αν αναζητηθούν δεδομένα για τον όρο «bmw» παγκοσμίως με ορισμένη χρονική περίοδο τους τελευταίους 12 μήνες λαμβάνεται το παρακάτω **Σχήμα 3–4** στο οποία φαίνεται το ενδιαφέρον για τον όρο ανά χώρα. Στην προκειμένη περίπτωση η μέγιστη τιμή 100 αντιστοιχεί στη χώρα της Ρουμανίας δηλαδή εκεί υπήρχε το μεγαλύτερο ενδιαφέρον για τον όρο συγκριτικά με τις άλλες χώρες.



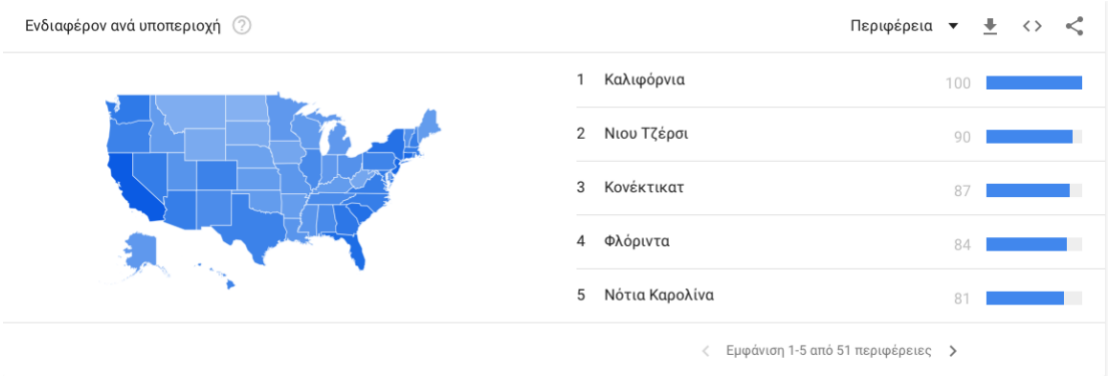
Σχήμα 3–4: Αναζήτηση όρου "bmw" ενδιαφέρον ανά χώρα (trends.google.com)

Επισημαίνεται πάλι ότι οι βαθμολογίες αντικατοπτρίζουν το σχετικό ενδιαφέρον. Για παράδειγμα αν για τον ίδιο όρο η χρονική περίοδος ορισθεί για τις τελευταίες 90 ήμερες παγκοσμίως προκύπτουν τα ακόλουθα στοιχεία όπου η μέγιστη τιμή 100 αντιστοιχεί στη χώρα του Λουξεμβούργου (**Σχήμα 3–5**).



Σχήμα 3–5: Όρος "bmw" ενδιαφέρον ανά χώρα τελευταίες 90 μέρες (trends.google.com)

Ακόμα αν αναζητηθούν δεδομένα για τον ίδιο όρο αλλά επιλεχθεί άλλη περιοχή συγκεκριμένα αυτή των Η.Π.Α για χρονικό διάστημα των τελευταίων 90 ημερών προκύπτουν τα ακόλουθα δεδομένα όπου η μέγιστη τιμή 100 αντιστοιχεί στην πολιτεία της Καλιφόρνιας (**Σχήμα 3–6**).



Σχήμα 3–6: Όρος "bmw" ενδιαφέρον ανά πολιτεία ΗΠΑ (trends.google.com)

Προφανώς σε κάθε περίπτωση από αυτές τις τρεις αναζητήσεις που παρουσιάστηκαν η τιμή 100 δεν αντικατοπτρίζει την ίδια δημοτικότητα μιας και αποτελεί όπως αναφέρθηκε ένα σχετικό μέγεθος. Τα δεδομένα του Google Trends αντικατοπτρίζουν τις αναζητήσεις

που κάνουν οι χρήστες στο Google καθημερινά, αλλά μπορεί επίσης να αντικατοπτρίζουν μη κανονική δραστηριότητα αναζήτησης, όπως αυτοματοποιημένες αναζητήσεις ή ερωτήματα που ενδέχεται να σχετίζονται με προσπάθειες εμφάνισης ανεπιθύμητου περιεχομένου στα αποτελέσματα αναζήτησης. Σύμφωνα με τη Google αν και υπάρχουν μηχανισμοί για τον εντοπισμό και το φιλτράρισμα μη κανονικής δραστηριότητας, αυτές οι αναζητήσεις ενδέχεται να διατηρηθούν στο Google Trends ως μέτρο ασφαλείας. Επεξηγείται ότι αν διενεργούταν φιλτράρισμα σε αυτές τις αναζητήσεις τότε θα γινόταν αντιληπτό στα άτομα που τις είχαν υποβάλει ότι αυτές είχαν εντοπιστεί. Συνεπώς θα ήταν δυσκολότερο να παραμείνει η δραστηριότητα αυτή φιλτραρισμένη από άλλα προϊόντα της Αναζήτησης Google, όπου τα δεδομένα αναζήτησης υψηλής πιστότητας θεωρούνται πιο σημαντικά. Αν ληφθούν υπόψη τα παραπάνω πρέπει να γίνει κατανοητό ότι τα δεδομένα του Google Trends δεν αντικατοπτρίζουν ακριβώς τη δραστηριότητα αναζήτησης.

Η Google δηλώνει πως πραγματοποιεί φιλτράρισμα σε ορισμένους τύπους αναζητήσεων, όπως:

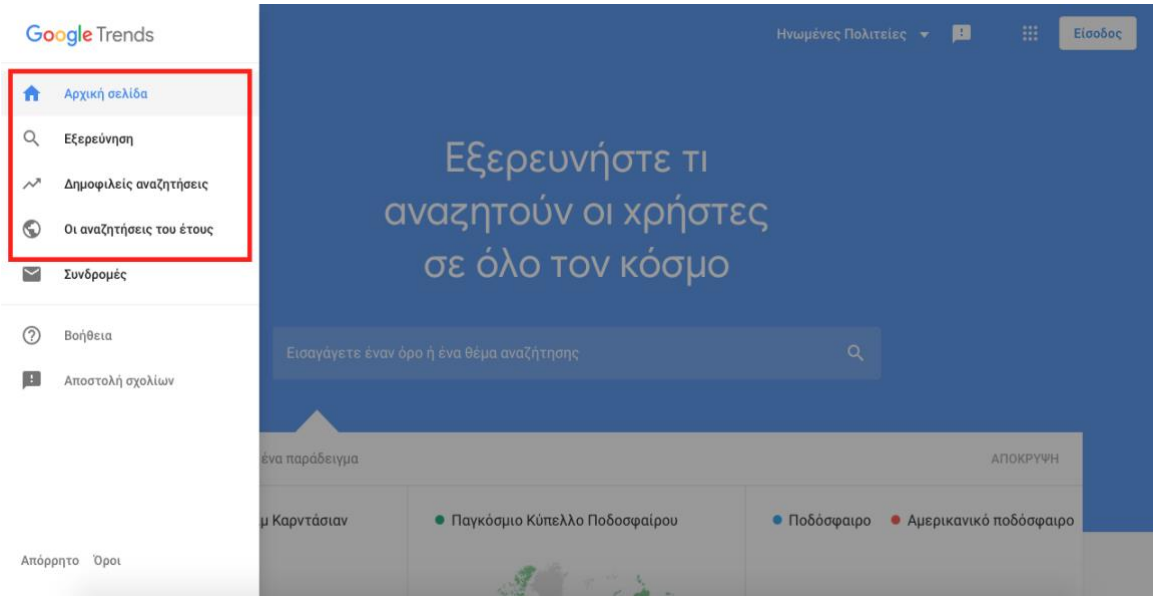
- Αναζητήσεις που πραγματοποιούνται από λίγα άτομα. Το Google Trends εμφανίζει μόνο δεδομένα για δημοφιλείς όρους. Συνεπώς οι όροι που έχουν χαμηλό όγκο αναζητήσεων εμφανίζονται ως «0»
- Διπλότυπες αναζητήσεις. Δηλαδή διαγράφονται οι επαναλαμβανόμενες αναζητήσεις που έχουν πραγματοποιηθεί από το ίδιο άτομο σε σύντομο χρονικό διάστημα.
- Ειδικούς χαρακτήρες. Δηλαδή απομακρύνονται ερωτήματα που περιέχουν αποστρόφους και άλλους ειδικούς χαρακτήρες (Google , 2020).

Από τα παραπάνω ενδιαφέρον έχει η μη κανονική δραστηριότητα συγκεκριμένα οι αυτοματοποιημένες αναζητήσεις ο ρόλος και οι επιπτώσεις τους στη διαδικασία της πρόγνωσης ζήτησης.

3.3 Περιβάλλον διεπαφής χρήστη

Το περιεχόμενο της σελίδας όπως φαίνεται και στο **Σχήμα 3–7** χωρίζεται στις εξής ενότητες:

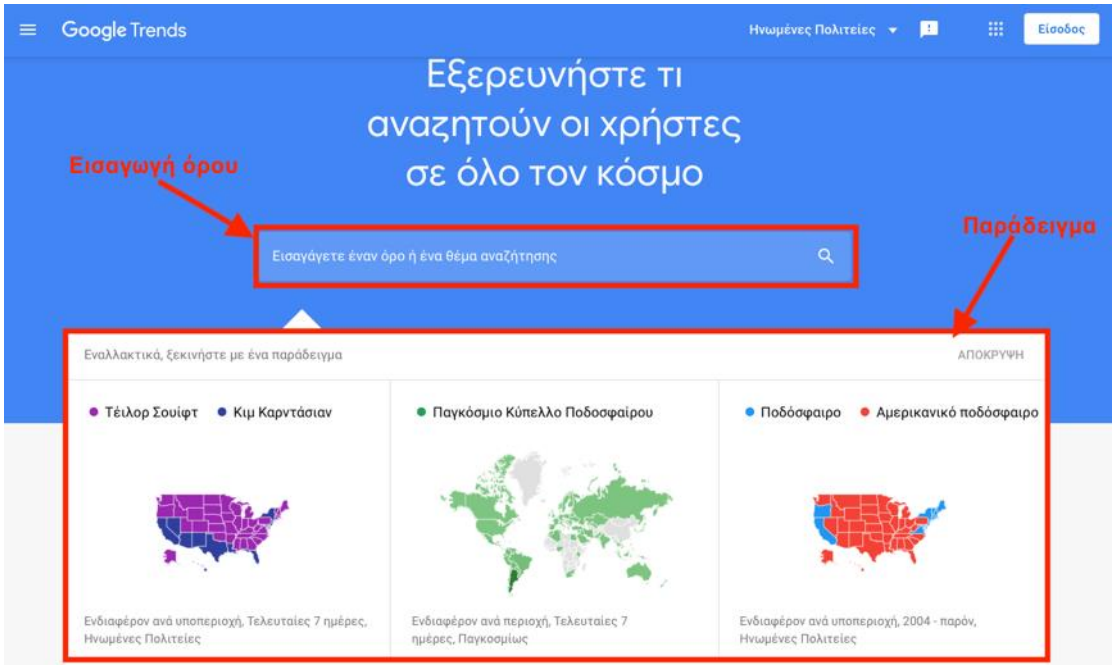
- Αρχική σελίδα
- Εξερεύνηση
- Δημοφιλείς αναζητήσεις
- Οι αναζητήσεις τους έτους



Σχήμα 3–7: Ενότητες Google Trends (trends.google.com)

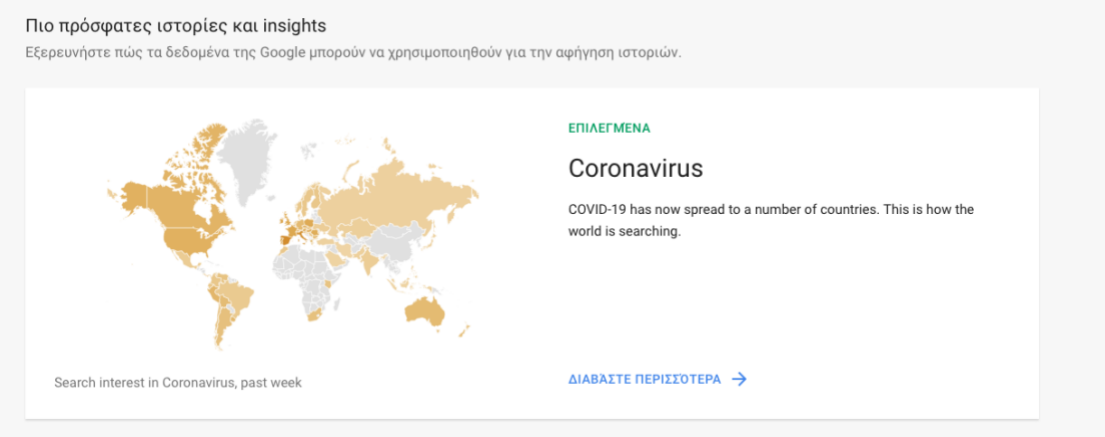
3.3.1 Αρχική σελίδα

Κατά την είσοδο στην ιστοσελίδα ο χρήστης βλέπει την Αρχική Σελίδα εκεί βρίσκονται τα πεδία και το ακόλουθο περιεχόμενο: Στην κορυφή της σελίδας όπως φαίνεται στο σχήμα υπάρχει πεδίο για εισαγωγή όρου αναζήτησης αλλά και η δυνατότητα επιλογής κάποιου παραδείγματος για εκμάθηση της υπηρεσίας (**Σχήμα 3–8**).



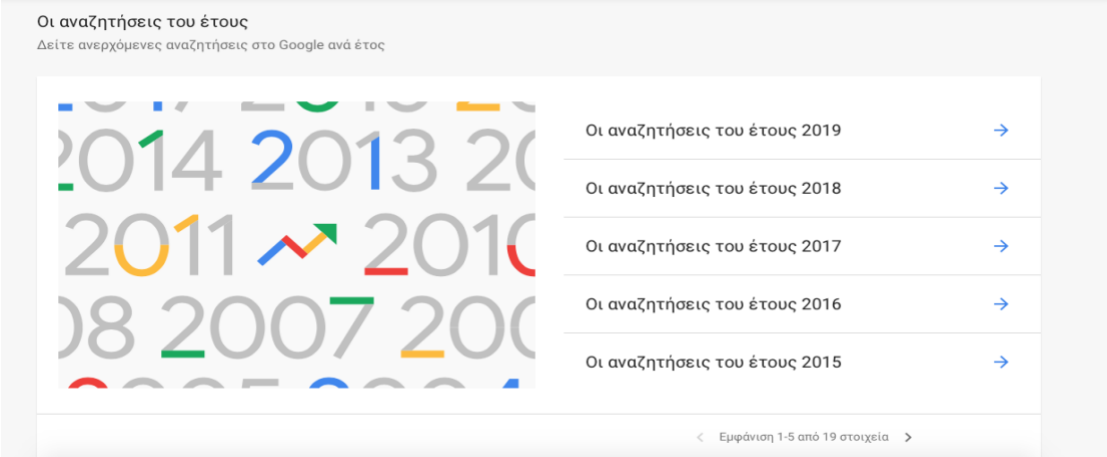
Σχήμα 3–8: Αρχική σελίδα εφαρμογής (trends.google.com)

Παρακάτω (**Σχήμα 3–9**) βρίσκεται περιεχόμενο σχετικό με πρόσφατες ιστορίες όπου υπάρχουν δεδομένα σχετικά με θέματα τα οποία έχουν αυξημένο ενδιαφέρον την παρούσα στιγμή με πρόσθετες πληροφορίες από την ομάδα της Google.

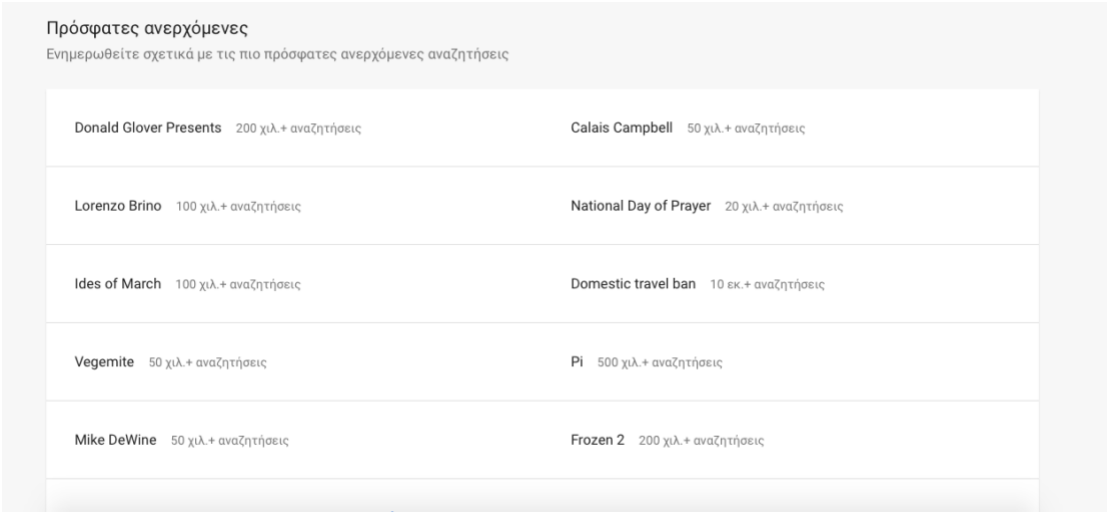


Σχήμα 3–9: Πρόσφατες ιστορίες αρχική σελίδα (trends.google.com)

Έπειτα υπάρχει περιεχόμενο σχετικό με τις πρόσφατες ανερχόμενες αναζητήσεις (**Σχήμα 3–10**) δηλαδή όπως πρωτύτερα έχει αναφερθεί τις αναζητήσεις οι οποίες είχαν τη μεγαλύτερη αύξηση δημοτικότητας συγκριτικά με την αμέσως προηγούμενη χρονική περίοδο.



Σχήμα 3–10: Πρόσφατες ανερχόμενες αναζητήσεις αρχική σελίδα (trends.google.com)



Σχήμα 3–11: Δημοφιλέστερες αναζητήσεις του έτους αρχική σελίδα (trends.google.com)

Τέλος στην αρχική σελίδα υπάρχει πεδίο που αναφέρεται στις δημοφιλέστερες αναζητήσεις κάθε έτους από το 2001 ως σήμερα (**Σχήμα 3–11**).

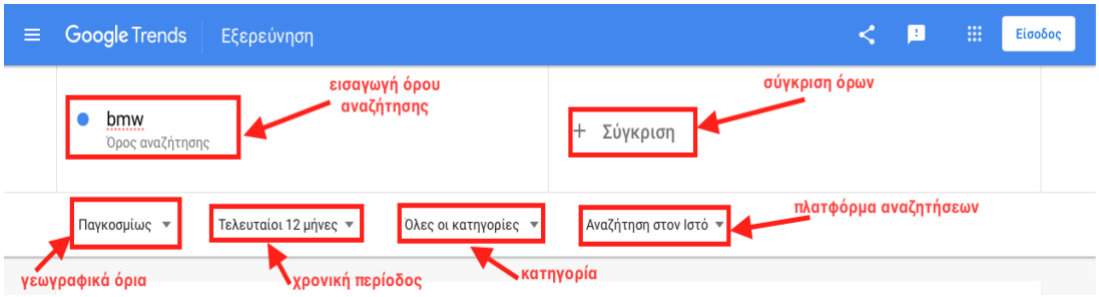
3.3.2 Εξερεύνηση

Η ενότητα η οποία έχει και μεγαλύτερη σημασία για τη συγκεκριμένη εργασία είναι αυτή της εξερεύνησης. Εκεί υπάρχει πλαίσιο για την εισαγωγή αναζητούμενων όρων αλλά και διάφορα άλλα πεδία και κουμπιά με τα οποία μπορεί να γίνει παραμετροποίηση της αναζήτησης όσον αφορά

- Τα γεωγραφικά όρια που αναζητήθηκε ο όρος
- Τη χρονική περίοδο αναζήτησης
- Την κατηγορία
- Την πηγή προέλευσης των αναζητήσεων

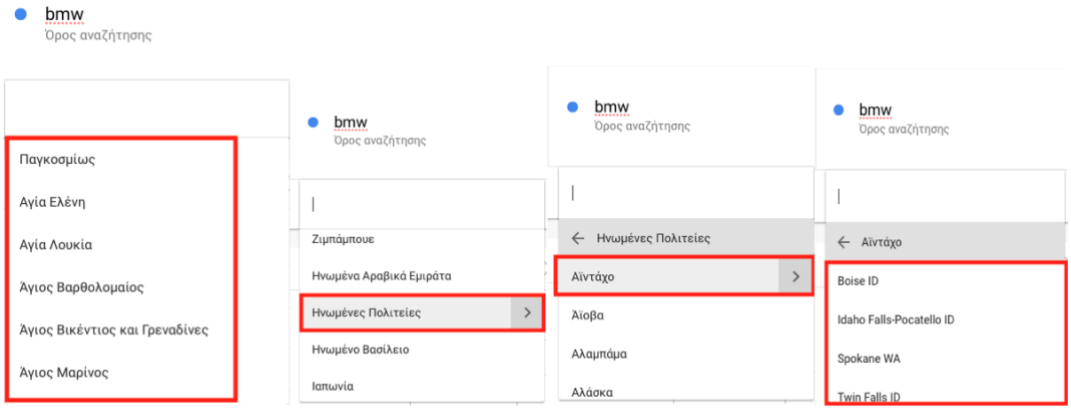
Επίσης αφού εισαχθεί κάποιος όρος εμφανίζεται και πρόσθετο πεδίο εισαγωγής όρων ώστε να γίνει σύγκριση. Αν εισαχθεί και δεύτερος όρος εμφανίζεται και άλλο πεδίο για σύγκριση τριών όρων αυτήν τη φορά και ούτω καθεξής με μέγιστο πλήθος τους πέντε.

Στο ακόλουθο (**Σχήμα 3–12**) φαίνονται επισημασμένα τα προαναφερθέντα πεδία αφού έχει εισαχθεί προς αναζήτηση δεδομένων ο όρος «bmw».



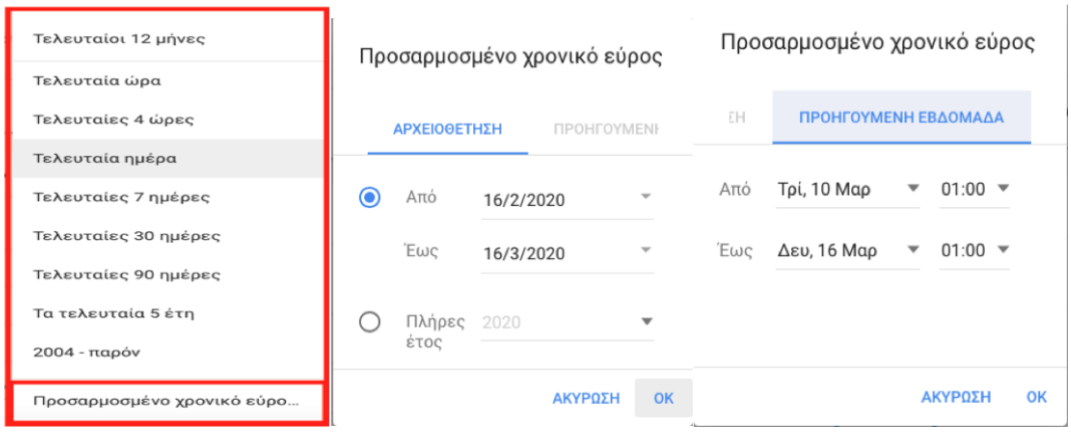
Σχήμα 3–12: Ενότητα Εξερεύνηση (trends.google.com)

Στο πεδίο των γεωγραφικών ορίων μπορεί να ορισθεί ο χώρος που έγιναν οι αναζητήσεις του όρου για τον οποίο ζητούνται δεδομένα. Πιο αναλυτικά μπορεί να ορισθεί να γίνει αναζήτηση δεδομένων σε παγκόσμια κλίμακα, ανά χώρα, ανά γεωγραφικό διαμέρισμα κάποιων χωρών και ανά πόλη η τελευταία επιλογή είναι δυνατή μόνο στις Ηνωμένες Πολιτείες (**Σχήμα 3–13**).



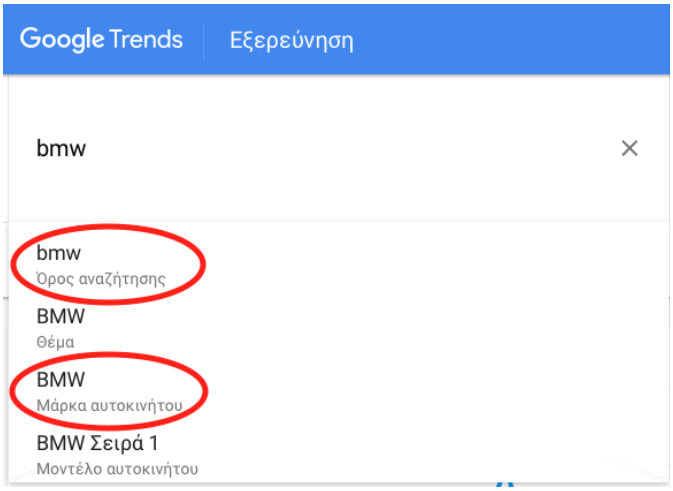
Σχήμα 3–13: Επιλογή περιοχής αναζητήσεων (trends.google.com)

Στο πεδίο χρονικής περιόδου ορίζεται το χρονικό εύρος κατά το οποίο εξάγονται δεδομένα σχετικά με τη δημοτικότητα των όρων αναζήτησης. Υπάρχουν επιλογές τυποποιημένων χρονικών περιόδων αλλά και προσαρμοσμένου χρονικού εύρους τόσο για δεδομένα πραγματικού όσο και για δεδομένα μη πραγματικού χρόνου, όπως ορίστηκαν παραπάνω (**Σχήμα 3–14**).

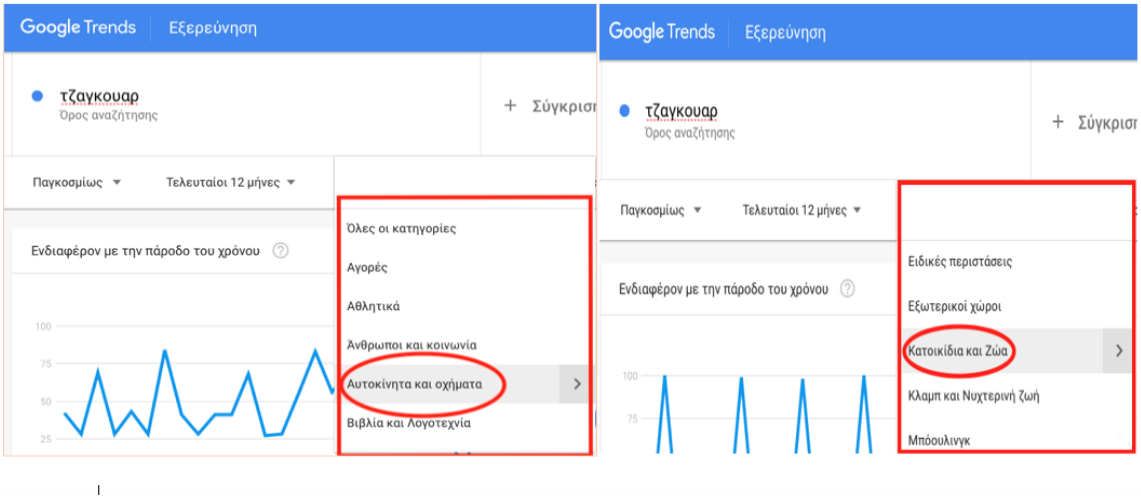


Σχήμα 3–14: Επιλογή χρονικού εύρους (trends.google.com)

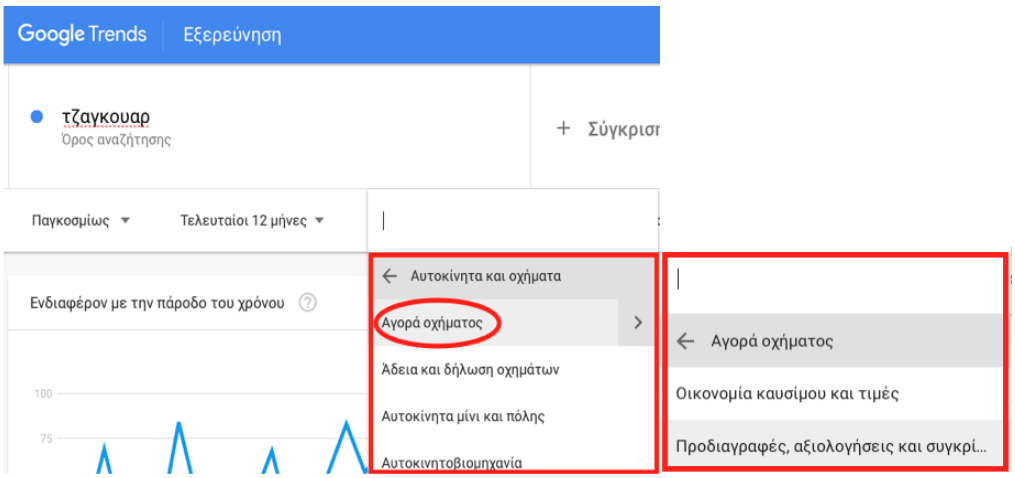
Αν χρησιμοποιηθεί το Google Trends για αναζήτηση μιας λέξης που έχει πολλές σημασίες, τα αποτελέσματα μπορούν να φιλτραριστούν σε μια συγκεκριμένη κατηγορία ώστε να εξαχθούν δεδομένα για τη σωστή έννοια της λέξης. Αυτό μπορεί να γίνει και κατά την εισαγωγή του όρου με λειτουργία αυτόματης επιλογής κατά την πληκτρολόγησή του στο πεδίο αναζήτησης (**Σχήμα 3–15**), αλλά και με τον ορισμό κάποιας κατηγορίας από το πεδίο κατηγορίες (**Σχήμα 3–16**). Για παράδειγμα αν αναζητηθεί ο όρος «τζάγκουαρ», μπορεί να προστεθεί μια κατηγορία για να διευκρινιστεί αν εννοείται το ζώο ή τον κατασκευαστή αυτοκινήτων. Υπάρχει μεγάλο πλήθος επιλογών κατηγορίας αλλά και υποκατηγοριών αυτής για κάθε όρο (**Σχήμα 3–17**), δε θα γίνει όμως λεπτομερής περιγραφή αυτού καθώς η Google αλλάζει τη σύνθεσή του κατά καιρούς.



Σχήμα 3–15: Αυτόματη επιλογή διευκρίνηση έννοιας όρου (trends.google.com)



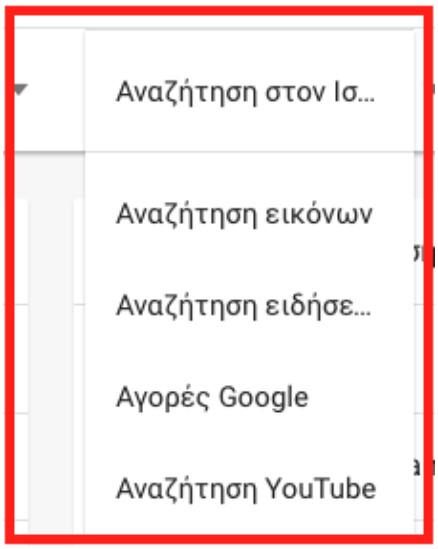
Σχήμα 3–16: Επιλογή κατηγορίας έννοιας όρου (trends.google.com)



Σχήμα 3–17: Κατηγορίες και υποκατηγορίες εννοιών (trends.google.com)

Στο πεδίο πλατφόρμα αναζητήσεων μπορεί να επιλεχθεί σε ποια πλατφόρμα έγινε αναζήτηση και τα δεδομένα προέρχονται από διάφορες πηγές όπως η Αναζήτηση Ιστού

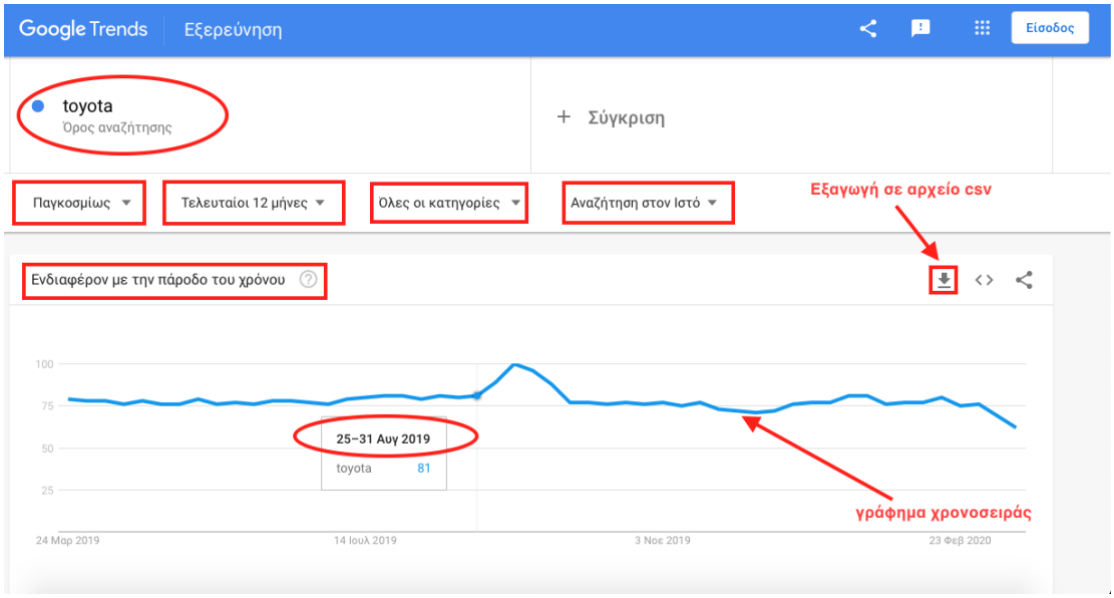
της Google, η Αναζήτηση Εικόνων, η Αναζήτηση Ειδήσεων, Αγορές Google, και η ιστοσελίδα YouTube (**Σχήμα 3–18**).



Σχήμα 3–18: Πλατφόρμα αναζήτησης (trends.google.com)

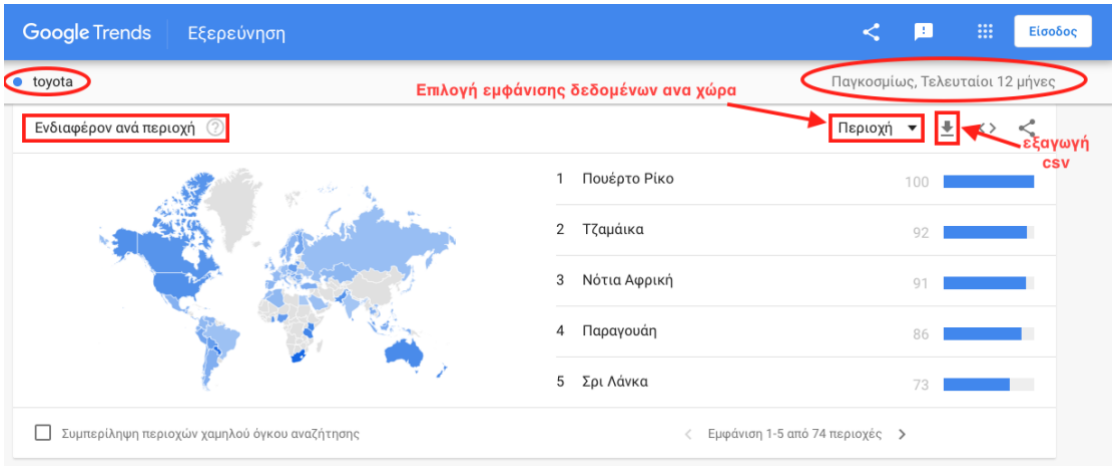
Μόλις γίνει λοιπόν καταχώρηση κάποιου όρου αναζήτησης και επιλογή των κατάλληλων παραμέτρων όπως αυτές αναφέρθηκαν προηγουμένως τότε στην ενότητα της Εξερεύνησης παρέχονται τα ακόλουθα δεδομένα:

Η χρονοσειρά που αντικατοπτρίζει το ενδιαφέρον για έναν όρο με την πάροδο του χρόνου όπως περιγράφηκε στις προηγούμενες σελίδες. Αυτή η χρονοσειρά παρουσιάζεται στην ιστοσελίδα ως γράφημα αλλά μπορεί να γίνει και εξαγωγή των δεδομένων της ως αρχείο csv με τη χρήση σχετικού κουμπιού στο παράθυρο (**Σχήμα 3–19**). Τα χρονικά σημεία ποικίλουν από μήνες, εβδομάδες, μέρες, ώρες κοκ. ανάλογα με την επιλογή του χρονικού εύρους στις παραμέτρους. Γίνεται λοιπόν καταχώρηση του όρου «toyota», σε παγκόσμια κλίμακα, για τους τελευταίους 12 μήνες, σε όλες τις κατηγορίες και με προέλευση δεδομένων την Αναζήτηση Ιστού Google. Εξάγεται το ακόλουθο γράφημα (**Σχήμα 3–19**) με χρονικά σημεία της χρονοσειράς ανά εβδομάδα. Στο σχήμα επίσης φαίνεται η επιλογή παραμέτρων και το κουμπί εξαγωγή των δεδομένων της χρονοσειράς σε αρχείο μορφής csv.

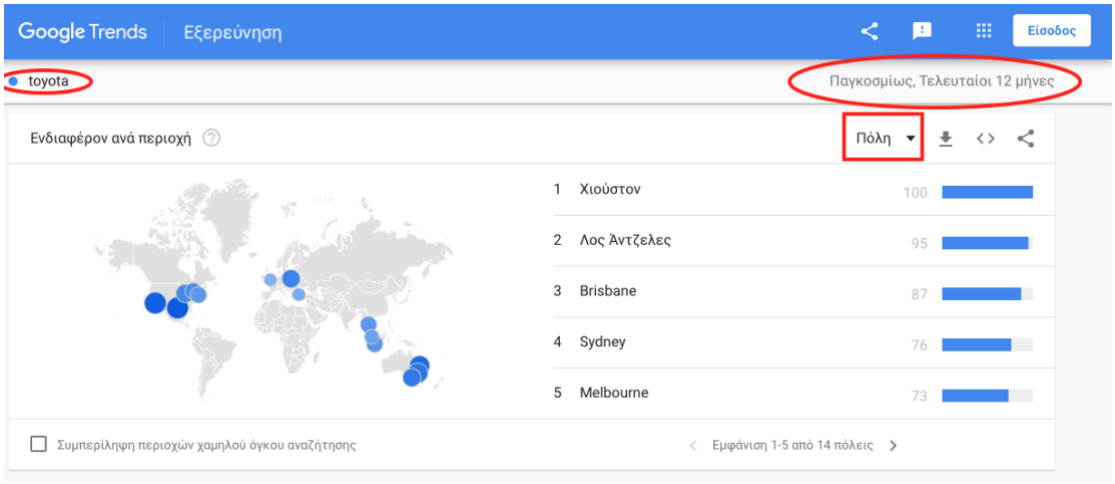


Σχήμα 3–19: Γράφημα χρονοσειράς (trends.google.com)

Τα δεδομένα ενδιαφέροντος ανά περιοχή τα οποία παρουσιάζονται ως ένας χάρτης με εντονότερα σκιασμένες τις υποπεριοχές με μεγαλύτερο ενδιαφέρον αλλά και ως λίστα με τις υποπεριοχές και την αντίστοιχη βαθμολογία καθεμιάς. Επιπλέον και εδώ είναι δυνατόν τα δεδομένα να εξαχθούν σε αρχείο csv με τη χρήση σχετικού κουμπιού στο παράθυρο. Οι υποπεριοχές μπορούν να φθάσουν μέχρι επίπεδο πόλης ανάλογα με τον ορισμό των γεωγραφικών ορίων στις παραμέτρους αναζήτησης. Οι υποπεριοχές με ανεπαρκή αριθμό αναζητήσεων θα έχουν βαθμολογία με τιμή 0 και δε θα αναφέρονται στη λίστα. Για την ίδια με την παραπάνω καταχώρηση λαμβάνονται τα εξής δεδομένα (**Σχήμα 3–20**) τα οποία παρουσιάζονται σε επίπεδο χώρας φαίνονται οι περιοχές με το μεγαλύτερο ενδιαφέρον οι οποίες είναι και πιο έντονα σκιασμένες στο χάρτη. Τώρα αν γίνει επιλογή παρουσίασης δεδομένων στο επίπεδο πόλης εξάγονται τα αντίστοιχα δεδομένα (**Σχήμα 3–21**) όπου παρουσιάζονται οι πόλεις με την μεγαλύτερη δημοτικότητα. Εδώ παρατηρείται επίσης η έννοια του σχετικού ενδιαφέροντος μιας και όπως είναι εμφανές στην ανά χώρα παρουσίαση τη μέγιστη τιμή δημοτικότητας κατέχει το Πουέρτο Ρίκο ενώ σε επίπεδο πόλης τη μέγιστη τιμή την κατέχει το Χιούστον των ΗΠΑ.

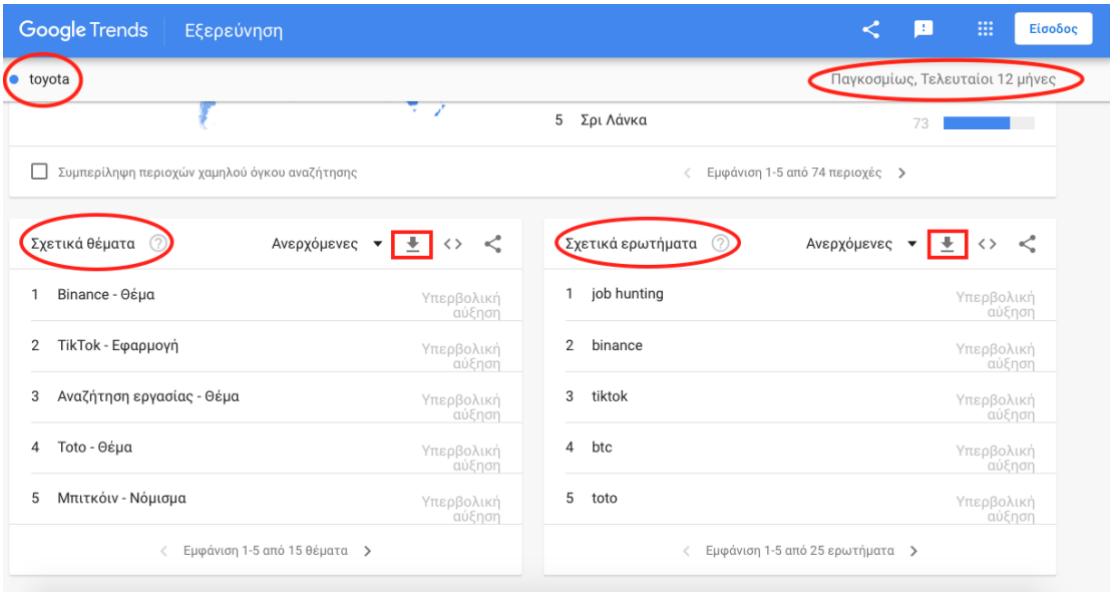


Σχήμα 3–20: Δεδομένα ενδιαφέροντος ανά χώρα (trends.google.com)



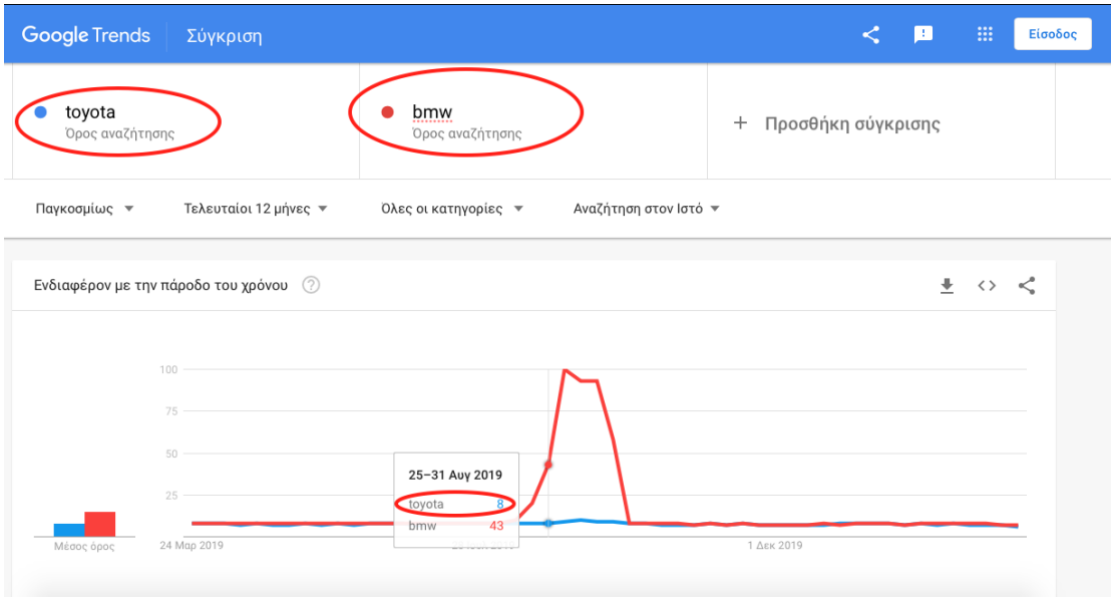
Σχήμα 3–21: Δεδομένα ενδιαφέροντος ανά πόλη (trends.google.com)

Επίσης στη σελίδα της Εξερεύνησης εμφανίζονται λίστες που περιλαμβάνουν σχετικά με τον αναζητούμενο όρο θέματα και ερωτήματα ανερχόμενα και κορυφαία έτσι όπως περιγράφηκαν σε προηγούμενη παράγραφο. Ακόμα και αυτές οι λίστες μπορούν να εξαχθούν σε αρχείο μορφής csv με χρήση κατάλληλου κουμπιού (**Σχήμα 3–22**).

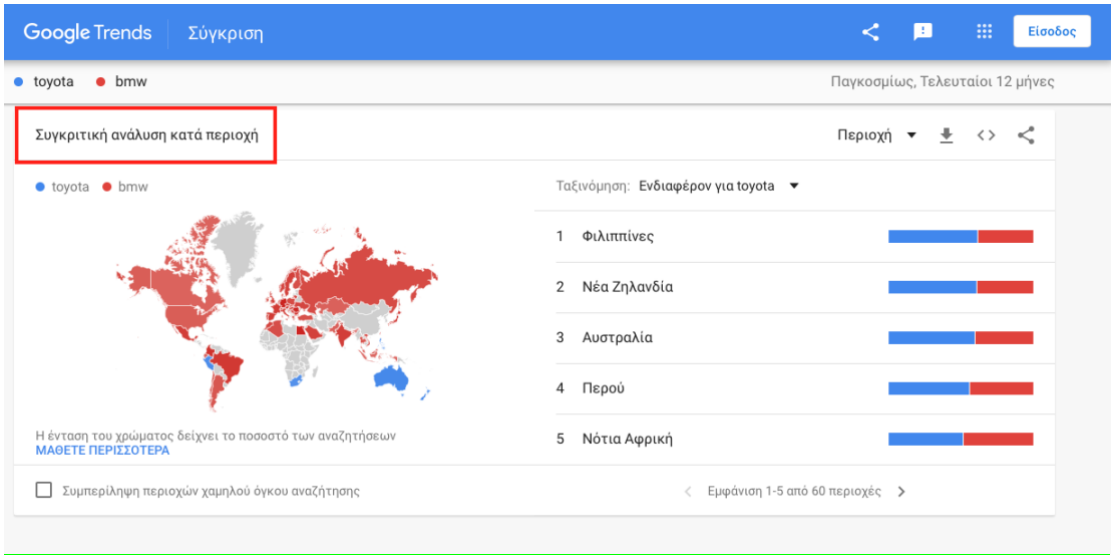


Σχήμα 3–22: Σχετικά θέματα κι ερωτήματα αναζητούμενου όρου (trends.google.com)

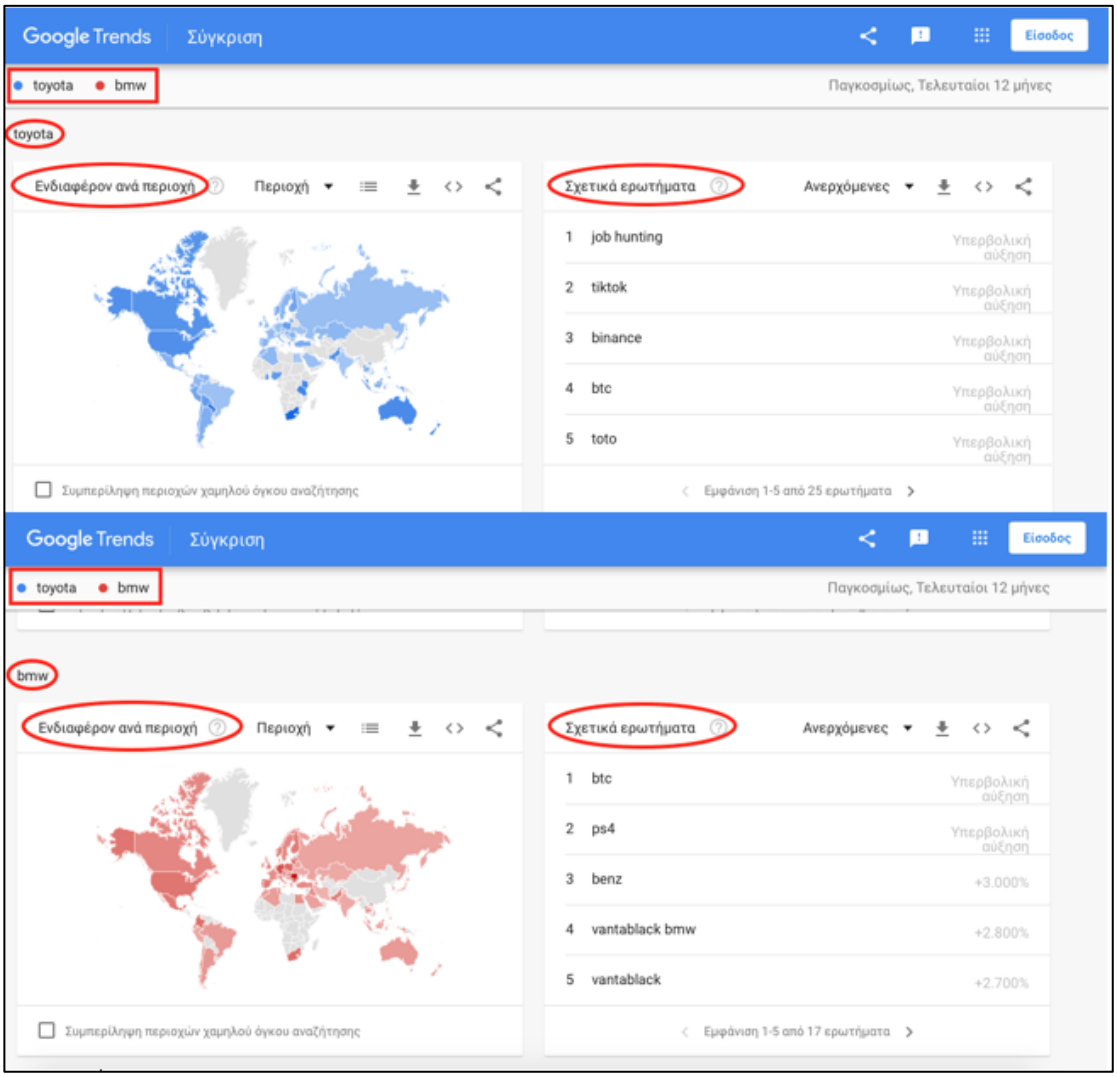
Επιπροσθέτως αν καταχωρηθούν πρόσθετοι όροι για σύγκριση μεταξύ τους τα δεδομένα που θα παραχθούν θα είναι πάλι χρονοσειρές και χάρτες που παρουσιάζουν τη δημοτικότητα των όρων (**Σχήμα 3–23**) και χάρτες συγκριτικής ανάλυσης (**Σχήμα 3–24**) σχετικά ερωτήματα κ.λπ. (**Σχήμα 3–25**). Απλά σε αυτήν την περίπτωση υπάρχει η διαφορά ότι οι βαθμολογίες που προκύπτουν για κάθε όρο, αντικατοπτρίζουν τη σχετική μεταξύ τους δημοτικότητα από το δείγμα που λήφθηκε. Για παράδειγμα αν καταχωρηθούν οι όροι «toyota» και «bmw» προς σύγκριση, στο κοινό διάγραμμα των χρονοσειρών φαίνεται ότι στο χρονικό σημείο 25-31 Αυγ 2019 ο όρος «toyota» έχει βαθμολογία ίση με 8 (**Σχήμα 3–23**) ενώ στο ίδιο χρονικό σημείο της χρονοσειράς όταν ζητούνται αποτελέσματα για τον όρο μόνο του φαίνεται ότι έχει βαθμολογία ίση με 81 (**Σχήμα 3–19**). Συνεπώς γίνεται ξεκάθαρη η διαφορά του σχετικού ενδιαφέροντος κατά τη σύγκριση δύο όρων. Επιπλέον για κάθε έναν από τους συγκρινόμενους όρους μπορεί να γίνει κάποια αλλαγή των παραμέτρων κι έτσι να συγκριθούν τοποθεσίες και χρονικές περίοδοι μεταξύ τους. Για παράδειγμα γίνεται καταχώρηση των όρων «χάμπουργκερ» και «χορτοφαγία» προς σύγκριση με τη διαφορά ότι στον πρώτο όρο έχει ορισθεί περιοχή αναζήτησης οι Ηνωμένες Πολιτείες ενώ στο δεύτερο η Γερμανία (**Σχήμα 3–26**). Ανάλογα μπορεί να αλλάξουν και οι χρονικές περίοδοι για κάθε όρο. Η αλλαγή παραμέτρων για κάθε όρο γίνεται από το κουμπί Αλλαγή φίλτρων στο πεδίο εισαγωγής όρου (**Σχήμα 3–27**).



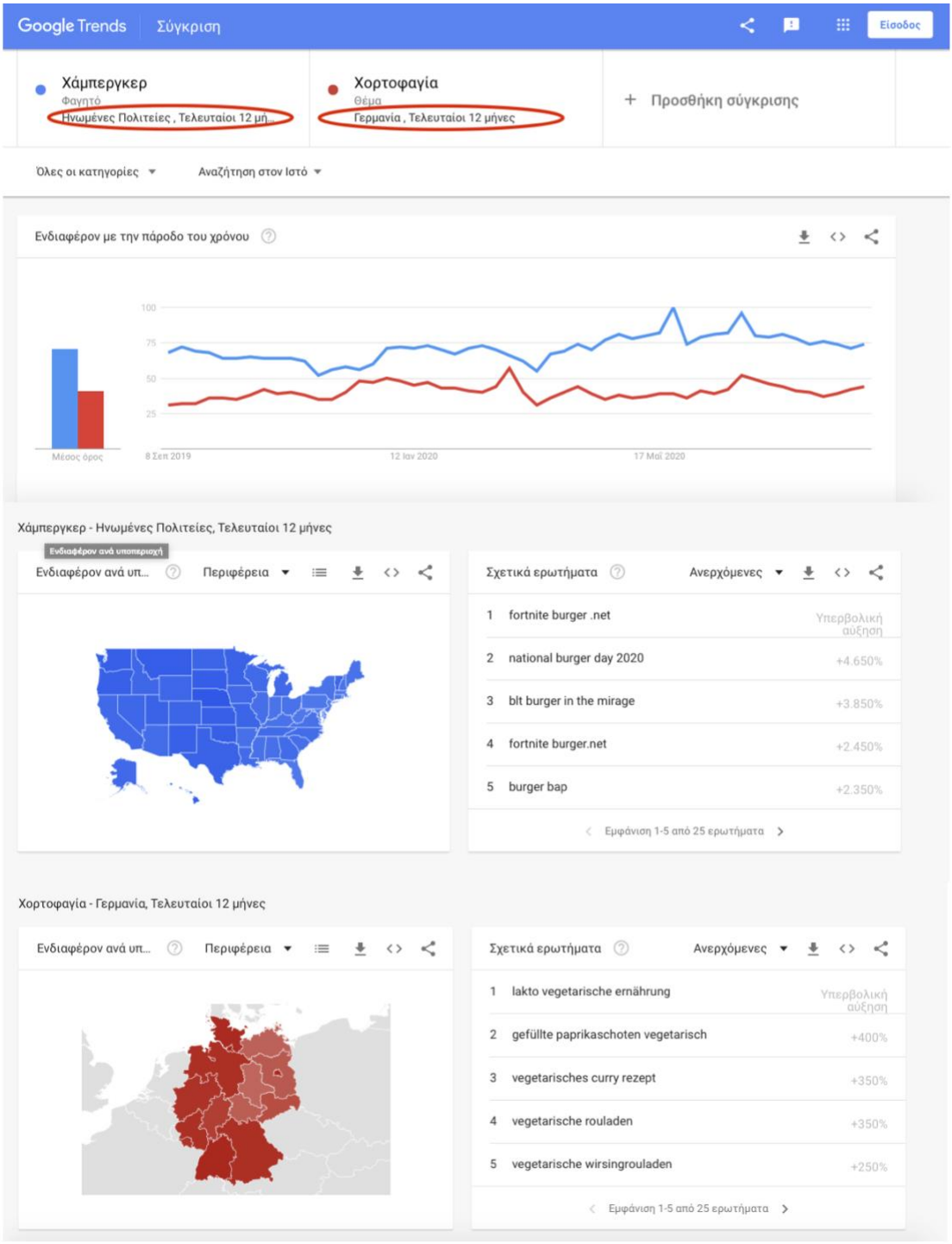
Σχήμα 3–23: Αποτελέσματα σύγκρισης δύο όρων στο χρόνο (trends.google.com)



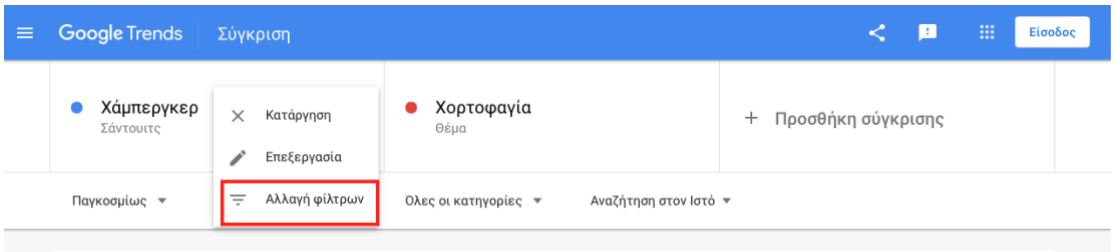
Σχήμα 3–24: Αποτελέσματα σύγκρισης δύο όρων κατά περιοχή (trends.google.com)



Σχήμα 3–25: Ενδιαφέρον ανά περιοχή και σχετικά ερωτήματα συγκρινόμενων όρων (trends.google.com)



Σχήμα 3–26: Σύγκριση όρων σε διαφορετικές περιοχές (trends.google.com)



Σχήμα 3–27: Αλλαγή φίλτρων σύγκρισης όρων (trends.google.com)

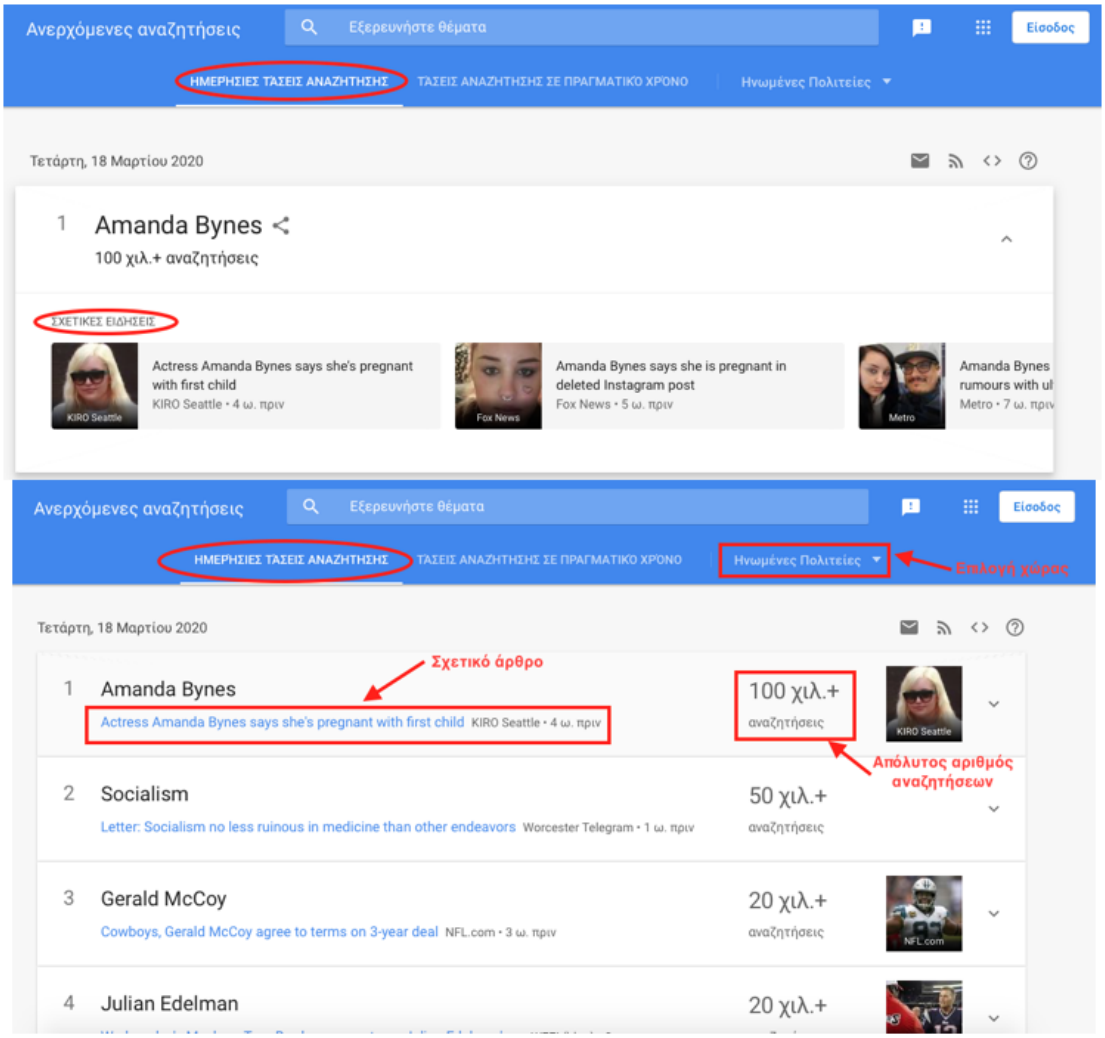
Τέλος ακόμα ένα φιλτράρισμα που μπορεί να γίνει στα αποτελέσματα του Google Trends είναι με τη χρήση κατάλληλης στίξης. Διευκρινιστικά αν παραδείγματος χάριν εισαχθεί ο όρος παπούτσια τένις χωρίς σημεία στίξης τότε τα αποτελέσματα θα περιλαμβάνουν αναζητήσεις που περιέχουν τους όρους «τένις» και «παπούτσια» με οποιαδήποτε σειρά και να περιλαμβάνουν ακόμα και αναζητήσεις «τένις χωρίς παπούτσια» κ.λπ. Αν πληκτρολογηθεί ο όρος ¨παπούτσια τένις¨ με διπλά εισαγωγικά τότε οι αναζητήσεις θα περιλαμβάνουν την ακριβή φράση μέσα στα διπλά εισαγωγικά, πιθανότατα με λέξεις πριν από και μετά τη φράση («κόκκινα παπούτσια τένις»). Αν πληκτρολογηθεί «τένις» + «σκουός» τότε τα αποτελέσματα θα περιλαμβάνουν αναζητήσεις που περιέχουν τις λέξεις «τένις» ή «σκουός». Αν πάλι πληκτρολογηθεί «παπούτσια» - «τένις» τα αποτελέσματα θα περιλαμβάνουν αναζητήσεις που περιέχουν τη λέξη «τένις», αλλά θα αποκλείουν τις αναζητήσεις με τη λέξη «παπούτσια». Αν πληκτρολογηθεί «καταχώριση» + «καταχώρηση» + «καταχώρειση», τα αποτελέσματα θα περιλαμβάνουν τις διαφορετικές ορθογραφίες δηλαδή το Google Trends αντιμετωπίζει κάθε εκδοχή μιας λέξης ως διαφορετική αναζήτηση, συμπεριλαμβανόμενων των ανορθόγραφων εκδοχών (Google , 2020).

3.3.3 Δημοφιλείς Αναζητήσεις

Στην ενότητα αυτή της ιστοσελίδας εμφανίζονται δημοφιλείς αναζητήσεις από όλον το κόσμο. Για κάθε μία από τις αναζητήσεις παρέχονται πρόσθετες πληροφορίες, όπως τα πιο συναφή άρθρα ή τα δημοφιλή ερωτήματα. Η σελίδα αυτή περιλαμβάνει τάσεις ημερήσιων αναζητήσεων και τάσεις αναζητήσεων σε πραγματικό χρόνο, όπως περιγράφονται παρακάτω:

Ημερήσιες τάσεις αναζήτησης

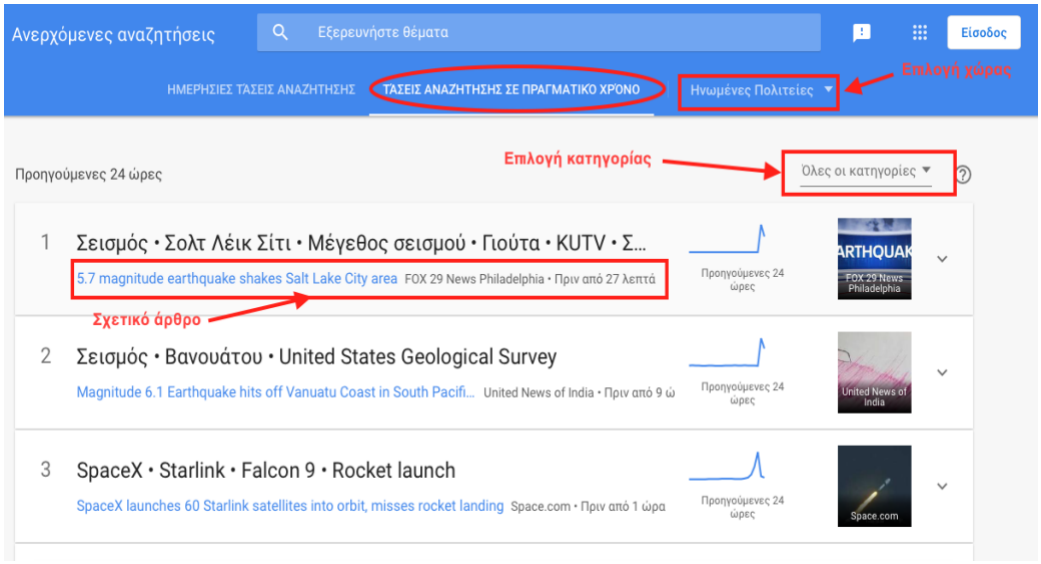
Οι Ημερήσιες τάσεις αναζήτησης επισημαίνουν αναζητήσεις οι οποίες παρουσίασαν σημαντικά αυξημένη επισκεψιμότητα μεταξύ όλων των αναζητήσεων κατά τις τελευταίες 24 ώρες και ενημερώνονται ανά μία ώρα. Αυτές οι τάσεις αναζήτησης δείχνουν τον τρόπο με τον οποίο αναζητούνται συγκεκριμένα ερωτήματα και τον απόλυτο αριθμό των αναζητήσεων που πραγματοποιούνται. Επίσης μπορεί να γίνει επιλογή της χώρας για την οποία εμφανίζονται πληροφορίες για τις ημερήσιες τάσεις (**Σχήμα 3–28**) .



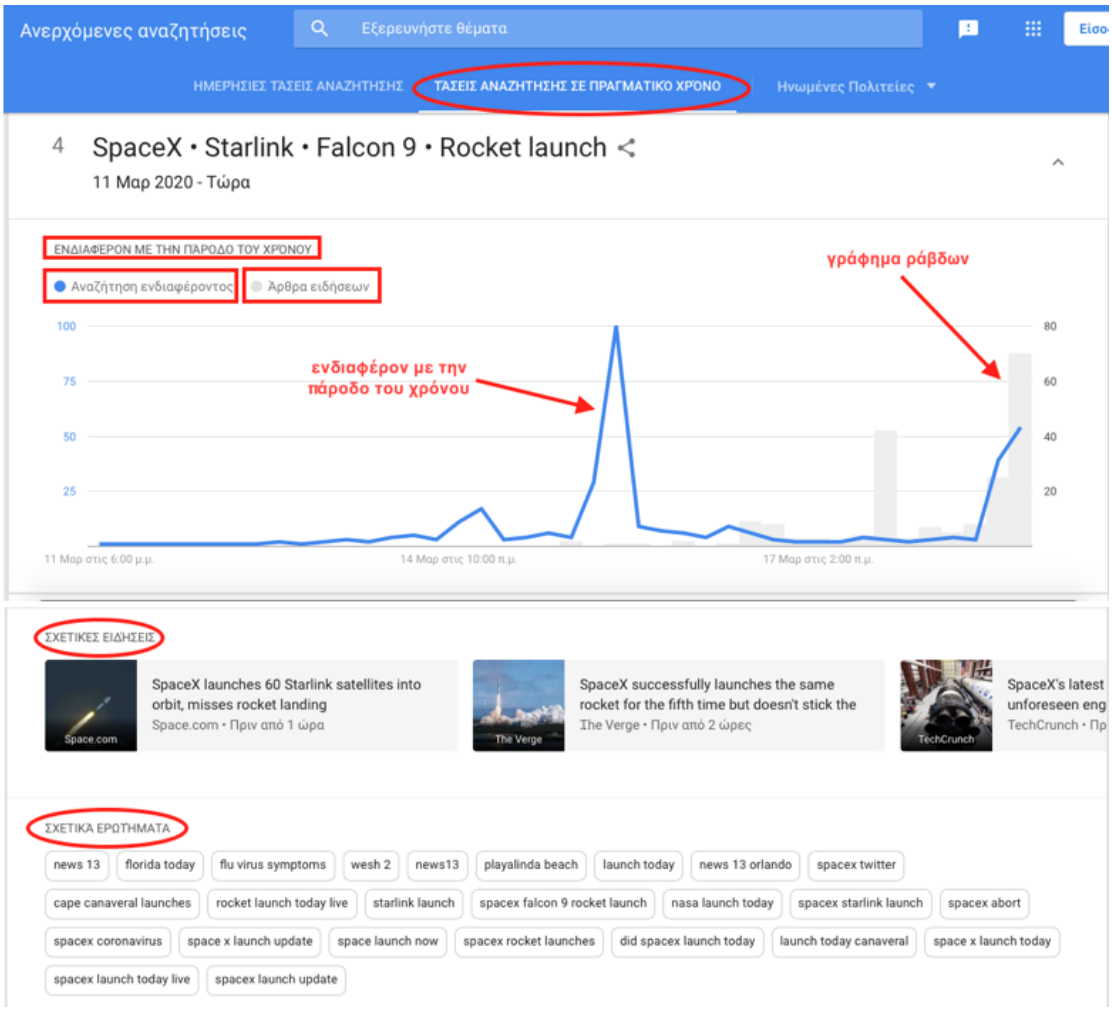
Σχήμα 3–28: Ημερήσιες τάσεις αναζήτησης (trends.google.com)

Τάσεις αναζήτησης σε πραγματικό χρόνο

Οι Τάσεις αναζήτησης σε πραγματικό χρόνο επισημαίνουν ιστορίες που είναι ανερχόμενες σε προϊόντα και υπηρεσίες Google κατά τις τελευταίες 24 ώρες και ενημερώνονται σε πραγματικό χρόνο. Αυτές οι ιστορίες αποτελούν μια συλλογή από θέματα του Γραφήματος Γνώσεων, το Ενδιαφέρον αναζήτησης, ανερχόμενα βίντεο YouTube ή/και άρθρα των Ειδήσεων Google που έχουν εντοπιστεί από τους αλγόριθμούς της Google (**Σχήμα 3–29**).



Σχήμα 3–29: Τάσεις αναζήτησης σε πραγματικό χρόνο (trends.google.com)



Σχήμα 3–30: Ανάπτυξη ιστορίας (trends.google.com)

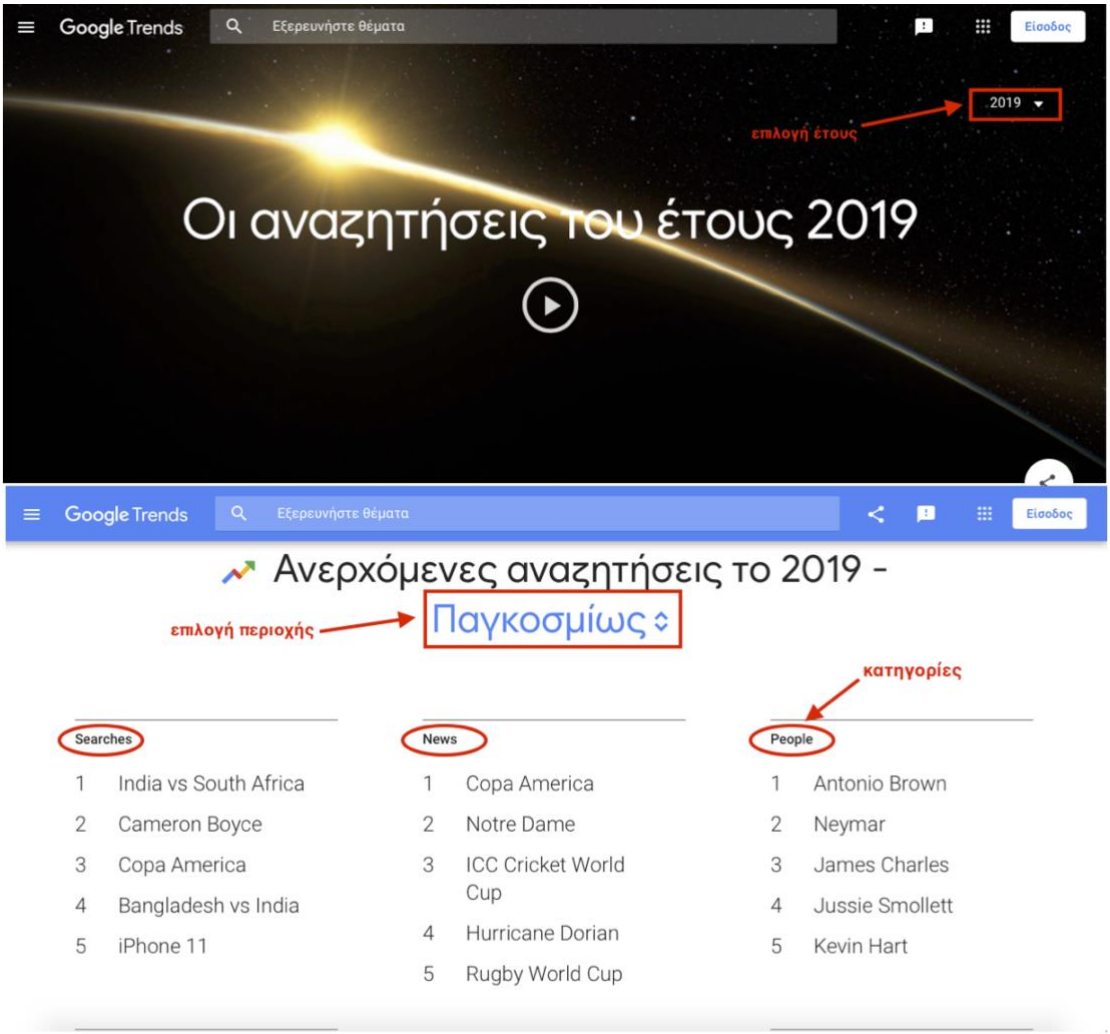
Στο **Σχήμα 3–30** το γράφημα ράβδων δείχνει τον αριθμό των άρθρων των Ειδήσεων Google που γράφονται ανά ώρα και αντιστοιχεί στον γκρι άξονα στα δεξιά. Το γράφημα γραμμής δείχνει το ενδιαφέρον στην Αναζήτηση Google με την πάροδο του χρόνου και

αντιστοιχεί στον μπλε άξονα στα αριστερά. Το γράφημα "Ενδιαφέρον με την πάροδο του χρόνου" δείχνει το ενδιαφέρον για τις αναζητήσεις σε σχέση με το υψηλότερο σημείο του γραφήματος, αλλά δεν εκφράζει τον απόλυτο όγκο των αναζητήσεων.

Το Γράφημα γνώσεων επιτρέπει στην τεχνολογία της Google τη σύνδεση αναζητήσεων με πράγματα και μέρη του πραγματικού κόσμου. Ο αλγόριθμος για τις τάσεις αναζητήσεων σε πραγματικό χρόνο ομαδοποιεί θέματα που είναι ταυτόχρονα ανερχόμενα στις Ειδήσεις και την Αναζήτηση Google και κατατάσσει τις ιστορίες με βάση τη σχετική κορύφωση σε όγκο και τον απόλυτο όγκο των αναζητήσεων. Η επιλογή των σχετικών άρθρων ειδήσεων που εμφανίζεται κάτω από τις αναζητήσεις γίνεται παρακολουθώντας την πλήρη κάλυψη των ιστοριών στις Ειδήσεις Google. Εάν διαπιστωθεί ότι οι ιστορίες αφορούν κυρίως θέματα τα οποία είναι ανερχόμενα τη δεδομένη στιγμή, επισημαίνονται τα κύρια άρθρα στην κάλυψη. Λαμβάνονται δεδομένα από την πλήρη κάλυψη των ιστοριών στις Ειδήσεις Google που σχετίζονται με ένα συμβάν. Έπειτα, χρησιμοποιείται η κατάταξη των Ειδήσεων Google, για να επιλεχθούν τα κορυφαία άρθρα για τη συγκεκριμένη δημοφιλή ιστορία.

3.3.4 Οι αναζητήσεις τους έτους

Αυτή η ενότητα της ιστοσελίδας δείχνει λίστες με όρους αναζήτησης που παρουσίασαν τη μεγαλύτερη κορύφωση κάθε έτος σε σχέση με το προηγούμενο. Αυτά τα δεδομένα μπορούν να εξαχθούν για χρονιές από το 2001 ως σήμερα ανάλογα με την επιλογή περιοχής στην οποία διεξήχθησαν οι αναζητήσεις. Οι λίστες αυτές παρουσιάζονται ανά κατηγορία όπως ειδήσεις, άνθρωποι, αναζητήσεις, ταινίες, ηθοποιοί κλπ. Κάθε λίστα αποτελείται από τις 10 κορυφαίες αναζητήσεις δηλαδή αυτές για τις οποίες παρατηρήθηκε η μεγαλύτερη δημοτικότητα σχετικά με προηγούμενες χρονικές περιόδους (**Σχήμα 3–31**).



Σχήμα 3–31: Αναζητήσεις του έτους (trends.google.com)

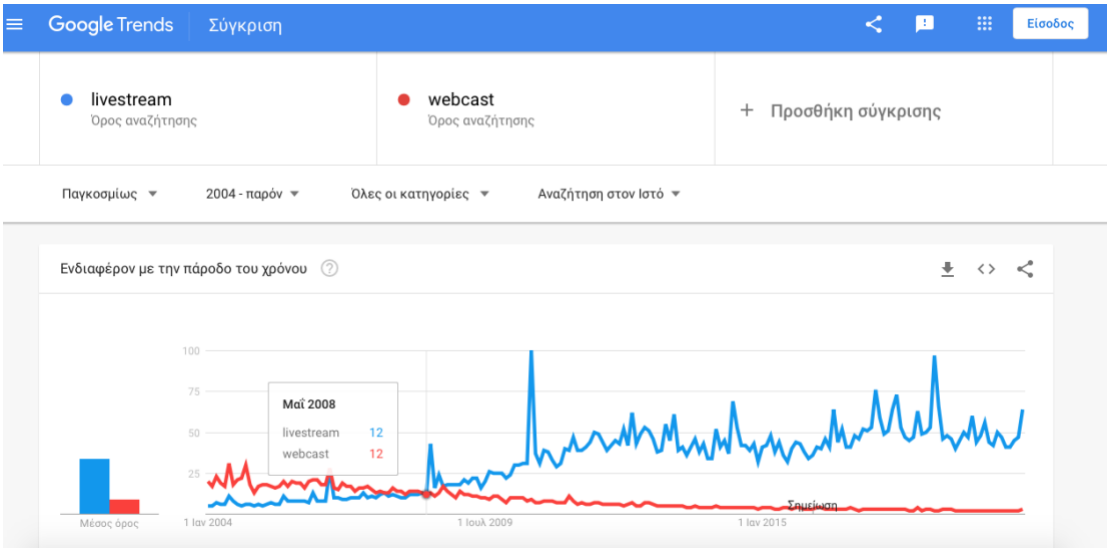
3.4 Εφαρμογές της πλατφόρμας Google Trends

3.4.1 Χρήσεις των δεδομένων της ιστοσελίδας

Οι αναζητήσεις στο διαδίκτυο αποτελούν μια τεράστια δεξαμενή δεδομένων. Η γνώση για το τι αναζητεί ο κόσμος στο διαδίκτυο παρέχει μια μοναδική προοπτική για ποια πράγματα δείχνει ενδιαφέρον και για ποια είναι περίεργος (Rogers, 2016). Οι μηχανές αναζήτησης έχουν γίνει συνήθη μέσα για την απόκτηση πολλών τύπων γνώσης, από την απάντηση σε συνηθισμένες ερωτήσεις ως την λεπτομερή περιγραφή ή την παροχή γενικών πληροφοριών κάθε θέματος (Broucke, 2016). Οι χρήστες κάνουν αναζητήσεις αυθόρμητα ενώ κάποιο μέρος των αναζητήσεών τους αφορά την αγορά προϊόντων, καθιστώντας τις αναζητήσεις στο διαδίκτυο κομμάτι της διαδικασίας για απόκτηση καταναλωτικών αγαθών. Συνεπώς δεδομένα σχετικά με τις αναζητήσεις μπορεί να φανούν χρήσιμα σε τομείς όπως το εμπόριο.

Το Google Trends το οποίο προσφέρει τέτοιου είδους δεδομένα μπορεί να χρησιμοποιηθεί για διάφορους σκοπούς. Παρακάτω παρατίθενται μερικοί:

- Βελτιστοποίηση ιστοσελίδων για τις μηχανές αναζήτησης (Search Engine Optimization). Πιο αναλυτικά γίνεται εύρεση περιεχομένου (λέξεις κλειδιά κ.λπ.) για το οποίο υπάρχει αυξημένο ενδιαφέρον στο Google Trends αυτό προστίθεται στις ιστοσελίδες ώστε αυτές να είναι καλύτερη η κατάταξή τους από τον αλγόριθμο των μηχανών αναζήτησης.
- Καθορισμός προϊόντων με μεγάλο ενδιαφέρον για ενασχόληση με το εμπόριο.
- Εύρεση σχετικών προϊόντων με αυτά στα οποία υπάρχει ενδιαφέρον για μελλοντική επέκταση επιχειρήσεων.
- Παρακολούθηση τοπικού ενδιαφέροντος για βελτίωση δικτύου διανομής. Παραδείγματος χάριν εξασφάλιση επάρκειας αποθέματος στις αποθήκες που βρίσκονται σε περιοχές με αυξημένη δημοτικότητα στα Google Trends.
- Παρακολούθηση της απόδοσης της στρατηγικής μάρκετινγκ (marketing) της επιχείρησης.
- Παρακολούθηση της θέσης ανταγωνιστών με χρήση της σύγκρισης όρων του Google Trends.
- Πιο αποτελεσματική επικοινωνία, όπως για ορολογίες με παρόμοια σημασία, μπορεί να γίνει έρευνα στο Google Trends για το ποιος όρος είναι πιο γνωστός ώστε να χρησιμοποιηθεί σε ένα τίτλο κειμένου κάποιο άρθρου κλπ. (Wolber, 2017). Για παράδειγμα οι όροι «livestream» και «webcast» εκφράζουν παρεμφερές νόημα, αλλά σήμερα όπως φαίνεται ο κόσμος αναζήτα περισσότερο για τον όρο «livestream» ενώ μέχρι το 2008 ο άλλος όρος ήταν πιο δημοφιλής (**Σχήμα 3–32**), συνεπώς αν χρησιμοποιηθεί ο πιο δημοφιλής όρος θα γίνει και περισσότερο αντιληπτός από το κοινό αλλά και θα βρεθεί και πιο εύκολα η ιστοσελίδα που το φιλοξενεί από τις μηχανές αναζήτησης.



Σχήμα 3–32: Σύγκριση όρων με παρεμφερείς σημασίες (trends.google.com)

3.4.2 Google Trends από ερευνητική σκοπιά

Καθώς οι αναζητήσεις στο διαδίκτυο και με τη σειρά του Google Trends αποτελούν μια μεγάλη και χρήσιμη πηγή πληροφοριών της οποίας τα δεδομένα είναι δυνατόν να χρησιμεύσουν σε διάφορους τομείς, έχει παρατηρηθεί μεγάλο ερευνητικό ενδιαφέρον όσον αφορά το εργαλείο του Google Trends. Η ερευνητική δραστηριότητα ουσιαστικά ξεκίνησε στον τομέα της επιδημιολογίας από τους Ginsberg, et al. (Ginsberg, et al., 2009) οι οποίοι χρησιμοποίησαν το Google Trends για να ιχνηλατήσουν την εβδομαδιαία έξαρση της εποχικής γρίπης στον πληθυσμό των Η.Π.Α. Ανακάλυψαν πως υπάρχει υψηλή συσχέτιση ανάμεσα στη σχετική δημοτικότητα ορισμένων όρων αναζήτησης και σε ποσοστά επισκέψεων γιατρών σε ασθενείς που παρουσιάζουν συμπτώματα ανάλογα με αυτά του ιού της γρίπης. Το μοντέλο που έφτιαξαν μετατράπηκε σε ένα σύστημα παρακολούθησης του ιού σε πραγματικό χρόνο το οποίο μπορεί να προβλέψει τη δραστηριότητα της γρίπης μια ή δύο εβδομάδες πριν την δημοσίευση αναφορών από τα Κέντρα Ελέγχου και Πρόληψης Ασθενειών της Αμερικής (CDC). Περαιτέρω δημοσιεύσεις επικεντρώνονται σε προβλέψεις και σε άλλους τομείς όπως αυτός της οικονομίας όπου οι Askitas και Zimmermann (Askitas & Zimmermann, 2009) χρησιμοποίησαν μηνιαία δεδομένα αναζητήσεων στη γερμανική γλώσσα και βρήκαν μια ισχυρή συσχέτιση ανάμεσα σε αναζητήσεις λέξεων κλειδιών και σε τιμές ανεργίας έτσι πρότειναν ένα τρόπο παρακολούθησης και πρόγνωσης των αλλαγών στις συνθήκες της οικονομίας.

Όπως ήδη έχει αναφερθεί σε αρχική ενότητα της παρούσας εργασίας, από τους πρώτους που εξέτασαν αν το εργαλείο του Google Trends έχει προγνωστική ισχύ ήταν οι Choi και Varian (Choi & Varian, 2012) οι οποίοι εισήγαγαν δεδομένα από την εφαρμογή σε παλινδρομικά μοντέλα προβλέψεων και κατάφεραν να επιτύχουν βελτίωση στην ακρίβεια της πρόγνωσης πωλήσεων σε κλάδους όπως το λιανικό εμπόριο, η αυτοκινητοβιομηχανία, η αγορά ακινήτων και ο τουρισμός. Σε αυτήν την εργασία βρέθηκε ότι η ενίσχυση των προγνωστικών μοντέλων με δεδομένα από το Google Trends είναι χρήσιμη περισσότερο για προβλέψεις κοντινών σε χρονικό ορίζοντα γεγονότων επειδή η εφαρμογή παρέχει πληροφορίες σχετικά με την αλλαγή στο ενδιαφέρον για κάποιο αναζητούμενο όρο, δηλαδή φανερώνει το σημείο καμπής στη δημοτικότητά του. Αν και η πρόγνωση πωλήσεων αποτελεί ένα σημαντικό πεδίο εφαρμογής των δεδομένων του Google Trends εκτενέστερη αναφορά στην ερευνητική δραστηριότητα σε αυτόν το τομέα και συγκεκριμένα στον κλάδο της αυτοκινητοβιομηχανίας γίνεται στο μέρος της βιβλιογραφικής ανασκόπησης της παρούσας εργασίας η οποία ασχολείται ειδικά με αυτό το θέμα.

Άλλοι τομείς για τους οποίους μελετήθηκε η εφαρμογή των δεδομένων Google Trends στην πρόγνωση είναι οι προβλέψεις της κίνησης των τιμών του χρηματιστηρίου. Μια τέτοια δημοσίευση είναι αυτή των Preis, Moat και Stanley (Preis, Moat, & Stanley, 2013) οι οποίοι έφτιαξαν μια στρατηγική εμπορίας (trading) του χρηματιστηριακού δείκτη Dow Jones Industrial Average (DIJA) η οποία προέβλεπε τις κινήσεις της τιμής του δείκτη βασιζόμενη σε δεδομένα του Google Trends που αφορούσαν αναζητήσεις λέξεων-κλειδιών της οικονομίας, δηλαδή όρους όπως «κρίση», «χρέος», «πληθωρισμός» κλπ. Ακόμα μελετήθηκε η χρήση του Google Trends και στην πρόγνωση κοινωνικοπολιτικών γεγονότων όπως το 2011 οι Lui, Metaxas, & Mustafaraj (Lui, Metaxas, & Mustafaraj,

2011) χρησιμοποίησαν δεδομένα αναζητήσεων για να αναλύσουν τις πιθανότητες νίκης υποψηφίων που συμμετείχαν στις εκλογές για την αμερικανική γερουσία το 2008 και το 2010. Σε αυτή τη μελέτη η ικανότητα του Google Trends στη πρόβλεψη των νικητών δεν ήταν σχετικά υψηλότερη σε σύγκριση με συμβατικές δημοσκοπήσεις της εφημερίδας New York Times κάτι που έδειξε και τα όρια της προγνωστικής ισχύος της εφαρμογής. Επίσης στο ίδιο πεδίο οι Mavragani και Tsagarakis (Mavragani & Tsagarakis, 2019) παρουσίασαν μια μεθοδολογία για πρόβλεψη αποτελεσμάτων δημοψηφισμάτων χρησιμοποιώντας το Google Trends. Η μέθοδος αυτή εφαρμόστηκε στο δημοψήφισμα της Σκωτίας το 2014, στο ελληνικό δημοψήφισμα του 2015 και σε άλλες τέσσερις περιπτώσεις. Η μέθοδος αυτή έδειξε καλές επιδόσεις και σε κάποιες περιπτώσεις ξεπέρασε και την ακρίβεια επίσημων δημοσκοπήσεων.

Στο τομέα του τουρισμού μελέτες έχουν γίνει όπως αυτή των Bangwayo-Skeete και Skeete (Bangwayo-Skeete & Skeete, 2014) που ανέλυσε κατά πόσο μπορεί η χρήση δεδομένων του Google Trends να βελτιώσει την πρόγνωση των αφίξεων τουριστών στην περιοχή της Καραϊβικής.

Όπως είναι φανερό υπάρχει μεγάλο εύρος εφαρμογών των δεδομένων του Google Trends σε διάφορους τομείς. Η συγκεκριμένη πλατφόρμα αποτελεί μια πλούσια πηγή πληροφοριών σχετικά με τη συμπεριφορά των χρηστών του διαδικτύου των οποίων ο αριθμός ολοένα και αυξάνεται. Συνεπώς στο μέλλον υπάρχει περιθώριο να δοκιμαστούν τα δεδομένα της ιστοσελίδας για πρόγνωση ή παρακολούθηση και σε νέους κλάδους χωρίς όμως αυτά να αποτελούν πάντα το πιο κρίσιμο κομμάτι στο σχεδιασμό συστημάτων ανάλυσης ή πρόβλεψης.

3.4.3 Μειονεκτήματα Google Trends

Μπορεί το Google Trends να παρουσιάστηκε παραπάνω ως ένα ισχυρό εργαλείο το οποίο αντικατοπτρίζει το ενδιαφέρον των χρηστών του διαδικτύου και τα δεδομένα που προκύπτουν από αυτό μπορούν ενισχύσουν διάφορα προβλεπτικά μοντέλα, αυτό όμως δε σημαίνει ότι δεν υπάρχουν και κάποια μειονεκτήματα σχετικά με τη χρήση του Google Trends τα οποία παρατίθενται σε αυτή την ενότητα. Αναλυτικότερα:

- Οι αλγόριθμοι με τους οποίους η Google συλλέγει τα δείγματα δεδομένων είναι άγνωστοι δηλαδή αποτελούν μαύρο κουτί (black box). Έχει αναφερθεί κιόλας από ερευνητές ότι η εταιρεία συχνά αλλάζει τους αλγορίθμους και κατ' επέκταση τον τρόπο που συλλέγονται τα δεδομένα κάνοντας τις προβλέψεις ασταθείς.
- Η δειγματοληψία γίνεται από την εταιρεία για εμπορικούς κι όχι επιστημονικούς σκοπούς με στόχο το κέρδος, άρα έτσι δε διαφυλάσσεται κάποιο υψηλό επίπεδο αξιοπιστίας τόσο των μεθόδων όσο και των δεδομένων.
- Λόγω της ομαλοποίησης ενός τόσο μεγάλου όγκου δεδομένων, με την κανονικοποίηση και την απόδοση βαθμολογίας κλπ. Υπάρχει κίνδυνος υπερπροσαρμογής (overfitting) των προβλεπτικών μοντέλων.
- Στην παρούσα εργασία που αφορά την πρόγνωση πωλήσεων δεν πρέπει να συσχετίζεται απόλυτα η αναζήτηση του χρήστη με την πρόθεση αγοράς του

συγκεκριμένου προϊόντος, για παράδειγμα μπορεί κάποιος χρήστης να κάνει αναζήτηση για διάφορα μοντέλα αυτοκινήτων και να καταλήξει στην αγορά ενός ή να μην αγοράσει καθόλου.

- Θα πρέπει επίσης να διευκρινισθεί πότε κάποιος χρήστης κάνει αναζήτηση για την αγορά ενός καινούργιου αυτοκινήτου κάτι το οποίο εξετάζει αυτή εργασία και πότε ενός μεταχειρισμένου. Συνεπώς υπάρχει ο κίνδυνος τα δεδομένα του Google Trends να προσδώσουν λανθασμένη πληροφορία στα μοντέλα.

4. Εφοδιαστική Αλυσίδα της Αυτοκινητοβιομηχανίας στην Ελλάδα

4.1 Ιστορία της αυτοκινητοβιομηχανίας στην Ελλάδα

Η Ελλάδα σήμερα έχει μια ισχνή βιομηχανική δραστηριότητα όσον αφορά την παραγωγή μηχανοκίνητων οχημάτων. Σύμφωνα με στοιχεία του 2018 (ACEA , 2018) μόλις το 0,6% των εργαζομένων του συνολικού βιομηχανικού κλάδου της χώρας απασχολείται άμεσα στην αυτοκινητοβιομηχανία. Η βιομηχανική αυτή δραστηριότητα περιορίζεται σε παραγωγή στρατιωτικών οχημάτων, λεωφορείων αστικών συγκοινωνιών, απορριμματοφόρων κλπ. από επιχειρήσεις όπως η Ελληνική Βιομηχανία Οχημάτων (ΕΛΒΟ) στην οποία εμπλέκεται το Ελληνικό δημόσιο. Η συγκεκριμένη παραγωγή βασίζεται κυρίως στην μετατροπή εισαγόμενων οχημάτων με σκοπό την ικανοποίηση αναγκών όπως αυτές του στρατού, του πυροσβεστικού σώματος κτλ.

Γενικά στην Ελλάδα είναι ανύπαρκτη η κατασκευή οχημάτων και ιδιαίτερα η παραγωγή επιβατικών αυτοκινήτων. Η κατάσταση αυτή δεν ήταν όμως πάντα έτσι, παλιότερα είχαν γίνει αξιόλογες προσπάθειες για δημιουργία επιχειρήσεων αυτοκινητοβιομηχανίας στο ελληνικό χώρο, κάποιες από αυτές δραστηριοποιήθηκαν με επιτυχία για κάποιες δεκαετίες, άλλες δεν κατάφεραν καν να αρχίσουν τη λειτουργία τους, ενώ μερικές υπήρξαν ιδιαίτερα καινοτόμες για την εποχή τους. Βέβαια η κατάληξη όλων αυτών των επιχειρήσεων ήταν ο τερματισμός των δραστηριοτήτων τους κυρίως λόγω κοινωνικοπολιτικών συγκυριών αυτόν τον καιρό, αφήνοντας έτσι την Ελλάδα με μηδαμινή σχεδόν δραστηριότητα στον τομέα της αυτοκινητοβιομηχανίας. Παρακάτω γίνεται μια συνοπτική αναφορά σε κάποιες από τις προσπάθειες για δημιουργία αυτοκινητοβιομηχανίας στην Ελλάδα.

Τη δεκαετία του 1970 ιδρύθηκε στη Θεσσαλονίκη η NAMCO (National Motor Company) όπου κατασκεύασε διάφορα μοντέλα αυτοκινήτων πολλά ως μετατροπές άλλων οχημάτων όπως Citroen κλπ. Το πιο γνωστό μοντέλο που κατασκεύασε είναι το Pony από το 1972 ως το 1980, βασισμένο σε μηχανικά μέρη του Citroen 2CV, με ονομασία Pony-Citroen (**Σχήμα 4–1**), παράχθηκαν πάνω από 18000 κομμάτια. Το μοντέλο αυτό επειδή ήταν προσιτό τόσο στο κόστος λειτουργίας όσο και στη συντήρησή του έγινε δημοφιλές στις Ελληνικές Ένοπλες Δυνάμεις. Ακολούθησαν η παραγωγή και άλλων εκδόσεων του Pony, καθώς και απόπειρα κατασκευής νέου εργοστασίου στο Κιλκίς. Η επένδυση όμως δεν ολοκληρώθηκε εξαιτίας κοινωνικοπολιτικών λόγων και οι μετέπειτα εκδόσεις του Pony πούλησαν ελάχιστα χωρίς να έχουν τη στήριξη του κράτους. Η εταιρεία σήμερα ακόμα παραμένει ζωντανή καλύπτοντας ανάγκες σε αναπτυσσόμενες αγορές της Αφρικής, της Ασίας και της Λατινικής Αμερικής όπου υπάρχει ζήτηση για στιβαρά αλλά προσιτά οικονομικά αυτοκίνητα.

Σχήμα 4–1: Pony-Citroen (el.wikipedia.org)

Η MAVA ήταν ο ελληνικός εισαγωγέας των αυτοκινήτων Renault. Το 1979 αποφάσισε να εισέλθει στο χώρο παραγωγής αυτοκινήτων, παρουσιάζοντας ένα επιβατηγό-βοηθητικό αυτοκίνητο, ένα είδος δημοφιλές στην Ελλάδα για λόγους φορολογικής κατηγοριοποίησης. Το μοντέλο αυτό ονομάστηκε Farma (**Σχήμα 4–2**) κι έφερε το λογότυπο της Renault, ενώ παράχθηκε σε διάφορες εκδόσεις, συμπεριλαμβανομένων των εκδόσεων «επιβατηγό» και «βαν». Τη δημιουργία του αυτοκινήτου ανέλαβε ο Έλληνας σχεδιαστής Γεώργιος Μιχαήλ, κατασκευάστηκαν 4500 κομμάτια διάφορων εκδόσεων. Από το 1985 η ελληνική νομοθεσία άλλαξε κι επηρέασε την αγορά για αυτό το είδος των οχημάτων, καθιστώντας έτσι την παραγωγή τους ασύμφορη. Έπειτα το 1985 σχεδιάστηκε ένα νέο μοντέλο πολύ πιο προηγμένο κι ελκυστικό το Farma Change, τότε όμως η MAVA σταμάτησε το έργο και μόνο ένα αυτοκίνητο, το πρωτότυπο του νέου μοντέλου, κατασκευάστηκε.

Σχήμα 4–2: MAVA-Renault Farma F (el.wikipedia.org)

Τη δεκαετία του 1970 έγινε κι ένα άλλο αξιοσημείωτο εγχείρημα για τη δημιουργία αυτοκινητοβιομηχανίας στην Ελλάδα. Πρόκειται για την κατασκευή του ηλεκτρικού αυτοκινήτου Enfield-Neorion 8000 (**Σχήμα 4–3**) στο χώρο των ναυπηγείων στο Νεώριο της Σύρου. Το καινοτόμο αυτό όχημα αναπτύχθηκε από της βρετανική εταιρεία Enfield Automotive η οποία αυτήν την εποχή άνηκε στον Έλληνα εφοπλιστή Γιάννη Γουλανδρή και καθώς η κατασκευή του οχήματος στην Αγγλία αποδείχθηκε πολύ ακριβή, αποφασίστηκε να γίνει στα ναυπηγεία του Νεωρίου τα οποία και αυτά ανήκαν στον εφοπλιστή εξ ου και το όνομα Enfield-Neorion. Η παραγωγή άρχισε στα τέλη του 1973

και διήρκεσε μέχρι τις αρχές του 1976 όποτε και διακόπηκε. Κατά τη διάρκεια αυτής της περιόδου κατασκευάστηκαν 120 αυτοκίνητα τα περισσότερα από τα οποία πουλήθηκαν στο Ηνωμένο Βασίλειο.

Σχήμα 4–3: Enfield Neorion E 8000 Bicini (el.wikipedia.org)

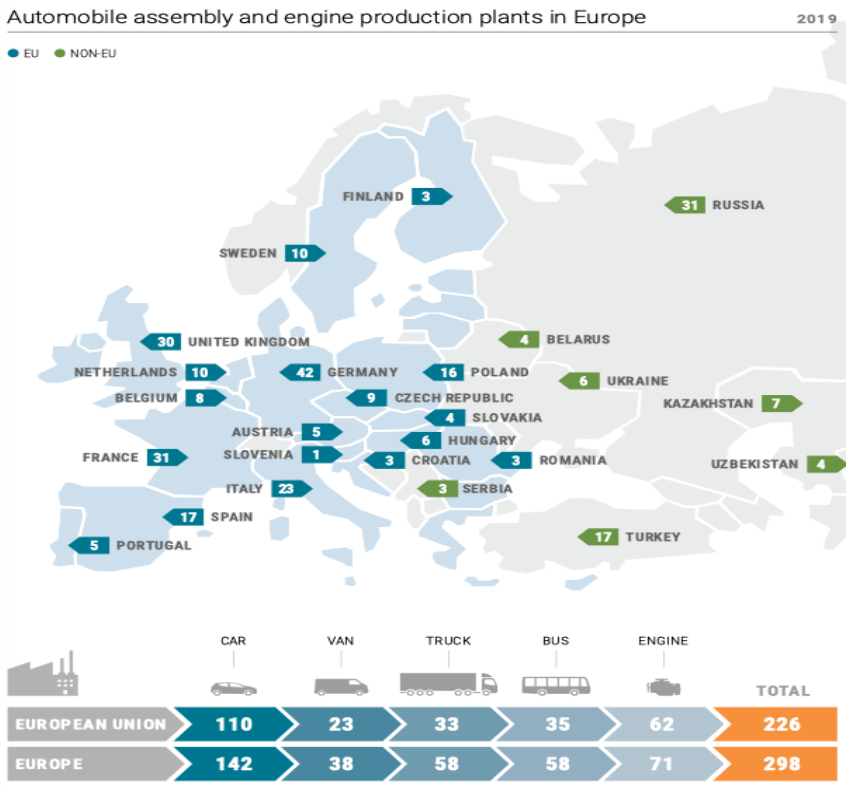
Μια πολλά υποσχόμενη απόπειρα δημιουργίας αυτοκινητοβιομηχανίας στην Ελλάδα είναι η ίδρυση της εταιρείας ΤΕΟΚΑΡ το 1976 από την οικογένεια Θεοχαράκη. Στην ΤΕΟΚΑΡ ανήκαν μονάδες συναρμολόγησης οχημάτων Nissan και το 1979 ανεγέρθηκε ένα εργοστάσιο συναρμολόγησης στη βιομηχανική περιοχή του Βόλου, το οποίο ήταν το δεύτερο μεγαλύτερο σε έκταση μετά τη Steyr ΕΛΒΟ με ταχύτερη συναρμολόγηση και προηγμένο τεχνολογικό εξοπλισμό. Η λειτουργία του εργοστασίου διήρκεσε ως το 1995 και συνολικά παράχθηκαν 170000 κομμάτια κατά τη διάρκεια αυτής της περιόδου. Αξίζει επίσης να αναφερθεί ότι η Nissan στις αρχές της δεκαετίας του 1980 πρότεινε στην οικογένεια Θεοχαράκη την από κοινού ίδρυση μεγαλύτερου εργοστασίου συναρμολόγησης, όμως η πρόταση τελικά αποθαρρύνθηκε από την τότε ελληνική νομοθεσία κι έτσι η ιαπωνική αυτοκινητοβιομηχανία δημιούργησε τη μονάδα αυτή στο Σάντερλαντ της Μεγάλης Βρετανίας.

Σε αυτήν την ενότητα λοιπόν αναφέρθηκαν κάποιες αξιοσημείωτες περιπτώσεις λειτουργίας ή απόπειρας δημιουργίας αυτοκινητοβιομηχανικού κλάδου στην Ελλάδα. Φυσικά και άλλες εταιρείες αποπειράθηκαν να δραστηριοποιηθούν σε αυτόν το τομέα και αξίζει να αναφερθούν η εταιρεία Μαλκότση η οποία παρήγαγε μηχανές και οι πετρελαιοκινητήρες της έβρισκαν εφαρμογή σε βάρκες μέχρι και τρακτέρ. Ακόμα εταιρεία Πετρόπουλος που κατασκεύαζε βιομηχανικά οχήματα από περονοφόρα μέχρι τρακτέρ. Επίσης εταιρείες που σήμερα είναι ενεργές στην εισαγωγή οχημάτων όπως η Σφακιανάκης και η Σαρακάκης κάποτε κατασκεύαζαν οχήματα κυρίως εμπορικής χρήσης (λεωφορεία κλπ.) συνήθως μετατρέποντας και ανακατασκευάζοντας εισαγόμενα μέρη οχημάτων.

Γενικά σχεδόν κάθε εγχείρημα κατέληξε ανεπιτυχές λόγω κοινωνικών και πολιτικών αιτιών έτσι η Ελλάδα δεν κατάφερε να διατηρήσει έναν υγιή κλάδο αυτοκινητοβιομηχανίας. Έτσι καλύπτει τις ανάγκες της σε οχήματα με εισαγωγές από άλλες χώρες όπως θα παρουσιασθεί σε επόμενη ενότητα αυτού του κεφαλαίου.

4.2 Αυτοκινητοβιομηχανία στην Ευρώπη

Αν και η Ελλάδα δε διαθέτει κάποια σημαντική δραστηριότητα στον κλάδο της κατασκευής αυτοκινήτων, η Ευρώπη αποτελούσε το μεγαλύτερο στον κόσμο παραγωγό αυτοκινήτων ενώ από το 2013 βρίσκεται στη δεύτερη θέση με ετήσια παραγωγή σχεδόν 20 εκατομμύρια οχήματα, το 86% περίπου των οποίων αποτελεί επιβατικά αυτοκίνητα. Όπως φαίνεται στο χάρτη (**Σχήμα 4–4**) υπάρχουν συνολικά στην Ευρώπη 298 μονάδες παραγωγής και συναρμολόγησης οχημάτων, οι 226 από αυτές βρίσκονται σε χώρες της Ευρωπαϊκής Ένωσης, ενώ οι 142 από αυτές παράγουν επιβατηγά αυτοκίνητα.



Σχήμα 4–4: Χάρτης εργοστασίων παραγωγής αυτοκινήτων και κινητήρων στην Ευρώπη (www.acea.be)

Σύμφωνα με στοιχεία του 2018 την Ευρώπη απασχολούνται 14,6 εκατομμύρια εργαζόμενοι στον τομέα της αυτοκινητοβιομηχανίας, οι 2,7 εκατομμύρια από αυτούς απασχολούνται άμεσα στην κατασκευή οχημάτων (κατασκευή κινητήρων, αμαξωμάτων κτλ.) ενώ οι υπόλοιποι 11,9 εκατομμύρια εργαζομένων απασχολούνται από την έμμεση κατασκευή (ηλεκτροκινητήρων, γρανάζιών κλπ.) μέχρι τις υπηρεσίες που συνοδεύουν την πώληση αμαξιών, την επισκευή και συντήρηση, την κατασκευή δρόμων κ.ο.κ. (ACEA, 2020).

Γενικά η Ευρωπαϊκή Ένωση αποτελεί έναν σημαντικό εξαγωγέα οχημάτων με εξαγωγές σε όλο τον κόσμο, η Ελλάδα όπως αναλύεται στην επόμενη ενότητα εισάγει μεγάλο μέρος οχημάτων από την Ευρώπη λόγω της εγγύτητάς της.

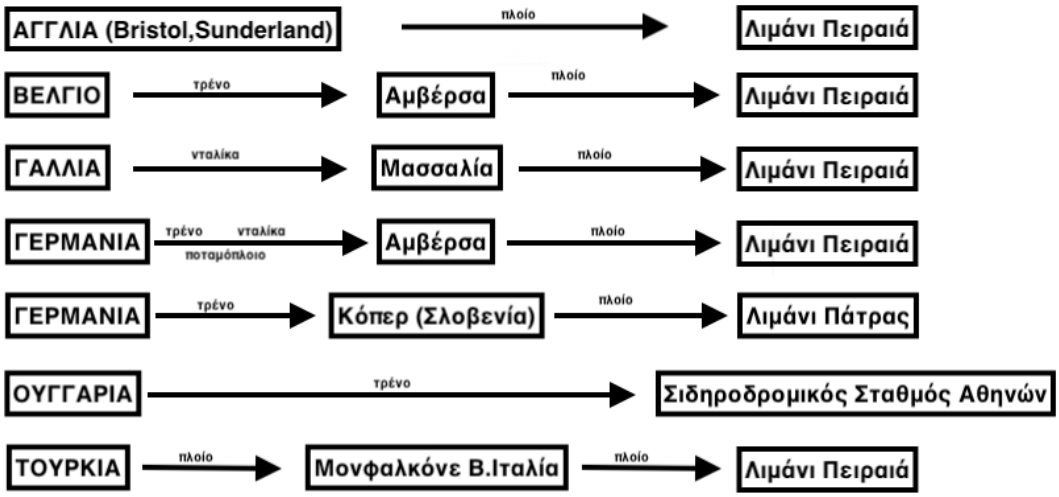
4.3 Προώθηση αυτοκινήτων στην ελληνική αγορά

Η τροφοδοσία της ελληνικής αλλά και άλλων χωρών αγοράς ποικίλλει. Η αποστολή των δρομολογίων καθορίζεται από την επικοινωνία των κατασκευαστικών οίκων με τις θυγατρικές εταιρείες τους σε κάθε χώρα ή με τους αντιπροσώπους τους εισαγωγείς όπως στην περίπτωση της Ελλάδας. Γενικά υπάρχουν διάφοροι τρόποι μεταφοράς των οχημάτων από εργοστάσιο συναρμολόγησης στη χώρα εισαγωγής. Ένα πολύ συνηθισμένο μέσο μεταφοράς είναι οι λεγόμενες αυτοκινητοφόρες, ειδικά φορτηγά που μπορούν να μεταφέρουν οχτώ ή εννέα αυτοκίνητα (**Σχήμα 4–5**).

Σχήμα 4–5: Αυτοκινητοφόρα (en.wikipedia.org)

Άλλος ένας τρόπος μεταφοράς είναι μέσω σιδηροδρόμων, ειδικά στις χώρες της κεντρικής Ευρώπης όπου οι σιδηροδρομικοί οργανισμοί τηρούν αυστηρά κριτήρια ασφαλείας και ταξιδεύουν με υψηλές ταχύτητες, αυτό το μέσο αποτελεί μια διαδεδομένη και συμφέρουσα επιλογή. Ακόμα ένας τρόπος προώθησης αποτελεί η μεταφορά αυτοκινήτων με πλοίο RoRo (Roll-on/Roll-off), πιο συγκεκριμένα χρησιμοποιούνται ειδικά πλοία που έχουν διαφορετικά καταστρώματα για κάθε είδος αυτοκινήτου. Τα αυτοκίνητα δένονται κατάλληλα ώστε σε περίπτωση κακοκαιρίας να μη σημειωθεί μετατόπιση φορτίου.

Σύμφωνα με παλιότερη έρευνα (Μαλινδρέτος, 2008) που μελετά τα φυσικά κανάλια διανομής των αυτοκινητοβιομηχανιών της Ευρώπης από το εργοστάσιο παραγωγής ως τους αντιπροσώπους στην Ελλάδα, σημειώνεται ότι σχεδόν όλα τα κανάλια καταλήγουν στο λιμάνι του Πειραιά με εξαίρεση ένα που καταλήγει στο λιμάνι της Πάτρας κι ένα στο σιδηροδρομικό σταθμό της Αθήνας. Είναι φανερό λοιπόν πως το κυριότερο μέσο μεταφοράς είναι το πλοίο, βέβαια στις περισσότερες περιπτώσεις υπάρχει συνδυασμός μέσων μεταφοράς με εναλλαγή μέσων σε διάφορους στρατηγικά τοποθετημένους σταθμούς. Έτσι το δεύτερο σκέλος της διαδρομής πραγματοποιείται με πλοίο ενώ το πρώτο μπορεί να γίνεται είτε με τρένο είτε με αυτοκινητοφόρα είτε ακόμα και με ποταμόπλοιο. Στο παρακάτω **Σχήμα 4–6** παρατίθενται ενδεικτικά κάποια φυσικά κανάλια διανομής από εργοστάσια παραγωγής της Ευρώπης με προορισμό την ελληνική αγορά. Σύμφωνα με την έρευνα υπάρχουν και αυτοκινητοβιομηχανίες που εισάγουν αυτοκίνητα στην Ελλάδα από εργοστάσια εκτός Ευρώπης (Ασία) χωρίς να γίνει ιδιαίτερη αναφορά.



Σχήμα 4–6: Φυσικά κανάλια διανομής αυτοκινήτων στην Ελλάδα

Τα εργοστάσια αποτελούν την αφετηρία των φυσικών καναλιών διανομής και η γεωγραφική τους θέση επηρεάζει καθοριστικά το κόστος και το χρόνο τροφοδοσίας των ευρωπαϊκών χωρών. Η χρονική διάρκεια της μεταφοράς των οχημάτων από τα εργοστάσια παραγωγής στην Ελλάδα κυμαίνεται από 3 ως 20 ημέρες χωρίς να συνυπολογίζεται ο χρόνος που απαιτείται να φτάσει το αυτοκίνητο από το σημείο εισαγωγής (λιμάνι Πειραιά κλπ.) ως τον αντιπρόσωπο που είναι 8 ως 12 εργάσιμες ημέρες. Δηλαδή από γίνεται λόγος για ένα χρονικό διάστημα λίγο μεγαλύτερο από 11 ημέρες ως ένα μήνα μέχρι να φτάσει το όχημα από το εργοστάσιο παραγωγής στον αντιπρόσωπο. Συνεπώς οι προβλέψεις των πωλήσεων πρέπει να καλύπτουν χρονικό ορίζοντα τουλάχιστον ενός μήνα (χωρίς να υπολογίζονται οι χρόνοι προγραμματισμού και κατασκευής) ώστε να υπάρξει ομαλή διαδικασία εφοδιασμού των αντιπροσώπων από τις αυτοκινητοβιομηχανίες.

Αφού τα οχήματα αφιχθούν στο λιμάνι ακολουθείται η εξής διαδικασία, αρχικά μεταφέρονται στον ειδικό τελωνειακό/αποθηκευτικό χώρο, οι χώροι αυτοί κανονικά βρίσκονται σε εγκαταστάσεις μεταφορικών εταιρειών 3PL (third party logistics) στις οποίες είτε έχουν παραχωρηθεί με κάποια σύμβαση είτε ενοικιάζονται. Σε αυτούς τους χώρους εκτελωνισμού λαμβάνουν χώρα όλες οι απαραίτητες διεργασίες σχετικά με την ομαλή, νόμιμη ένταξη των αυτοκινήτων στις αγορές κι αποτελούν τεράστιες εκτάσεις που ξεκινούν από 10 στρέμματα και μπορεί να ξεπεράσουν και τα 200. Στη συνέχεια της διαδικασίας, μόλις δοθεί μια παραγγελία ακολουθεί ο εκτελωνισμός του αυτοκινήτου στο όνομα του πελάτη κι αναφέρεται στην αρχική τιμή που έχει συμφωνηθεί με τον εκάστοτε έμπορα. Ύστερα αφού ολοκληρωθεί η διεργασία του εκτελωνισμού ακολουθεί η τιμολόγηση και η εξόφληση από τον πελάτη. Έπειτα διενεργείται ο προκαθορισμένος τεχνικός έλεγχος PDI κατά τον οποίο προετοιμάζεται το όχημα, αφαιρούνται τα αυτοκόλλητα προστασίας κατά τη μεταφορά με πλοίο, γίνεται έλεγχος για τυχόν ζημιές, πλύσιμο και πλέον το αυτοκίνητο είναι έτοιμο για παράδοση στον έμπορο από όπου θα το παραλάβει ο πελάτης.

Οι εταιρείες στοχεύουν να πουλήσουν τα αυτοκίνητα μέσα σε ένα χρονικό διάστημα 6-8 μηνών από την στιγμή που θα φθάσουν στο τελωνείο, από εκεί και πέρα υποχρεούνται να εκδώσουν πινακίδες και να τα πουλήσουν σε χαμηλότερες τιμές. Από εδώ προκύπτει κι άλλος ένας λόγος για την κρισιμότητα των προβλέψεων ώστε οι προμηθευτές να κάνουν μια παραγγελία τέτοιου μεγέθους που να καλύπτει και τη μελλοντική ζήτηση αλλά και να μη πουληθούν τα οχήματα σε χαμηλότερες τιμές.

5. Μεθοδολογία

Όπως αναφέρθηκε σκοπός της παρούσας εργασίας είναι η δημιουργία μοντέλων πρόγνωσης με την εισαγωγή δεδομένων διαδικτυακών αναζητήσεων Google Trends και η διεξαγωγή προβλέψεων για τις μηνιαίες πωλήσεις αυτοκινήτων στην Ελλάδα σε **χρονικό ορίζοντα 12 μηνών μπροστά**. Σε αυτό το κεφάλαιο της εργασίας θα περιγραφεί βήμα-βήμα η διαδικασία που ακολουθήθηκε για την κατασκευή αυτών των μοντέλων, για τη διεξαγωγή των προβλέψεων αλλά και για την αξιολόγησή τους.

5.1 Εργαλεία υλοποίησης

Η υλοποίηση των προγνωστικών μοντέλων και η διεξαγωγή των προβλέψεων έγινε από την αρχή ως το τέλος αποκλειστικά με τη χρήση της γλώσσας προγραμματισμού Python 3.7 και των ενσωματωμένων βιβλιοθηκών της στο ολοκληρωμένο περιβάλλον ανάπτυξης (IDE) Jupyter Lab. Η γλώσσα Python τα τελευταία χρόνια έχει αναδειχθεί ως ιδανικό εργαλείο για την διεκπεραίωση τέτοιου είδους εργασιών. Οι βιβλιοθήκες της, τις οποίες ενσωματώνει στο πακέτο διανομής της, διευκολύνουν σε μεγάλο βαθμό διάφορα κομμάτια της Επιστήμης των Δεδομένων (Data Science), από τη συλλογή και την επεξεργασία δεδομένων, μέχρι την κατασκευή μοντέλων προβλέψεων και την αξιολόγησή τους.

Στην παρούσα εργασία χρησιμοποιήθηκαν αρκετές δημοφιλείς βιβλιοθήκες της γλώσσας Python σε διάφορα βήματα της διαδικασίας διεξαγωγής προβλέψεων. Πιο συγκεκριμένα για τη συλλογή και αρχική επεξεργασία δεδομένων απαιτούμενα για τα μοντέλα πρόβλεψης χρησιμοποιήθηκαν οι βιβλιοθήκες pandas και NumPy. Τα μοντέλα που χρησιμοποιήθηκαν ως μέτρο σύγκρισης (benchmark) υλοποιήθηκαν κυρίως με τη βοήθεια των βιβλιοθηκών statsmodels, pmdarima και SKTIME. Τα υπόλοιπα μοντέλα στα οποία έγινε εισαγωγή δεδομένων Google Trends και βασίζονται σε τεχνικές μηχανικής μάθησης (Machine Learning) κι εξόρυξης δεδομένων (Data Mining), υλοποιήθηκαν με τη χρήση της δημοφιλούς βιβλιοθήκης scikit-learn. Η ίδια βιβλιοθήκη επίσης χρησιμοποιήθηκε και για την αξιολόγηση όλων των προβλέψεων με διάφορα μέτρα σφάλματος. Όσον αφορά την απεικόνιση των δεδομένων, γραφημάτων διαγραμμάτων κ.λπ. έγινε χρήση των βιβλιοθηκών Matplotlib και seaborn.

5.2 Συλλογή κι επεξεργασία δεδομένων

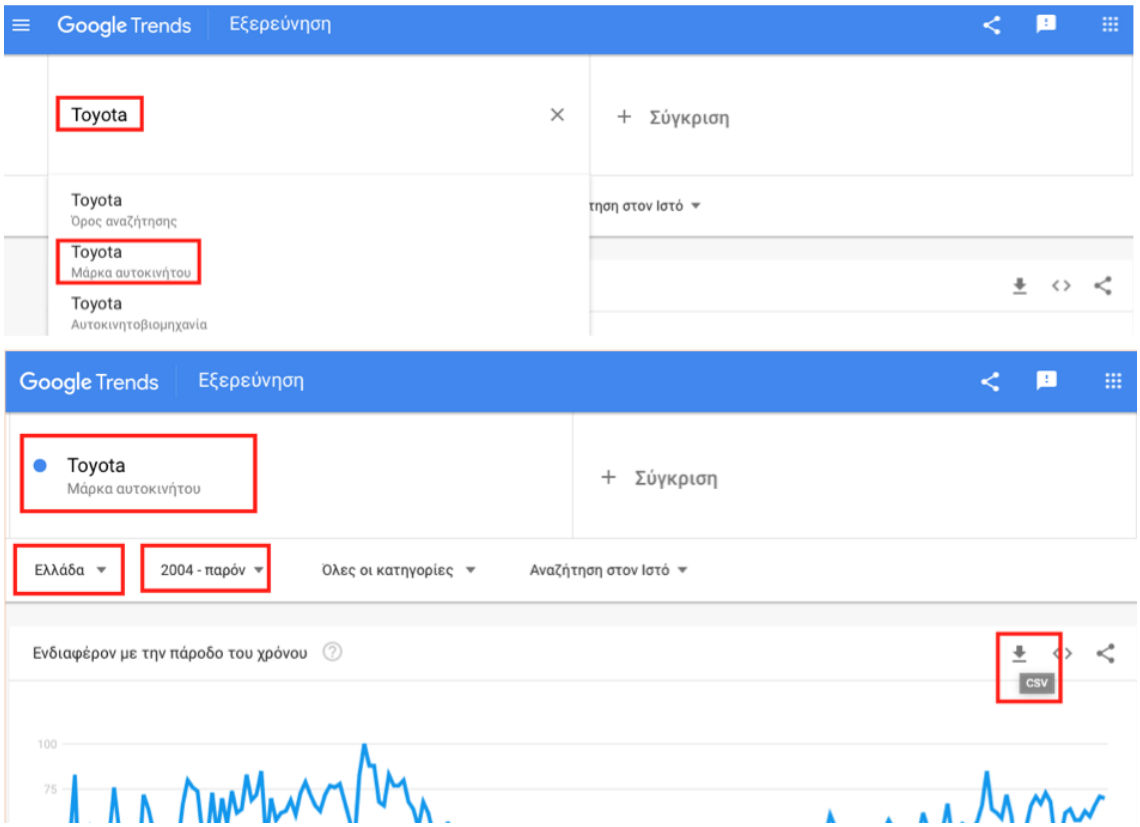
Ουσιαστικά το πρώτο βήμα της διαδικασίας είναι η συλλογή των απαραίτητων δεδομένων για την κατασκευή των προγνωστικών μοντέλων. Η απόφαση για το τι είδους δεδομένα χρειάζονται για τη διεξαγωγή τέτοιας διαδικασίας στηρίχθηκε εν μέρει σε κάποιες προτεινόμενες ιδέες της βιβλιογραφίας.

Τα δεδομένα που συλλέχθηκαν αφορούν τιμές μηνιαίων παρατηρήσεων (όποιες δεν ήταν, μετατράπηκαν σε μηνιαίες) οι οποίες ξεκινούν από τον Ιανουάριο του 2004 και φτάνουν ως τον Ιούλιο του 2021 (πλην ορισμένων εξαιρέσεων). Τα δεδομένα αυτά αναπαριστούν διάφορα μεγέθη που από εδώ και στο εξής θα χαρακτηρίζονται ως

μεταβλητές (ανεξάρτητες κι εξαρτημένη). Παρακάτω γίνεται μια αναλυτικότερη παρουσίαση του τρόπου συλλογής των δεδομένων κάθε μεταβλητής.

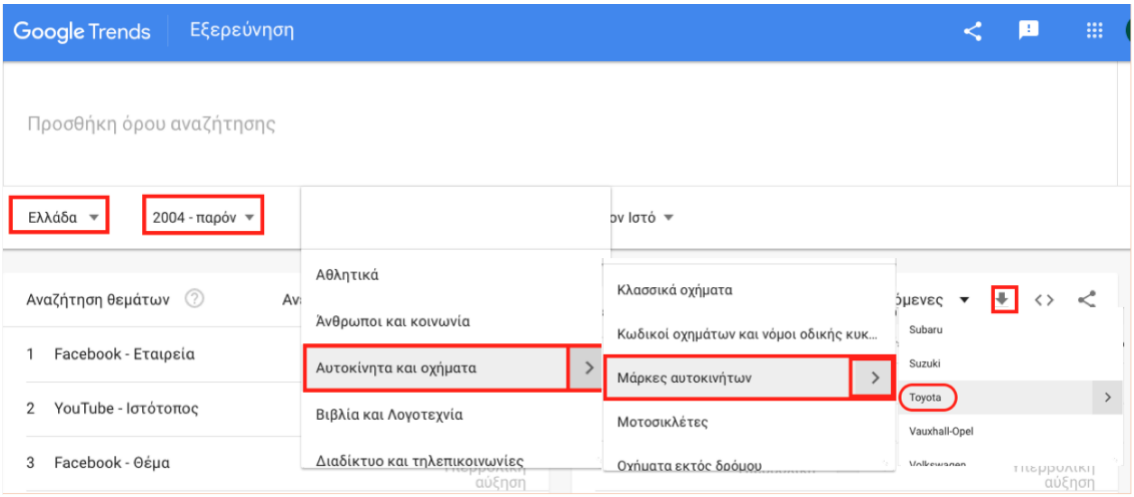
Αρχικά αποφασίστηκε η μελέτη περιπτώσεων για τέσσερις συγκεκριμένες μάρκες αυτοκινήτων, οι οποίες είναι οι **TOYOTA, AUDI, OPEL, FORD**. Για κάθε μια από αυτές τις μάρκες συλλέχθηκαν οι **μηνιαίες πωλήσεις** τους στην Ελλάδα. Η πηγή αυτών των στοιχείων ήταν η ιστοσελίδα του Συνδέσμου Εισαγωγέων Αντιπροσώπων Αυτοκινήτων (ΣΕΑΑ) https://seaa.gr. Για να μην υπάρξουν παρανοήσεις, τα στοιχεία αυτά αφορούν μόνο νέες ταξινομήσεις επιβατικών οχημάτων και όχι μεταπωλήσεις μεταχειρισμένων.

Επίσης για κάθε μια από τις παραπάνω μάρκες έγινε εξαγωγή δεδομένων από την πλατφόρμα Google Trends, που αφορούσαν τη δημοτικότητα της αναζήτησης του ονόματος κάθε μάρκας αυτοκινήτου στη μηχανή αναζήτησης της Google. Για κάθε μάρκα αυτοκινήτου εξάχθηκαν **δύο χρονοσειρές** από τα **Google Trends** ακολουθώντας για την καθεμία διαφορετική προσέγγιση. Η πρώτη προσέγγιση εύρεσης δεδομένων από την εφαρμογή έγινε πληκτρολογώντας το όνομα της μάρκας στο κατάλληλο πεδίο της ενότητας Εξερεύνηση, φυσικά ρυθμίζοντας κατάλληλα τις παραμέτρους τοποθεσίας και χρονικού διαστήματος. Αφού πληκτρολογήθηκε το όνομα, εμφανίζεται ένα πλαίσιο με επιλογές αυτόματης συμπλήρωσης. Στη συγκεκριμένη περίπτωση επιλέχθηκε η ένδειξη «Μάρκα αυτοκινήτου», όπως φαίνεται στο **Σχήμα 5–1**, έπειτα με το πάτημα του κουμπιού CSV εξάγεται η χρονοσειρά.



Σχήμα 5–1: Εξαγωγή δεδομένων Google Trends με την πρώτη προσέγγιση (trends.google.com)

Με το δεύτερο τρόπο που εξήχθησαν δεδομένα από το Google Trends δεν χρειάστηκε να γίνει πληκτρολόγηση κάποιου όρου, απλά από την ενότητα Εξερεύνηση και αφού καθοριστούν κατάλληλα η τοποθεσία και το χρονικό παράθυρο, στο πεδίο «Κατηγορίες» επιλέχθηκε « Αυτοκίνητα και οχήματα» έπειτα «Μάρκες αυτοκινήτων» και μετά η μάρκα για την οποία πρέπει να εξαχθεί η χρονοσειρά, όπως φαίνεται στο **Σχήμα 5–2**.



Σχήμα 5–2: Εξαγωγή δεδομένων Google Trends με την δεύτερη προσέγγιση (trends.google.com)

Η συλλογή δύο διαφορετικών χρονοσειρών Google Trends για κάθε μάρκα αποφασίστηκε με το σκεπτικό ότι ίσως κάθε μία από αυτές τις χρονοσειρές δίνει διαφορετικές πληροφορίες στο προγνωστικό μοντέλο.

Εκτός από τις παραπάνω μεταβλητές που αφορούν κάθε μάρκα ξεχωριστά, κρίθηκε σκόπιμο —λαμβάνοντας υπόψιν και προτάσεις στη βιβλιογραφία— να συλλεχθούν για το ίσιο χρονικό διάστημα και άλλες μεταβλητές κυρίως οικονομικής φύσεως που να μπορούν να αποτυπώσουν την οικονομική κατάσταση που επικρατεί στη χώρα άρα συνεπώς και στην αγοραστική δύναμη των καταναλωτών. Έτσι λοιπόν έγινε συλλογή δεδομένων όπως το μηνιαίο **ποσοστό ανεργίας** στην Ελλάδα και το τριμηνιαίο **Ακαθάριστο Εγχώριο Προϊόν** (μετατράπηκε σε μηνιαίες παρατηρήσεις) τα οποία εξάχθηκαν από την ιστοσελίδα της Ελληνικής Στατιστικής Αρχής (ΕΛΣΤΑΤ) https://www.statistics.gr. Ακόμα συλλέχθηκαν μεταβλητές όπως οι **τιμές της βενζίνης** στην Ελλάδα **με φόρο** και **χωρίς φόρο**, η πηγή αυτών των δεδομένων ήταν ο διαδικτυακός τόπος της Ευρωπαϊκής Επιτροπής και συγκεκριμένα η ηλεκτρονική διεύθυνση https://ec.europa.eu/energy/data-analysis/weekly-oil-bulletin_en. Επίσης έγινε συλλογή των μηνιαίων τιμών κλεισίματος του **Γενικού Δείκτη του Χρηματιστηρίου Αθηνών** από τον ιστότοπο capital.gr.

Τέλος σε μια προσπάθεια να αποτυπωθεί η οικονομική κρίση του 2009 κι έπειτα, συλλέχθηκαν πάλι χρονοσειρές από την πλατφόρμα του Google Trends, αυτήν τη φορά σχετικά με την αναζήτηση των όρων **«οικονομική κρίση» (ως θέμα)**, **«μνημόνιο» (ως θέμα)**, **«χρέος» (ως θέμα)**, **«περικοπές»** και **«απολύσεις»**.

Όλες οι παραπάνω μεταβλητές χαρακτηρίζονται ως ανεξάρτητες διότι βοηθούν στην πρόγνωση της εξαρτημένης που είναι οι πωλήσεις. Βέβαια και οι πωλήσεις μπορούν να χαρακτηριστούν ως ανεξάρτητες αν γίνει λόγος για παρελθούσες πωλήσεις πχ. πωλήσεις 12 μήνες πριν από την τιμή πρέπει να προβλεφθεί.

Συνοψίζοντας λοιπόν και για λόγους ευκρίνειας, στον παρακάτω πίνακα (**Πίνακας 5-1**) συγκεντρώνονται οι μεταβλητές που συλλέχθηκαν για τις ανάγκες της παρούσας

εργασίας. Στην τρίτη στήλη σημειώνεται η ονομασία που έχει η κάθε μεταβλητή στον κώδικα ή στις στήλες του ολοκληρωμένου πίνακα δεδομένων.

Πίνακας 5-1: Συγκεντρωτική παράθεση μεταβλητών που συλλέχθηκαν

#	**Μεταβλητή**	**Σύμβολο στον κώδικα**
1α, 1β, 1γ, 1δ	Μηνιαίες πωλήσεις για κάθε μάρκα	TOYOTA, AUDI, OPEL, FORD
2α, 2β, 2γ, 2δ	Δεδομένα Google Trends για κάθε μάρκα (1ος τρόπος εξαγωγής)	gt1_toyota, gt1_audi, gt1_opel, gt1_ford
3α, 3β, 3γ, 3δ	Δεδομένα Google Trends για κάθε μάρκα (2ος τρόπος εξαγωγής)	gt2_toyota, gt2_audi, gt2_opel, gt2_ford
4	Ποσοστό Ανεργίας	unemployment
5	Μηνιαίο ΑΕΠ	Qgdp_Maverage
6	Τιμή βενζίνης με φόρο	gas_wt
7	Τιμή βενζίνης χωρίς φόρο	gas_wot
8	Γενικός Δείκτης Χρηματιστηρίου Αθηνών	gd_at
9	Google Trends «οικονομική κρίση» (ως θέμα)	gt_crisisTh
10	Google Trends «μνημόνιο» (ως θέμα)	gt_memTh
11	Google Trends «χρέος» (ως θέμα)	gt_debtTh
12	Google Trends «περικοπές»	gt_cuts
13	Google Trends «απολύσεις»	gt_fir

Ολόκληρο το σύνολο των δεδομένων παρατίθεται στο **Παράρτημα Ι** της παρούσας εργασίας.

5.3 Διαχώριση δεδομένων – δεδομένα ελέγχου

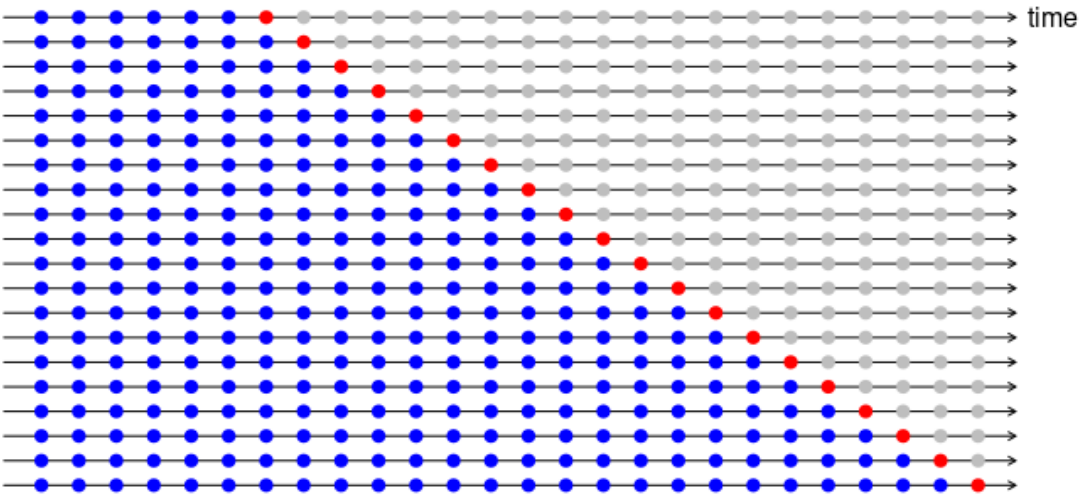
Όπως αναφέρθηκε προηγουμένως, για τους σκοπούς της παρούσας εργασίας συλλέχθηκαν δεδομένα από τον Ιανουάριο του 2004 ως τον Ιούλιο του 2021. Για τη διεξαγωγή όμως των προβλέψεων και κυρίως για την αξιολόγησή τους τα δεδομένα αυτά πρέπει να χωρισθούν σε δύο μέρη: τα δεδομένα εκπαίδευσης (training data) και τα δεδομένα ελέγχου (test data ή αλλιώς hold out data). Τα δεδομένα εκπαίδευσης είναι τα δεδομένα στα οποία το μοντέλο θα προσπαθήσει να προσαρμοστεί ρυθμίζοντας κατάλληλα τις παραμέτρους του. Τα δεδομένα αυτά είναι «γνωστά» στο μοντέλο και συνήθως αποτελούν ποσοστό από τα δεδομένα ελέγχου. Τα δεδομένα ελέγχου «κρατούνται στην άκρη», είναι «άγνωστα» στο μοντέλο και προορίζονται για την αξιολόγηση του μετρώντας την απόκλιση των προβλέψεων από τις πραγματικές τιμές που περιέχονται σε αυτό το σύνολο δεδομένων.

Για τη διεξαγωγή των προβλέψεων στην παρούσα εργασία, τα δεδομένα χωρίστηκαν ως εξής: από τον Ιανουάριο 2004 ως τον Αύγουστο του 2018 ορίστηκαν τα δεδομένα εκπαίδευσης και από το Σεπτέμβριο του 2018 ως τον Ιούλιο του 2021 ορίστηκαν τα

δεδομένα ελέγχου. Δηλαδή τα μοντέλα που κατασκευάστηκαν πραγματοποίησαν μηνιαίες προβλέψεις για το χρονικό 09-2018 ως 07-2021 και τα μέτρα σφάλματος που υπολογίστηκαν κατά την αξιολόγησή τους στην ουσία μετρούν την απόκλιση των προβλέψεων από τις πραγματικές τιμές αυτό το διάστημα. Κάθε προγνωστικό μοντέλο διεξαγάγει προβλέψεις για το ίδιο χρονικό διάστημα (δεδομένα ελέγχου) ώστε να είναι δυνατή η σύγκριση της απόδοσής του με αυτήν των υπολοίπων. Σε αυτήν την περίπτωση που αποτελεί την πιο απλή προσέγγιση σχετικά με το διαχωρισμό των δεδομένων με σκοπό την αξιολόγηση του μοντέλου, το μέγεθος των δύο τμημάτων παραμένει σταθερό (**Σχήμα 5–3**). Υπάρχουν και πιο περίπλοκες προσεγγίσεις σχετικά με τους τρόπους διαχωρισμούς των δεδομένων για τις οποίες δεν αξίζει να γίνει αναφορά στην παρούσα εργασία, εκτός ίσως από την προσέγγιση του διερυνόμενου παραθύρου (expanding window).

Σχήμα 5–3: Απλός διαχωρισμός δεδομένων σε δεδομένα εκπαίδευσης κι ελέγχου (https://otexts.com/fpp2/accuracy.html)

Η προσέγγιση του διερυνόμενου παραθύρου (expanding window) χρησιμοποιείται συχνά στα προβλήματα προβλέψεων που έχουν να κάνουν με το χρόνο, δηλαδή διαχειρίζονται δεδομένα χρονοσειρών. Σε αυτήν την περίπτωση τα μεγέθη των τμημάτων στα οποία χωρίστηκαν τα δεδομένα μεταβάλλονται με το χρόνο. Κάθε φορά που γίνεται μια πρόβλεψη, υποτίθεται ότι η πραγματική τιμή της προηγούμενης παρατήρησης γίνεται γνωστή κι έτσι προστίθεται στα δεδομένα εκπαίδευσης, συνεπώς το μέγεθος των δεδομένων εκπαίδευσης αυξάνεται κατά ένα χρονοσημείο κάθε φορά που γίνεται πρόβλεψη. Στην ουσία κάθε φορά ένα διαφορετικό μοντέλο προσαρμόζεται στο ολοένα και διερυνόμενο σετ δεδομένων εκπαίδευσης. Το σφάλμα υπολογίζεται ως ο μέσος όρος των σφαλμάτων κάθε πρόβλεψης. Η μέθοδος αυτή αναπαρίσταται σχηματικά παρακάτω στο **Σχήμα 5–4** .



Σχήμα 5–4: Διαχωρισμός δεδομένων με την προσέγγιση διερυνόμενου παραθύρου - expanding window (https://otexts.com/fpp2/accuracy.html)

Η μέθοδος αυτή χρησιμοποιήθηκε στην παρούσα εργασία για τη διεξαγωγή προβλέψεων με τα μοντέλα μέτρου σύγκρισης (benchmarks), συγκεκριμένα στα μοντέλα εκθετικής εξομάλυνσης και ARIMA, για αυτό και περιγράφεται σε αυτήν την ενότητα.

5.4 Μοντέλα μέτρου σύγκρισης (benchmarks)

Για να μπορέσει να γίνει αξιολόγηση των μοντέλων που δημιουργήθηκαν με τα δεδομένα Google Trends, και για να γίνει εύκολα αντιληπτό αν αυτά τα μοντέλα πέτυχαν το στόχο τους ο οποίος είναι η διεξαγωγή προβλέψεων με βελτιωμένη ακρίβεια, πρέπει να ορισθεί κάποιο μέτρο σύγκρισης. Έτσι θα μπορεί να διαπιστωθεί αν η αποτελεσματικότητα των εξεταζόμενων μοντέλων είναι ικανοποιητική.

Ως μέτρα σύγκρισης (benchmarks) λοιπόν στην παρούσα εργασία, δημιουργήθηκαν κάποια μοντέλα βασισμένα σε παραδοσιακές μεθόδους προβλέψεων όπως η εκθετική εξομάλυνση και τα μοντέλα ARIMA, οι οποίες χρησιμοποιούνται εδώ και πολλά χρόνια σε προβλήματα πρόγνωσης ζήτησης.

Πιο συγκεκριμένα ως μέτρα σύγκρισης δημιουργήθηκαν δύο μοντέλα εκθετικής εξομάλυνσης, το πρώτο είναι τύπου Holt-Winters' με προσθετική εποχικότητα 12 μηνών και υλοποιήθηκε με τη βοήθεια της βιβλιοθήκης statsmodels. Το δεύτερο είναι μια αυτοματοποιημένη μορφή εκθετικής εξομάλυνσης που επιλέγει τις παραμέτρους του μοντέλου ανάλογα με τα εκάστοτε δεδομένα, λέγεται AutoETS και προέρχεται από τη βιβλιοθήκη της Python SKTIME. Το δεύτερο μοντέλο δημιουργήθηκε περισσότερο σαν έλεγχος για το πρώτο για να φανεί αν η επιλογή του παραμέτρων του πρώτου ήταν η βέλτιστη. Τα δύο αυτά μοντέλα μετέπειτα στην εργασία θα αναφέρονται με τα ονόματα «Μοντέλο 1» και «Μοντέλο 2» αντίστοιχα.

Επίσης δημιουργήθηκε ένα αυτοπαλινδρομικό μοντέλο κινητού μέσου όρου (ARIMA) το οποίο υλοποιήθηκε με τη βοήθεια της βιβλιοθήκης pmdarima. Η βιβλιοθήκη αυτή παρέχει τη δυνατότητα για δημιουργία μοντέλου ARIMA στο οποίο οι παράμετροί του (p,d,q)

επιλέγονται αυτόματα ανάλογα με τα εκάστοτε δεδομένα διενεργώντας διάφορους ελέγχους για το αν ικανοποιούνται ορισμένα κριτήρια. Έτσι η όλη διαδικασία καθίσταται λιγότερο χρονοβόρα κι εξαλείφεται ως ένα βαθμό η αμφιβολία για την επιλογή της βέλτιστης τάξης (p,d,q) του μοντέλου. Η βιβλιοθήκη pmdarima βασίζεται στην βιβλιοθήκη statsmodels. Το μοντέλο αυτό θα αναφέρεται μετέπειτα στην εργασία ως «Μοντέλο 3».

Οι προβλέψεις που διεξάχθηκαν με τα παραπάνω μοντέλα έγιναν ακολουθώντας την τεχνική expanding window, έτσι για την διεξαγωγή κάθε μοναδικής πρόβλεψης εφαρμόστηκε μοντέλο με διαφορετικές παραμέτρους για αυτό το λόγο και δεν αναφέρονται τα ακριβή χαρακτηριστικά των παραπάνω μοντέλων.

Τέλος δημιουργήθηκε ακόμα ένα μοντέλο για μέτρο σύγκρισης, λαμβάνοντας υπόψιν πρακτικές που ακολουθήθηκαν στη βιβλιογραφία. Πρόκειται για ένα απλό αυτοπαλινδρομικό μοντέλο, το οποίο θεωρεί ότι η τιμή της ζητούμενης πρόβλεψης είναι γραμμικός συνδυασμός της προηγούμενης πραγματικής τιμής και αυτής 12 μηνών πριν συμπεριλαμβανομένου ενός μεγέθους σφάλματος. Το παραπάνω μοντέλο έχει προταθεί ως μέτρο σύγκρισης σε διάφορες μελέτες αν και περιέχει την πληροφορία της αμέσως προηγούμενης πραγματικής παρατήρησης από την πρόβλεψη κάτι που δε συμβαδίζει με τα υπόλοιπα μοντέλα των οποίων ο χρονικός ορίζοντας πρόβλεψης είναι 12 μήνες. Το συγκεκριμένο μοντέλο που μετέπειτα στην εργασία θα αναφέρεται ως «Μοντέλο 4» προσαρμόστηκε με την μέθοδο ελάχιστών τετραγώνων στα δεδομένα εκπαίδευσης με τη συνδρομή της βιβλιοθήκης scikit-learn και η σχέση που το περιγράφει εμφανίζεται στην ακόλουθη εξίσωση:

$$S_t = \beta_0 + \beta_1 S_{t-1} + \beta_2 S_{t-12} + \varepsilon_t$$

Όλα τα παραπάνω μοντέλα διεξήγαγαν προβλέψεις με χρονικό ορίζοντα 12 μηνών για τη χρονική περίοδο Σεπτεμβρίου 2018 ως Ιούλιο του 2021 (περίοδος δεδομένων ελέγχου). Η απόκλιση μεταξύ των προβλέψεων και των πραγματικών παρατηρήσεων υπολογίστηκε με τη χρήση μέτρων σφαλμάτων και τα αποτελέσματα παρατίθενται για σύγκριση σε επόμενο κεφάλαιο.

5.5 Επιλογή μεθόδων πρόβλεψης

Καθώς στην παρούσα εργασία διερευνάται αν η μεταβολή της δημοτικότητας των αναζητούμενων όρων στη μηχανή της Google μπορεί να συσχετιστεί με το πιθανό ενδιαφέρον των καταναλωτών για την αγορά ορισμένων προϊόντων, έγινε η υπόθεση ότι τα δεδομένα Google Trends (όπως και οι υπόλοιπες οικονομικές μεταβλητές) επεξηγούν κατά ένα μέρος το μέγεθος των πωλήσεων. Έτσι για αυτόν τον λόγο κρίθηκε σκόπιμο να δημιουργηθούν προγνωστικά μοντέλα βασισμένα σε αιτιοκρατικές ή αλλιώς επεξηγηματικές μεθόδους πρόβλεψης αλλά και σε προηγμένες (στην πραγματικότητα πρόκειται για αιτιοκρατικές με χρήση πιο περίπλοκων τεχνικών υλοποίησης) σύμφωνα με την κατηγοριοποίηση που έχει γίνει σε προηγούμενο κεφάλαιο.

Το προγνωστικά μοντέλα που δημιουργήθηκαν για τις ανάγκες αυτής της διπλωματικής εργασίας βασίζονται στις ακόλουθες μεθόδους προβλέψεων:

- **Παλινδρόμηση με δένδρα αποφάσεων** (Decision Tree Regression) υλοποιήθηκε χρησιμοποιώντας το scikit-learn, το μέγιστο βάθος του δένδρου ορίστηκε στο 4 και η τυχαία κατάσταση στο 0, οι υπόλοιπες παράμετροι παρέμειναν προκαθορισμένες
- **Παλινδρόμηση με τυχαία δάση** (Random Forest Regression) υλοποιήθηκε χρησιμοποιώντας το scikit-learn, το μέγιστο βάθος κάθε δένδρου ορίστηκε στο 4 και η τυχαία κατάσταση στο 0, οι υπόλοιπες παράμετροι παρέμειναν προκαθορισμένες
- **Πολλαπλή γραμμική παλινδρόμηση** (Multiple Linear Regression) υλοποιήθηκε χρησιμοποιώντας το scikit-learn, όλες οι παράμετροι παρέμειναν προκαθορισμένες
- **Παλινδρόμηση με μηχανές διανυσμάτων υποστήριξης** (Support Vector Regression) υλοποιήθηκε χρησιμοποιώντας το scikit-learn, η παράμετρος C ορίστηκε 15, η epsilon 0.4, ο πυρήνας «γραμμικός», οι υπόλοιπες παράμετροι παρέμειναν προκαθορισμένες. Επίσης έγινε μετασχηματισμός των δεδομένων για καλύτερη απόδοση του μοντέλου.

5.6 Επιλογή ανεξάρτητων μεταβλητών

Ο αρχικός αριθμός των ανεξάρτητων μεταβλητών που συλλέχθηκε για την κατασκευή των προγνωστικών μοντέλων όπως περιγράφηκε στην ενότητα 5.2 είναι αρκετά μεγάλος. Σε αυτές τις περιπτώσεις συνήθως ενδείκνυται η μείωση του αριθμού των ανεξάρτητων μεταβλητών, εξετάζοντας με κάποιο τρόπο την προγνωστική τους ισχύ και αποφασίζοντας αν πρέπει να συμπεριληφθούν στο μοντέλο ή όχι.

Στην παρούσα εργασία η απόφαση συμπερίληψης κάθε μιας μεταβλητής βασίστηκε εν μέρει στην ικανοποίηση δυο συγκεκριμένων κριτηρίων. Αρχικά υπολογίστηκε ο συντελεστής συσχέτισης Pearson κάθε μίας από τις ανεξάρτητες μεταβλητές με αυτήν του αριθμού πωλήσεων (οι πωλήσεις μετατοπίστηκαν κατά 12 μήνες λόγω του χρονικού ορίζοντα πρόβλεψης), την εξαρτημένη δηλαδή. Η ικανοποίηση του άλλου κριτηρίου εξετάστηκε με τη βοήθεια της λειτουργίας Recursive Feature Elimination with Cross-Validation της βιβλιοθήκης scikit-learn. Πρόκειται για έναν αλγόριθμο ο οποίος διεξαγάγει προβλέψεις για κάποια τμήματα των δεδομένων εκπαίδευσης. Στην αρχή οι προβλέψεις διεξάγονται με όλες τις μεταβλητές στο μοντέλο και μετριέται κάποιο σφάλμα, ύστερα αφαιρείται μια μεταβλητή και ξαναγίνονται οι προβλέψεις και ξανά μετριέται το σφάλμα, αυτή η διαδικασία επαναλαμβάνεται μέχρι να μείνει μόνο μια μεταβλητή στο τέλος. Μόλις γίνει αυτό από τη σύγκριση των μέτρων σφάλματος κάθε προβλέψεις επιλέγεται ο βέλτιστος συνδυασμός ανεξάρτητων μεταβλητών για το μοντέλο. Η σειρά με την οποία θα αφαιρεθούν οι μεταβλητές καθορίζεται με βάση κάποια βαθμολογία που ο αλγόριθμος χαρακτηρίζει την κάθε μεταβλητή.

Λαμβάνοντας υπόψιν τα δύο κριτήρια που περιεγράφηκαν παραπάνω αποφασίστηκε μια διαλογή των μεταβλητών που θα παρουσιαστεί αναλυτικά στην επόμενη ενότητα του κεφαλαίου.

5.7 Τελικά προγνωστικά μοντέλα

Στην ενότητα αυτή παρουσιάζονται τα προγνωστικά μοντέλα που χρησιμοποιήθηκαν σε αυτήν την εργασία για τη διεξαγωγή προβλέψεων. Μετά από διάφορες δοκιμές και λαμβάνοντας υπόψιν το περιεχόμενο των ενοτήτων 5.5 και 5.6 κατασκευάστηκαν 4 προγνωστικά μοντέλα με εισαγωγή Google Trends και άλλα 4 ταυτόσημα των οποίων η μόνη τους διαφορά με τα πρώτα είναι η αφαίρεση των μεταβλητών Google Trends. Αυτό έγινε για να εξεταστεί πιο αναλυτικά η συνεισφορά των δεδομένων Google Trends στην απόδοση των μοντέλων.

Στον παρακάτω πίνακα (**Πίνακας 5-2**) λοιπόν συγκεντρώνονται τα προγνωστικά μοντέλα της εργασίας με λεπτομέρειες για τη μέθοδο στην οποία βασίστηκαν και για τις ανεξάρτητες μεταβλητές τις οποίες τελικά συμπεριέλαβαν. Η ανεξάρτητη μεταβλητή Πωλήσεις* αναφέρεται φυσικά στην εισαγωγή παρελθοντικών δεδομένων πωλήσεων και οι υπόλοιπες μεταβλητές είναι μετατοπισμένες έχοντας καθυστέρηση (lag) 12 μηνών (Δηλαδή οι πωλήσεις του τρέχοντος μήνα εξαρτώνται από τις ανεξάρτητες μεταβλητές 12 μηνών πριν). Το gt1 αναφέρεται στα δεδομένα Google Trends σχετικά με τη μάρκα αυτοκινήτου τα οποία εξάχθηκαν με την πρώτη προσέγγιση όπως αναφέρεται στην παρούσα εργασία ενώ το gt2 σε αυτά που εξάχθηκαν με τη δεύτερη. Το gt(απολύσεις) αναφέρεται στις αναζητήσεις του όρου «απολύσεις» όπως περιγράφηκε στην ενότητα 5.2.

Πίνακας 5-2: Προγνωστικά μοντέλα

Όνομα	**Μέθοδος Πρόβλεψης**	**Ανεξάρτητες Μεταβλητές**
Μοντέλο 1	Εκθετική Εξομάλυνση	Πωλήσεις*
Μοντέλο 2	Εκθετική Εξομάλυνση	Πωλήσεις*
Μοντέλο 3	ARIMA	Πωλήσεις*
Μοντέλο 4	Απλό Αυτοπαλινδρομικό	Πωλήσεις*
Μοντέλο 5	Παλινδρόμηση με Δέντρο Αποφάσεων	Πωλήσεις*, gt1, gt2, ανεργία, ΑΕΠ, gt(απολύσεις)
Μοντέλο 6	Παλινδρόμηση με Δέντρο Αποφάσεων	Πωλήσεις*, ανεργία, ΑΕΠ
Μοντέλο 7	Παλινδρόμηση με Τυχαία Δάση	Πωλήσεις*, gt1, gt2, ανεργία, ΑΕΠ, gt(απολύσεις)
Μοντέλο 8	Παλινδρόμηση με Τυχαία Δάση	Πωλήσεις*, ανεργία, ΑΕΠ
Μοντέλο 9	Γραμμική Παλινδρόμηση	Πωλήσεις*, gt1, ανεργία, ΑΕΠ
Μοντέλο 10	Γραμμική Παλινδρόμηση	Πωλήσεις*, ανεργία, ΑΕΠ
Μοντέλο 11	Παλινδρόμηση με διανύσματα υποστήριξης	Πωλήσεις*, gt1, ανεργία, ΑΕΠ
Μοντέλο 12	Παλινδρόμηση με διανύσματα υποστήριξης	Πωλήσεις*, ανεργία, ΑΕΠ

Τα παραπάνω μοντέλα αφού εκπαιδεύτηκαν με τα δεδομένα της περιόδου Ιανουάριος 2004 ως Αύγουστος 2018 ή με ένα τμήμα αυτών των δεδομένων, ύστερα διενέργησαν προβλέψεις για το χρονικό διάστημα Σεπτέμβριος 2019 ως Ιούλιος 2021.

Ολόκληρος ο κώδικας με τον οποίο υλοποιήθηκε η διεξαγωγή προβλέψεων παρουσιάζεται στο **Παράρτημα ΙΙ** της παρούσας εργασίας.

5.8 Αξιολόγηση

Για την αξιολόγηση της απόδοσης κάθε μοντέλου υπολογίστηκαν κάποια μέτρα σφάλματος για την περίοδο διεξαγωγής των προβλέψεων Σεπτέμβριος 2019 ως Ιούλιος 2021. Τα σφάλματα αυτά μετρούν την απόκλιση μεταξύ των προβλέψεων και των πραγματικών τιμών. Όσο μικρότερη η τιμή του σφάλματος τόσο μεγαλύτερη η ακρίβεια του μοντέλου συνεπώς και καλύτερη η απόδοση του. Τα μέτρα σφάλματος που υπολογίστηκαν με τη βοήθεια της βιβλιοθήκης scikit-learn στην παρούσα εργασία είναι τα εξής:

- Ρίζα μέσου τετραγωνικού σφάλματος (**RMSE**)
- Μέσο απόλυτο ποσοστιαίο σφάλμα (**MAPE**)
- Μέσο απόλυτο σφάλμα (**MAE**)

6. Παρουσίαση Αποτελεσμάτων

Σε αυτό το κεφάλαιο παρουσιάζονται —για κάθε μάρκα που εξετάζεται στην παρούσα εργασία— οι τιμές των σφαλμάτων που υπολογίστηκαν για τις προβλέψεις κάθε μοντέλου στο χρονικό διάστημα Σεπτέμβριος 2018 ως Ιούλιος 2021. Επίσης παρατίθενται επεξηγήσεις για τον τρόπο που παρουσιάζονται τα αποτελέσματα μαζί με κάποια σχόλια.

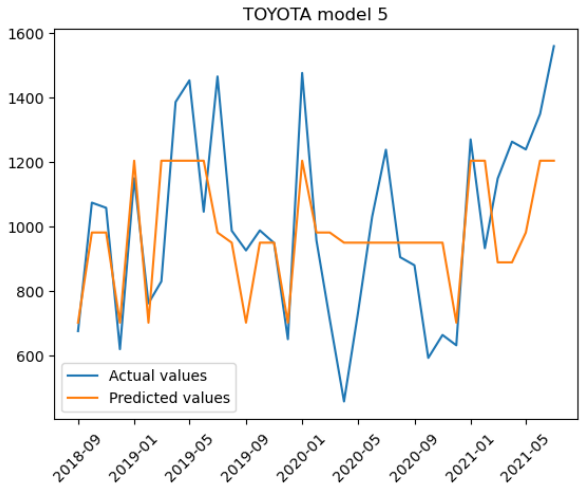
Στον παρακάτω πίνακα (**Πίνακας 6-1**) λοιπόν συγκεντρώνονται όλες οι τιμές σφάλματος που προέκυψαν κατά τη διενέργεια προβλέψεων από το κάθε διαφορετικό μοντέλο.

Πίνακας 6-1: Συγκεντρωτικός πίνακας σφαλμάτων

		Benchmarks				GT		GT		GT		GT	
TOYOTA	Σφάλμα \ Μοντέλο	**1**	2	3	4	**5**	6	7	8	9	10	11	12
	RMSE	**275**	286	306	276	**228**	286	262	281	293	296	282	283
	MAPE	**24,1%**	24,4%	27,7%	23,6%	**20,2%**	24,9%	23,2%	23,6%	28,6%	29,5%	24,2%	24,8%
	MAE	**213**	222	241	220	**183**	226	212	227	228	235	215	219
AUDI	Σφάλμα \ Μοντέλο	1	2	3	4	5	6	**7**	**8**	9	10	11	12
	RMSE	166	167	158	155	148	155	**139**	**144**	160	153	151	149
	MAPE	56,5%	58,2%	60,6%	74,0%	65,7%	72,1%	**58,3%**	**63,1%**	73,0%	63,0%	63,9%	64,5%
	MAE	123	123	121	109	106	118	**101**	**106**	124	116	114	114
OPEL	Σφάλμα \ Μοντέλο	1	2	3	**4**	5	6	**7**	8	9	10	11	12
	RMSE	247	250	246	**195**	224	229	**198**	206	220	243	208	220
	MAPE	50,5%	46,7%	47,2%	**38,6%**	44,9%	43,7%	**41,5%**	44,9%	43,5%	59,2%	47,4%	52,1%
	MAE	193	183	183	**150**	180	175	**153**	165	163	200	164	179
FORD	Σφάλμα \ Μοντέλο	1	2	3	**4**	5	6	7	**8**	9	10	11	12
	RMSE	103	117	124	**90**	107	157	100	**97**	122	122	115	99
	MAPE	43,6%	50,4%	52,2%	**37,1%**	44,1%	56,7%	44,1%	**42,7%**	54,4%	54,4%	50,9%	43,6%
	MAE	78	89	98	**75**	90	124	80	**76**	97	97	88	79

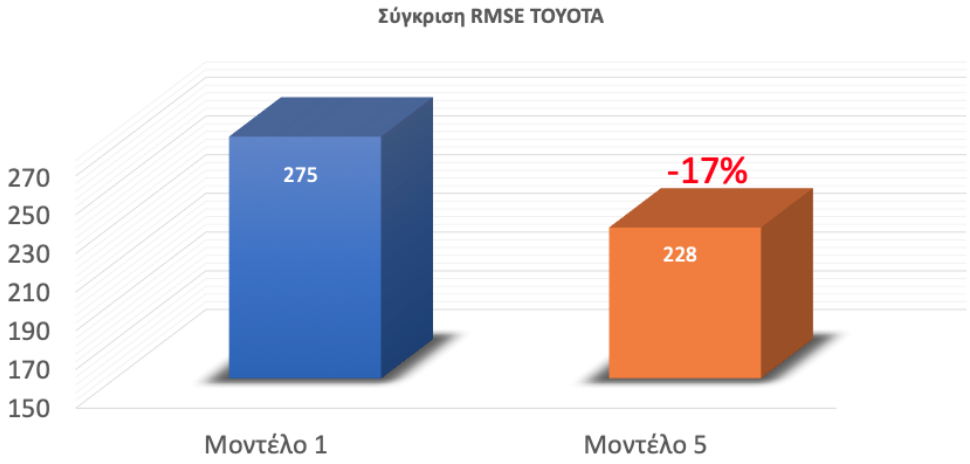
Στην περίπτωση της μάρκας αυτοκινήτων **TOYOTA**

- Γενικά οι προβλέψεις όλων των μοντέλων κρίνονται ικανοποιητικές αφού το MAPE κυμαίνεται σε χαμηλά επίπεδα αγγίζοντας ως και το 20%. Στο παρακάτω **Σχήμα 6–1** φαίνεται πώς το καλύτερο σε απόδοση μοντέλο προσεγγίζει τις πραγματικές τιμές πωλήσεων στο σύνολο ελέγχου (test set).



Σχήμα 6–1: Διάγραμμα τιμών προβλέψεων και πραγματικών παρατηρήσεων 09-2018 ως 07-2021 TOYOTA

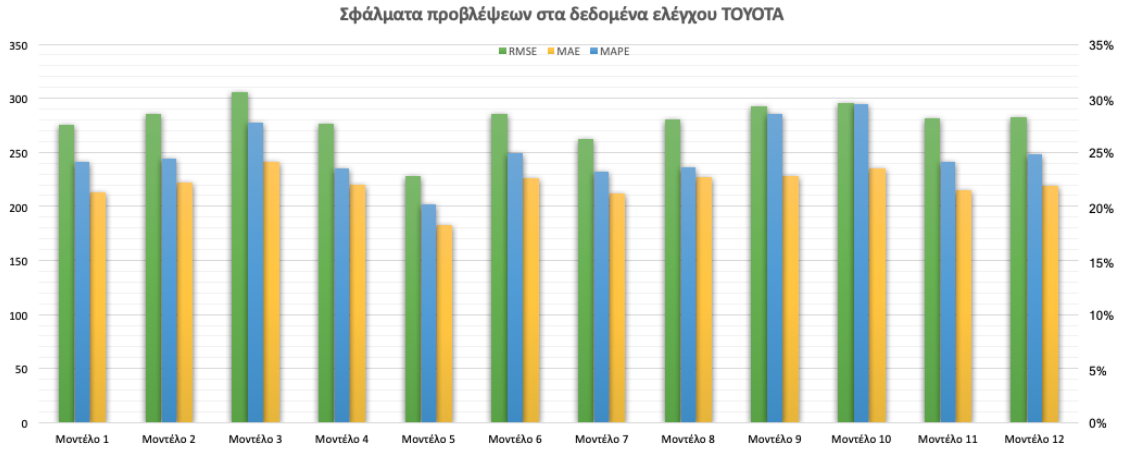
– Την καλύτερη απόδοση την είχε το Μοντέλο 5, μοντέλο το οποίο περιείχε δεδομένα Google Trends. Ξεπέρασε κατά πολύ το αμέσως επόμενο μοντέλο χωρίς Google Trends το οποίο είναι το Μοντέλο 1 που ανήκει στα μέτρα σύγκρισης. Πιο συγκεκριμένα το ξεπέρασε κατά 17% (**Σχήμα 6–2**) όσον αφορά το RMSE και αποτελεί την καλύτερη απόπειρα πρόγνωσης που έγινε σε αυτήν την εργασία.



Σχήμα 6–2: Σύγκριση των δύο καλύτερων σε απόδοση μοντέλων TOYOTA

– Ακόμα παρατηρείται ότι όλα τα μοντέλα με υποστήριξη Google Trends (Μοντέλα 5, 7 ,9, 11) ξεπέρασαν τα αντίστοιχά τους χωρίς Google Trends (Μοντέλα 6, 8, 10, 12) (**Σχήμα 6–3**)

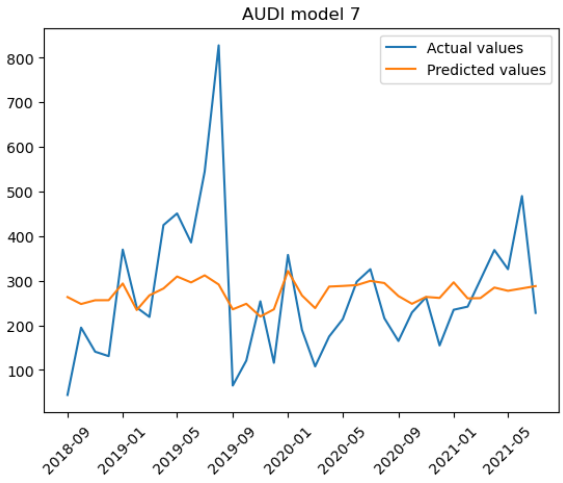
– Τέλος αν τα μοντέλα εξεταστούν σε ζεύγη με βάση τη μέθοδο πρόβλεψης που στηρίχτηκαν (5 και 6, 7 και 8, 9 και 10, 11 και 12), είναι φανερό (**Σχήμα 6–3**) πως στα μοντέλα με δέντρα απόφασης και τυχαία δάση υπήρξε μεγαλύτερη βελτίωση της απόδοσης κατά την εισαγωγή Google Trends, σε σχέση με τα υπόλοιπα. Συνεπώς εικάζεται ότι τα δεδομένα Google Trends συνδυάζονται καλύτερα με τέτοιου είδους τεχνικές προβλέψεων.



Σχήμα 6–3: Γραφική αναπαράσταση τιμών σφαλμάτων TOYOTA

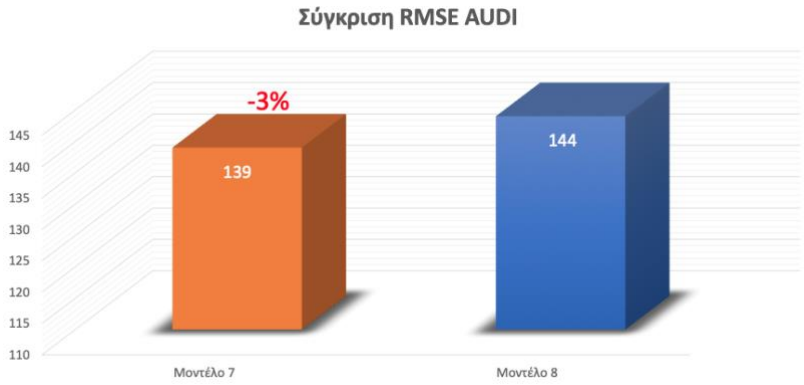
Στην περίπτωση της μάρκας αυτοκινήτων **AUDI**

– Σε γενικές γραμμές η απόδοση όλων των προβλέψεων δεν κρίνεται ικανοποιητική, το MAPE κυμαίνεται κοντά στο 60% κάποιες φορές και πολύ παραπάνω. Οπότε οι προβλέψεις μπορούν να χαρακτηριστούν και αναξιόπιστες. Στο παρακάτω **Σχήμα 6–4** φαίνεται η απόκλιση των προβλέψεων του καλύτερου μοντέλου από τις πραγματικές παρατηρήσεις.

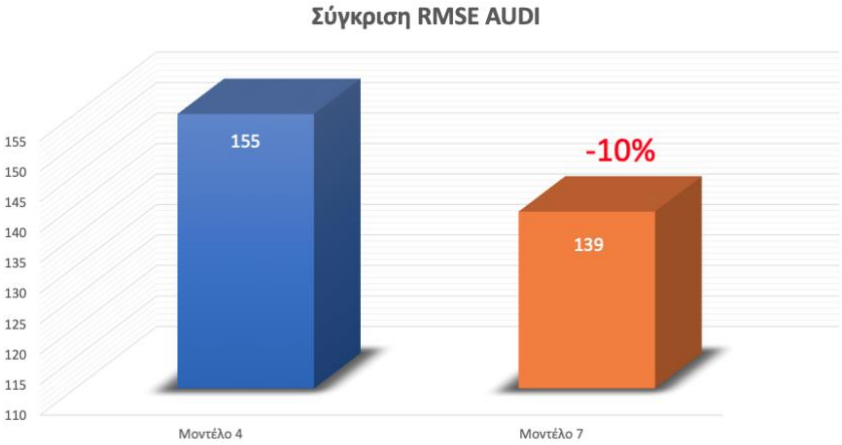


Σχήμα 6–4: Διάγραμμα τιμών προβλέψεων και πραγματικών παρατηρήσεων 09-2018 ως 07-2021 AUDI

– Όμως παρόλη την μειωμένη αποτελεσματικότητα όλων των μοντέλων, είναι φανερό πως το Μοντέλο 7 με την εισαγωγή Google Trends ξεπέρασε το επόμενο καλύτερο χωρίς Google Trends Μοντέλο 8 κατά ένα μικρό ποσοστό του 3% στο RMSE (**Σχήμα 6–5**) αλλά και το καλύτερο από τα benchmarks κατά 10%(**Σχήμα 6–6**). Συνεπώς είναι εμφανές ότι τα δεδομένα Google Trends έδωσαν χρήσιμη πληροφορία στο μοντέλο και βελτίωσαν την απόδοσή του, σε κάποιο βαθμό έστω.

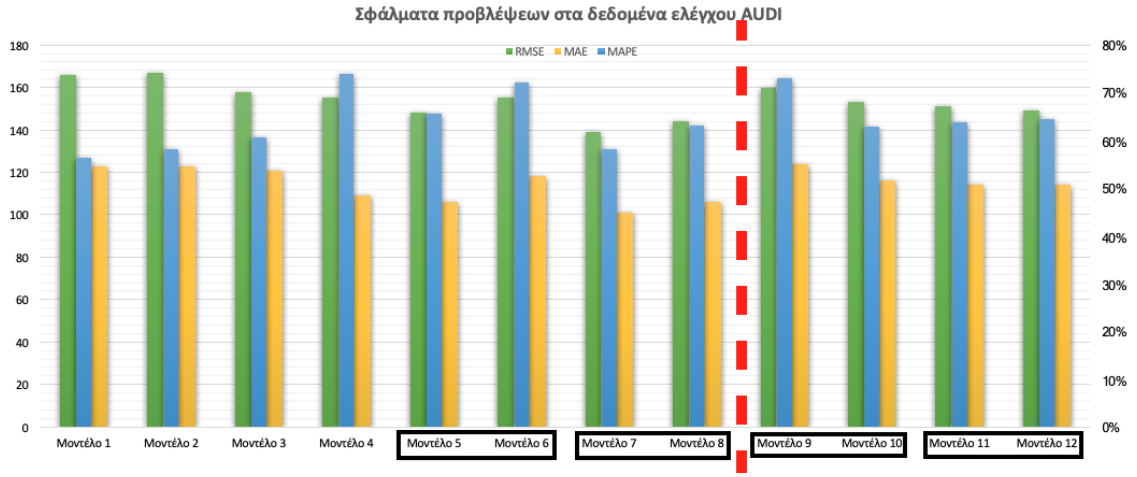


Σχήμα 6–5: Σύγκριση των δύο καλύτερων σε απόδοση μοντέλων AUDI



Σχήμα 6–6: Σύγκριση του καλύτερου μοντέλου με το καλύτερο benchmark AUDI

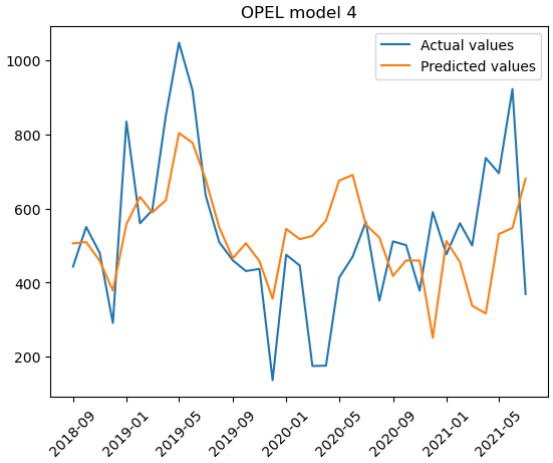
– Επίσης και σε αυτήν την περίπτωση φαίνεται ότι οι μέθοδοι των δένδρων αποφάσεων και τυχαίων δασών εκμεταλλεύτηκαν την πληροφορία των Google Trends με πιο αποδοτικό τρόπο (**Σχήμα 6–7**).



Σχήμα 6–7: Γραφική αναπαράσταση τιμών σφαλμάτων AUDI

Στην περίπτωση της μάρκας **OPEL**

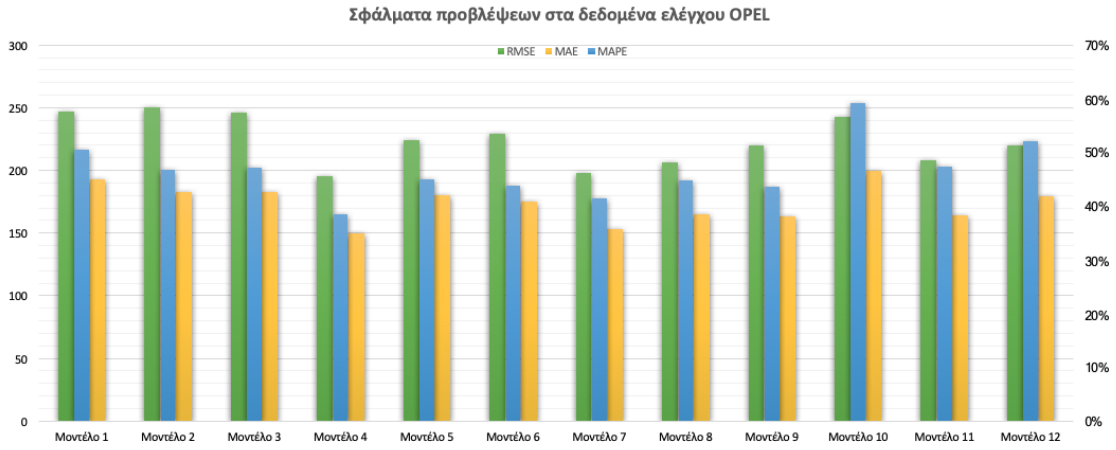
– Η γενική απόδοση των μοντέλων είναι σίγουρα καλύτερη από αυτή της AUDI αλλά όχι καλύτερη της TOYOTA, τα MAPE κυμαίνονται γύρω από το 40%. Στη συνέχεια φαίνεται η προσέγγιση του καλύτερου μοντέλου στις πραγματικές τιμές (**Σχήμα 6–8**).



Σχήμα 6–8: Διάγραμμα τιμών προβλέψεων και πραγματικών παρατηρήσεων 09-2018 ως 07-2021 OPEL

– Το Μοντέλο 7 είναι το μοντέλο Google Trends με την καλύτερη απόδοση, δεν κατάφερε όμως να ξεπεράσει το Μοντέλο 4 από τα benchmarks το οποίο βέβαια περιέχει δεδομένα πωλήσεων ενός μήνα πριν από την πρόβλεψη, οπότε η ίσως η σύγκριση δεν είναι ωφέλιμη. Κατά τ 'άλλα σημείωσε μικρή βελτίωση σε σχέση με το Μοντέλο 8, αν και ξεπέρασε κατά πολύ τα υπόλοιπα μοντέλα benchmarks με ποσοστό περίπου 20% στο RMSE.

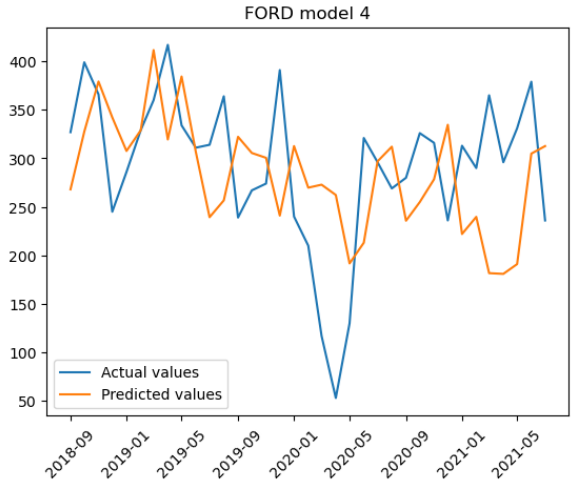
– Κάθε μοντέλο Google Trends ξεπέρασε το αντίστοιχό του χωρίς Google Trends έστω και με μικρή διαφορά όπως φαίνεται και στο **Σχήμα 6–9**.



Σχήμα 6–9: Γραφική αναπαράσταση τιμών σφαλμάτων OPEL

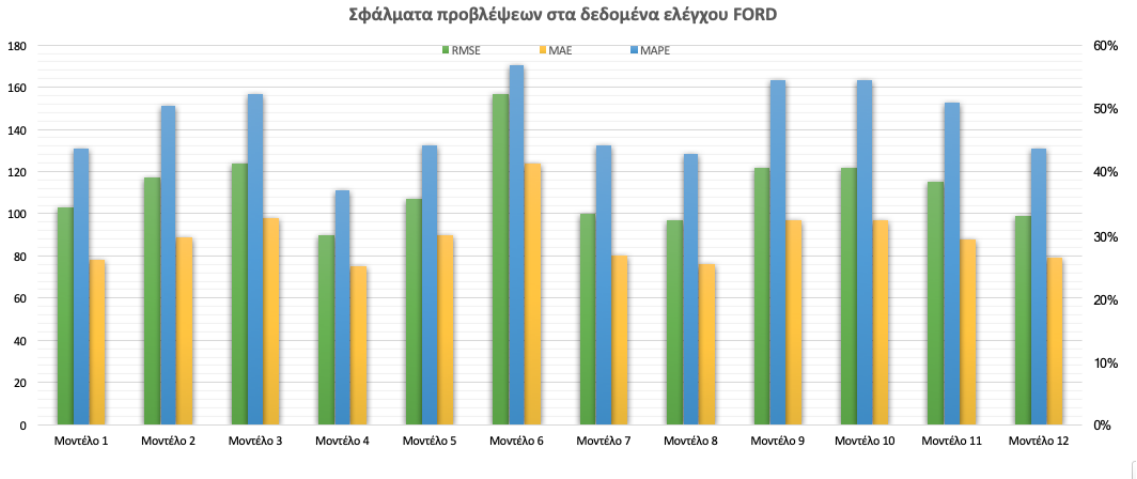
Στην περίπτωση της μάρκας **FORD**

– Σε γενικές γραμμές η απόδοση των μοντέλων δεν ήταν κάτι αξιοσημείωτο, ήταν παρόμοια με αυτή των μοντέλων της OPEL. Στη συνέχεια φαίνεται η προσέγγιση του καλύτερου μοντέλου στις πραγματικές τιμές (**Σχήμα 6–10**).



Σχήμα 6–10: Διάγραμμα τιμών προβλέψεων και πραγματικών παρατηρήσεων 09-2018 ως 07-2021 FORD

– Στη συγκεκριμένη περίπτωση τα δεδομένα Google Trends δεν φαίνεται να συνεισφέραν στη βελτίωση της απόδοσης των μοντέλων. Τα μοντέλα Google Trends σε αυτήν την περίπτωση παρήγαγαν παρόμοιες προβλέψεις με τα μοντέλα χωρίς Google Trends με αποδόσεις σχεδόν ίδιες ή και χειρότερες όπως φαίνεται και στο **Σχήμα 6–11**.



Σχήμα 6–11: Γραφική αναπαράσταση τιμών σφαλμάτων FORD

Αναλυτικά οι προβλέψεις που πραγματοποιήθηκαν στη συγκεκριμένη εργασία για την περίοδο Σεπτέμβριος 2018 ως Ιούλιος 2021 (test set δηλαδή) παρουσιάζονται σε μορφή γραφημάτων σε αντιστοιχία με τις πραγματικές παρατηρήσεις των πωλήσεων για την ίδια περίοδο στο **Παράρτημα III** στο τέλος αυτού του τεύχους.

7. Συμπεράσματα και Μελλοντικές Προεκτάσεις

7.1 Συμπεράσματα

Λαμβάνοντας υπόψιν τους στόχους της παρούσας διπλωματικής εργασίας αλλά και τα αποτελέσματα που προέκυψαν από αυτήν όπως παρουσιάζονται στο έκτο κεφάλαιο, μπορούν να εξαχθούν τα εξής συμπεράσματα:

- Είναι εφικτό να κατασκευαστούν αποτελεσματικά προγνωστικά μοντέλα με την υποστήριξη των δεδομένων Google Trends σε συνδυασμό με άλλες οικονομικές μεταβλητές και χρήση ποικίλων μεθόδων πρόβλεψης. Πιο συγκεκριμένα στην περίπτωση της μάρκας ΤΟΥΟΤΑ, το καλύτερο μοντέλο κατασκευάστηκε με την τεχνική των δένδρων αποφάσεων και περιλάμβανε δεδομένα Google Trends. Αυτό το μοντέλο κατάφερε ξεπεράσει τις αποδόσεις των υπολοίπων, πετυχαίνοντας μείωση των σφαλμάτων κατά 17% στο RMSE, 16% στο MAPE και 14% στο MAE συγκριτικά με τις τιμές σφαλμάτων του επόμενου καλύτερου μοντέλου χωρίς Google Trends. Αυτό αποτελεί μια πολλά υποσχόμενη ένδειξη για την προγνωστική ισχύ των δεδομένων Google Trends και τη συνεισφορά τους στα προγνωστικά μοντέλα.
- Στις περισσότερες περιπτώσεις αν και δεν κατάφεραν να φτάσουν τα αποτελέσματα της περίπτωσης της ΤΟΥΟΤΑ, τα δεδομένα Google Trends επέφεραν μια μικρή βελτίωση έστω στην απόδοσή των μοντέλων σχετικά με τα υπόλοιπα. Αυτό δείχνει ότι τα δεδομένα Google Trends τελικά περιέχουν χρήσιμη πληροφορία σχετικά με το ενδιαφέρον των καταναλωτών που μπορεί να αποτυπωθεί στις πωλήσεις.
- Υπήρχαν και κάποιες περιπτώσεις που η εισαγωγή των δεδομένων Google Trends δεν επέδρασε θετικά στη βελτίωση της απόδοσης των μοντέλων. Συνεπώς εικάζεται ότι η δυνατότητα βελτίωσης της ακρίβειας των προβλέψεων με εισαγωγή Google Trends, εξαρτάται από τα χαρακτηριστικά της μάρκας. Για παράδειγμα στην περίπτωση της ΤΟΥΟΤΑ που επετεύχθη και δημιουργήθηκε το πιο αποτελεσματικό μοντέλο πρόβλεψης — αν και όλα τα μοντέλα απέδωσαν καλύτερα για αυτήν τη μάρκα— ενδεχομένως να παίζει ρόλο ότι η ΤΟΥΟΤΑ είναι ή πρώτη σε πωλήσεις αυτοκινήτων στην Ελλάδα, με τους αριθμούς των οχημάτων που πουλάει να ξεπερνούν κατά πολύ τις άλλες μάρκες.
- Στις περισσότερες περιπτώσεις παρατηρήθηκε ότι ο συνδυασμός δεδομένων Google Trends και μεθόδων πρόβλεψης βασισμένα σε δένδρα αποφάσεων ή τυχαία δάση είχε καλύτερα αποτελέσματα αναφορικά με την απόδοση των μοντέλων, σε σχέση με τις υπόλοιπες τεχνικές προβλέψεων. Έτσι συμπεραίνεται ότι η προγνωστική ισχύς των δεδομένων Google Trends αξιοποιείται καλύτερα με τη χρήση προηγμένων μεθόδων πρόβλεψης όπως αυτές που στηρίζονται στη χρήση δένδρων αποφάσεων και άλλες λογικές προσεγγίσεις. Στην παρούσα εργασία υπολογίζεται ότι η μέση βελτίωση απόδοσης που επετεύχθη κατά την εισαγωγή δεδομένων Google Trends σε μοντέλα με δένδρα αποφάσεων και τυχαία δάση είναι

της τάξης του 8% ενώ στα άλλα κυρίως γραμμικά μοντέλα είναι 1%. Αυτό ισχυροποιεί το παρόν συμπέρασμα.

7.2 Μελλοντικές προεκτάσεις

Κατά την εκπόνηση της παρούσας διπλωματικής εργασίας προέκυψαν κάποια ενδιαφέροντα θέματα για περαιτέρω διερεύνηση που ενδεχομένως θα μπορούσαν να βελτιώσουν ακόμα περισσότερο την ακρίβεια των προβλέψεων. Τα σημεία που θα μπορούσαν να επικεντρωθούν μελλοντικές μελέτες είναι τα εξής:

- Χρήση πιο εξελιγμένων τεχνικών πρόβλεψης όπως τα Τεχνητά Νευρωνικά Δίκτυα, ίσως μπορούν να αξιοποιήσουν ακόμα καλύτερα την προγνωστική ισχύ των δεδομένων Google Trends.
- Εξαντλητική διερεύνηση για τη ρύθμιση υπερπαραμέτρων του μοντέλου με σκοπό τη βελτιστοποίηση της ακρίβειας των προβλέψεων.
- Διερεύνηση αν τα χαρακτηριστικά κάθε μάρκας μπορούν να επηρεάσουν την αποτελεσματικότητα του μοντέλου, και αν γίνεται να προστεθούν ως επιπλέον πληροφορία σε αυτό με σκοπό τη βελτίωση της ακρίβειας.
- Κατασκευή γενικού μοντέλου που να εκπαιδεύεται με δεδομένα από όλες τις μάρκες συνολικά και να εκτελεί επιμέρους προβλέψεις.
- Συλλογή διαδικτυακών δεδομένων και από άλλες πλατφόρμες (αριθμός αναφορών στο Twitter).
- Αναζήτηση για τη βέλτιστη τιμή του χρονικού ορίζοντα που να πετυχαίνει την καλύτερη απόδοση.
- Διαφορετικός τρόπος συλλογής δεδομένων από το Google Trends, ένα χρονοσημείο για κάθε βήμα πρόβλεψης, επιλογή διαφορετικού χρονοπαραθύρου.
- Με την πάροδο του χρόνου όλο και περισσότεροι άνθρωποι, συνεπώς τα δεδομένα Google Trends γίνονται πιο αντιπροσωπευτικά του πληθυσμού. Άρα προτείνεται χρήση πιο πρόσφατων δεδομένων (για εκπαίδευση κλπ.) στα μοντέλα.
- Βελτιστοποίηση διαδικασίας διανομής με τη χρήση των χωρικών δεδομένων της πλατφόρμας Google Trends.
- Προσπάθεια διαχωρισμού δεδομένων Google Trends που αναφέρονται σε πωλήσεις καινούργιων αυτοκινήτων και σε μεταχειρισμένων.

Κατάλογος Αναφορών

1. ACEA . (2018). *Share of direct automotive employment in the EU, by country*. Ανάκτηση από acea.be: https://www.acea.be/statistics/tag/category/share-of-direct-automotive-employment-in-eu-by-country
2. ACEA. (2020). *The Automobile Industry Pocket Guide.* Ανάκτηση από acea.be: https://www.acea.be/uploads/publications/ACEA_Pocket_Guide_2020-2021.pdf
3. Armstrong, J. (2002). *Principles of Forecasting.* New York, USA: Kluwer Academic Publishers.
4. Askitas, N., & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *IZA Discussion Papers, 4201*.
5. Bangwayo-Skeete, P. F., & Skeete, R. W. (2014). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management, 46*, 454-464.
6. Box, G. E., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: forecasting and control.* Upper Saddle River, New Jersey: Prentice-Hall.
7. Broucke, S. v. (2016, Apr 14). *Forecasting with Google Trends*. Ανάκτηση 2020, από Medium DataMiningApps Articles: https://medium.com/dataminingapps-articles/forecasting-with-google-trends-114ab741bda4
8. Choi, H., & Varian, H. (2012, June 27). Predicting the Present with Google Trends. *THE ECONOMIC RECORD, 88*.
9. Chopra, S., & Meindl, P. (2015). *Supply Chain Management.* Pearson Publications.
10. ELSTAT. (2020, Νοέμβριος 10). ΕΡΕΥΝΑ ΧΡΗΣΗΣ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΗΣΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΣ ΑΠΟ ΝΟΙΚΟΚΥΡΙΑ ΚΑΙ ΑΤΟΜΑ – ΧΡΗΣΗ ΗΛΕΚΤΡΟΝΙΚΟΥ ΕΜΠΟΡΙΟΥ – ΑΠΟΡΡΗΤΟ ΚΑΙ ΠΡΟΣΤΑΣΙΑ ΠΡΟΣΩΠΙΚΩΝ ΔΕΔΟΜΕΝΩΝ : Έτος 2020. Πειραιάς, Αττική, Ελλάδα.
11. Fantazzini, D., & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics, 170*(Part A), 97-135.
12. FSU NTUA. (2019). *Διαφάνειες Μαθήματος.* Ανάκτηση Sep 5, 2020, από Forecasting and Strategic Unit: https://www.fsu.gr/el/component/jdownloads/finish/6/1271
13. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009, February 19). Detecting influenza epidemics using search engine query data. *nature, 457*.

14. Google . (2020). *FAQ about Google Trends data*. Ανάκτηση Sep 7, 2020, από support.google.com: https://support.google.com/trends/answer/4365533?hl=el&ref_topic=6248052

15. Hyndman, R. J., & Athanasopoulos, G. (2020, August 19). *Forecasting: principles and practice.* Ανάκτηση Sep 5, 2020, από OTexts: https://otexts.com/fpp2/

16. icrossing. (2015). *Search engine infographic 2015: The countries that stand between Google and total world domination*. Ανάκτηση Sep 6, 2020, από icrossing: https://www.icrossing.com/uk/ideas/search-engine-infographic-2015-countries-stand-between-google-and-total-world-domination

17. internet live stats. (2020). *internet live stats*. Ανάκτηση Sep 6, 2020, από internet live stats: https://www.internetlivestats.com

18. Kim, D., Woo, J., Shin, J., Lee, J., & Kim, Y. (2019). Can search engine data improve accuracy of demand forecasting for new products? Evidence from automotive market. *Industrial Management & Data Systems, 119*(5), 1089-1103.

19. Larose, C. D., & Larose, D. T. (2019). *Data science using Python and R.* Hoboken: Wiley.

20. Lui, C., Metaxas, P. T., & Mustafaraj, E. (2011). On the predictability of the U.S. elections through search volume activity. *IADIS International Conferences*.

21. Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting Methods and Applications.* John Wiley & Sons, Inc.

22. Mavragani, A., & Tsagarakis, K. P. (2019). Predicting referendum results in the Big Data Era. *Journal of Big Data, 6*(3).

23. Preis, T., Moat, H., & Stanley, H. (2013, April 25). Quantifying Trading Behavior in Financial Markets Using Google Trends. *Sci Rep 3*(1684).

24. Rogers, S. (2016, Jul 1). *What is Google Trends data — and what does it mean?* Ανάκτηση 2020, από Medium Google News Lab: https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8

25. statcounter. (2020). *Search Engine Market Share Worldwide*. Ανάκτηση Sep 6, 2020, από statcounter GlobalStats: https://gs.statcounter.com/search-engine-market-share

26. statista. (2020). *Global digital population as of July 2020*. Ανάκτηση Jul 6, 2020, από statista.com: https://www.statista.com/statistics/617136/digital-population-worldwide/

27. Wachter, P., Widmer, T., & Klein, A. (2019). Predicting Automotive Sales using Pre-Purchase Online Search Data. *Proceedings of the Federated Conference on Computer Science and Information Systems, 18*, 569–577.

28. Wijnhoven, F., & Plant, O. (2017). Sentiment Analysis and Google Trends Data for Predicting Car Sales. *38th International Conference on Information Systems*, (σ. 1).

29. Wikipedia. (2020). *Google Search*. Ανάκτηση Sep 6, 2020, από Wikipedia: https://en.wikipedia.org/wiki/Google_Search

30. Wikipedia. (2020). *Google Trends* . Ανάκτηση Sep 6, 2020, από Wikipedia: https://en.wikipedia.org/wiki/Google_Trends

31. Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine LEarning Tools and Techniques.* San Francisco, CA: Elsevier.

32. Wolber, A. (2017, May 31). *5 ways to use Google Trends for work*. Ανάκτηση 2020, από TechRepublic: https://www.techrepublic.com/article/5-ways-to-use-google-trends-for-work/

33. Woo, J., & Owen, A. (2019). Forecasting private consumption with Google Trends data. *Journal of Forecasting, 38*(2), 81-91.

34. Γεωργούλη, Κ. (2015). *Τεχνητή Νοημοσύνη.* Ζωγράφου: ΣΕΑΒ Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών.

35. Μαλινδρέτος, Γ. Π. (2008). *Σχεδιασμός και Οργάνωση της Εφοδιαστικής Αλυσίδας Αυτοκινήτων και Ανταλλακτικών στην Ελλάδα. Πρακτικές Καναλιών Διανομής, Διαχείρισης και Πληροφοριακής Υποστήριξης Διαδικασιών Logistics.* Χαροκόπειο Πανεπιστήμιο Αθηνών. ΕΣΔΟ.

Παράρτημα Ι

Στο συγκεκριμένο Παράρτημα παρατίθεται ολόκληρο το αρχικό σετ δεδομένων που συλλέχθηκε για τις ανάγκες της παρούσας διπλωματικής εργασίας. Τα ονόματα των στηλών αντιστοιχούν στις μεταβλητές των μοντέλων όπως περιγράφεται στην ενότητα 5.2.

#	date	TOYOTA	AUDI	OPEL	FORD	gt1_toyota	gt1_audi	gt1_opel	gt1_ford	gt2_toyota	gt2_audi	gt2_opel	gt2_ford	unemployment	Qgdp_Maverage	gas_wt	gas_wot	gd_at	gt_crisisTh	gt_memTh	gt_debtTh	gt_cuts	gt_fir
1	2004-01	3938	654	1753	2083	31	69	0	100	44	48	16	10	12%	14959	737	324	2433	55	27	0	0	0
2	2004-02	2253	383	1396	1397	77	73	99	96	9	27	51	29	11%	14959	742	328	2452	0	0	0	0	0
3	2004-03	2789	526	1876	1802	77	49	33	88	33	17	60	12	11%	14959	771	352	2371	0	0	0	0	0
4	2004-04	3105	564	1656	1461	52	30	95	20	25	39	97	28	11%	15868	791	368	2518	0	0	0	0	0
5	2004-05	2318	473	2097	1124	67	64	65	90	36	50	78	41	10%	15868	837	407	2424	0	0	0	0	0
6	2004-06	2261	628	3143	1720	50	32	43	98	36	62	78	16	10%	15868	809	384	2349	0	0	0	0	0
7	2004-07	1858	504	2955	2012	84	35	84	45	60	29	72	34	10%	16938	817	390	2319	0	0	0	0	0
8	2004-08	1728	486	1961	1169	28	62	59	72	35	54	57	32	10%	16938	831	402	2314	0	0	0	0	0
9	2004-09	1349	525	1792	1219	37	56	38	64	58	20	45	36	10%	16938	827	399	2328	0	0	0	0	0
10	2004-10	1834	343	2052	1627	79	63	60	39	79	46	82	52	10%	16806	855	422	2489	0	0	0	0	0
11	2004-11	2222	537	1752	1477	51	68	65	49	67	36	78	60	11%	16806	826	398	2655	0	0	0	0	0
12	2004-12	786	357	1152	893	23	90	58	92	74	62	35	14	11%	16806	773	354	2786	0	0	0	0	0
13	2005-01	2808	571	2465	2181	62	56	100	72	76	75	66	80	11%	15242	769	350	2920	0	0	0	0	0
14	2005-02	2037	490	2053	1127	33	48	32	89	92	97	32	58	10%	15242	806	381	3145	0	0	0	0	0
15	2005-03	2443	623	2141	2124	52	49	67	77	48	70	53	76	11%	15242	817	391	2855	0	0	0	0	0
16	2005-04	1603	536	1999	2114	40	27	41	67	70	45	37	60	10%	16392	878	437	2868	0	10	0	0	0
17	2005-05	1608	508	2259	1856	51	74	65	52	61	61	83	21	10%	16392	867	427	2960	0	0	0	0	0
18	2005-06	1864	586	2129	2054	53	79	51	58	80	62	31	48	10%	16392	885	441	3061	0	0	53	0	0
19	2005-07	2087	520	1996	2002	57	48	39	45	81	74	57	43	10%	17250	932	480	3272	0	0	0	0	0
20	2005-08	1684	426	1288	1305	39	87	72	41	74	90	71	68	10%	17250	947	492	3231	0	0	0	0	0
21	2005-09	1687	450	1249	1118	50	80	57	70	76	78	60	57	10%	17250	1010	544	3382	15	0	0	0	0
22	2005-10	1805	462	1590	1526	55	58	78	63	74	69	82	48	10%	17529	998	534	3307	0	0	0	0	0
23	2005-11	1616	430	1454	1543	50	66	61	69	78	57	48	18	10%	17529	902	455	3442	0	0	0	0	0
24	2005-12	719	553	892	689	37	42	62	79	100	89	44	56	10%	17529	898	451	3664	0	6	0	0	0
25	2006-01	2772	480	2081	1848	62	72	64	67	93	100	60	68	10%	16698	936	482	3978	0	11	0	0	0
26	2006-02	2179	408	1444	1305	49	63	67	68	42	41	48	48	11%	16698	925	473	4203	0	0	0	0	0
27	2006-03	2172	486	1840	1839	69	68	51	78	48	37	62	27	9%	16698	944	489	4122	0	0	0	0	0
28	2006-04	2251	439	1394	2250	75	68	92	57	58	45	86	29	9%	17997	992	528	4140	0	0	0	0	0
29	2006-05	2497	511	1870	2162	41	60	77	56	53	86	34	40	9%	17997	1033	563	3753	0	0	0	0	0
30	2006-06	2801	510	1748	2002	70	56	47	44	50	31	70	21	8%	17997	1025	556	3694	15	0	0	0	0
31	2006-07	2638	546	1675	2063	67	59	76	78	46	55	37	46	8%	18641	1081	589	3748	0	0	0	0	0
32	2006-08	1889	467	1084	1503	69	74	55	53	31	42	71	45	9%	18641	1082	586	3869	0	0	0	0	0
33	2006-09	1781	490	1298	1141	39	92	77	56	63	66	46	47	8%	18641	962	488	3931	0	0	0	0	0
34	2006-10	1954	397	1366	1641	69	69	83	53	70	69	100	36	8%	19284	906	441	4129	0	0	33	0	0
35	2006-11	1759	565	1516	1176	60	60	67	60	65	50	48	55	9%	19284	909	443	4221	10	0	31	0	0
36	2006-12	821	910	817	1425	74	57	54	48	69	51	49	55	10%	19284	911	445	4394	0	0	0	0	42
37	2007-01	3770	411	2869	1744	62	77	63	78	54	82	58	54	9%	17610	902	420	4710	0	4	0	0	0
38	2007-02	2216	405	1755	1280	72	52	82	65	71	49	93	49	9%	17610	905	422	4504	0	0	0	0	0
39	2007-03	2772	570	2040	1970	73	71	83	51	77	67	70	37	10%	17610	960	468	4643	0	0	23	0	0
40	2007-04	2348	438	1965	1550	70	58	62	62	42	49	67	50	9%	19453	1006	506	4737	0	4	0	0	0
41	2007-05	2569	578	2344	2448	69	69	57	59	86	60	51	53	8%	19453	1054	545	4972	7	0	0	0	0
42	2007-06	2590	658	2504	2031	46	38	58	59	61	63	78	69	8%	19453	1059	549	4844	14	0	0	0	0
43	2007-07	2905	611	2556	1917	59	74	56	73	75	65	77	73	8%	19976	1053	545	4918	0	7	0	0	0
44	2007-08	2124	488	1600	1525	63	69	88	61	47	53	77	46	8%	19976	1023	520	4913	7	0	0	0	0
45	2007-09	1520	458	1639	1353	61	87	40	52	51	82	58	50	9%	19976	1020	518	5123	0	0	0	0	0

#	date	TOYOTA	AUDI	OPEL	FORD	gt1_toyota	gt1_audi	gt1_opel	gt1_ford	gt2_toyota	gt2_audi	gt2_opel	gt2_ford	unemployment	Qgdp_Maverage	gas_wt	gas_wot	gd_at	gt_crisisTh	gt_memTh	gt_debtTh	gt_cuts	gt_fir
46	2007-10	2272	581	1694	1318	59	86	58	51	67	71	69	61	8%	20525	1023	520	5335	0	0	0	0	0
47	2007-11	1851	498	1621	1005	56	55	57	64	66	74	51	52	8%	20525	1072	560	5054	5	2	47	0	21
48	2007-12	762	479	738	916	61	81	46	50	64	90	58	40	9%	20525	1069	558	5179	0	3	0	0	0
49	2008-01	3598	573	2836	1879	77	86	57	76	42	67	57	59	8%	18626	1099	564	4363	15	0	0	0	0
50	2008-02	2425	482	1755	990	73	96	48	71	60	90	61	60	8%	18626	1089	556	4133	5	2	0	0	0
51	2008-03	2131	558	1646	1312	61	77	61	60	48	73	53	57	9%	18626	1118	580	3986	5	2	15	42	0
52	2008-04	2457	712	2199	1701	68	100	66	78	61	82	79	62	8%	20248	1128	588	4214	0	3	32	0	0
53	2008-05	2565	623	2087	1418	57	69	56	55	60	64	64	52	7%	20248	1199	647	4177	0	0	15	0	0
54	2008-06	2241	686	2313	1389	67	60	43	46	50	66	53	36	8%	20248	1256	693	3440	0	0	0	0	0
55	2008-07	3063	611	2477	1674	59	65	46	45	61	49	51	61	7%	21026	1264	701	3395	0	0	0	44	42
56	2008-08	1987	647	1436	1173	53	82	59	59	45	67	42	37	7%	21026	1197	645	3293	0	0	0	0	0
57	2008-09	1720	511	1223	1022	64	82	54	54	64	70	50	47	8%	21026	1182	633	2856	5	2	0	0	18
58	2008-10	2167	613	1497	1154	55	74	60	85	48	61	47	56	7%	20763	1084	552	2060	100	0	0	0	0
59	2008-11	1448	545	837	1109	60	80	50	56	51	39	54	41	8%	20763	951	442	1914	92	0	0	0	56
60	2008-12	670	329	564	931	62	65	46	40	41	51	39	35	9%	20763	857	364	1787	35	2	60	0	0
61	2009-01	2058	676	1752	1157	83	68	69	73	49	43	83	43	10%	17795	839	340	1779	38	0	14	0	0
62	2009-02	1718	480	1155	1387	100	54	64	70	75	31	71	49	9%	17795	884	377	1536	62	2	0	0	62
63	2009-03	1582	502	1038	1269	63	60	62	78	54	44	64	55	10%	17795	884	377	1684	41	0	14	0	56
64	2009-04	1671	654	1085	1053	98	78	77	73	50	49	55	41	10%	20073	922	409	2054	30	0	0	0	0
65	2009-05	2446	923	1634	1599	61	68	64	61	58	45	73	33	9%	20073	950	431	2327	22	0	26	0	0
66	2009-06	2072	772	1673	1444	70	67	39	54	59	38	38	23	9%	20073	1021	480	2210	13	2	13	0	18
67	2009-07	3426	1543	2798	2084	79	76	44	60	38	56	47	49	10%	20418	1060	472	2362	19	0	28	0	0
68	2009-08	1581	1095	2409	1279	76	74	50	53	44	36	54	41	9%	20418	1099	504	2466	11	0	16	0	0
69	2009-09	551	407	1380	714	64	68	54	66	50	47	47	31	9%	20418	1071	482	2661	12	4	47	0	0
70	2009-10	1177	292	797	887	79	71	58	61	46	41	68	41	10%	20891	1047	462	2686	22	0	11	0	15
71	2009-11	2026	356	1153	1337	56	56	48	46	39	33	45	34	11%	20891	1082	490	2263	14	0	11	0	14
72	2009-12	699	204	600	1132	58	60	42	62	36	38	36	28	11%	20891	1071	481	2196	22	2	22	0	14
73	2010-01	3884	545	2084	1806	50	55	55	47	40	31	48	46	12%	18043	1100	505	2048	23	2	19	28	13
74	2010-02	1857	268	1330	942	65	56	44	45	50	40	53	50	12%	18043	1179	511	1913	17	0	81	0	0
75	2010-03	2541	515	1634	1698	58	49	36	46	40	32	32	58	12%	18043	1379	546	2067	33	0	27	100	48
76	2010-04	1539	315	1071	800	45	42	38	41	39	25	43	64	12%	18859	1446	574	1870	46	0	100	55	67
77	2010-05	1337	209	968	943	38	36	30	37	26	31	30	41	12%	18859	1503	573	1551	38	31	19	80	39
78	2010-06	1541	226	1378	1091	39	37	31	41	27	29	27	40	12%	18859	1492	552	1434	23	45	10	27	26
79	2010-07	1106	211	1168	619	51	47	36	29	33	19	34	37	12%	18829	1508	545	1682	27	16	20	0	27
80	2010-08	693	206	686	440	41	40	38	39	18	35	32	39	13%	18829	1502	540	1555	15	18	22	0	30
81	2010-09	524	99	530	601	38	47	34	33	24	31	39	43	13%	18829	1493	533	1471	9	22	17	24	35
82	2010-10	782	220	737	458	29	42	36	37	21	20	30	42	14%	18977	1495	534	1547	14	20	25	24	34
83	2010-11	622	173	1033	552	32	34	33	38	26	27	34	26	14%	18977	1513	549	1420	41	34	33	0	33
84	2010-12	290	154	384	216	50	35	36	33	27	24	28	42	15%	18977	1564	590	1414	17	22	17	24	23
85	2011-01	1417	175	1251	555	44	42	24	36	61	63	23	85	16%	16254	1614	630	1593	21	10	37	19	9
86	2011-02	1120	154	677	422	41	32	29	40	62	55	32	100	16%	16254	1619	634	1577	14	20	21	0	27
87	2011-03	995	226	847	894	42	31	32	34	62	43	33	71	17%	16254	1667	672	1535	19	9	25	26	34
88	2011-04	943	174	1068	763	33	33	29	34	54	46	33	89	16%	17303	1690	691	1435	24	14	21	48	28
89	2011-05	842	284	1211	560	33	31	32	32	58	54	25	81	17%	17303	1701	700	1309	16	15	37	26	21
90	2011-06	783	320	1159	530	36	31	26	27	64	53	22	70	16%	17303	1675	679	1279	14	39	32	9	13
91	2011-07	1094	235	1107	603	40	31	26	35	60	44	26	63	17%	17357	1679	682	1204	14	14	14	19	14
92	2011-08	1238	174	956	478	44	32	31	35	50	48	32	60	19%	17357	1690	691	916	13	8	21	10	14
93	2011-09	884	198	904	416	37	27	28	28	64	50	30	57	18%	17357	1692	692	798	20	10	15	17	100
94	2011-10	598	201	486	359	37	27	28	23	55	44	34	50	19%	16855	1675	679	809	25	4	37	24	30
95	2011-11	934	249	1218	417	36	30	30	29	51	40	26	60	22%	16855	1657	664	682	27	10	26	16	27
96	2011-12	1278	123	556	430	37	25	34	29	61	43	32	55	22%	16855	1640	651	680	18	6	6	32	35
97	2012-01	631	233	1374	409	39	28	38	28	78	41	36	59	22%	14929	1703	700	796	23	15	15	7	41
98	2012-02	438	106	452	223	39	30	28	28	69	44	26	56	23%	14929	1739	730	744	31	100	21	15	57

#	date	TOYOTA	AUDI	OPEL	FORD	gt1_toyota	gt1_audi	gt1_opel	gt1_ford	gt2_toyota	gt2_audi	gt2_opel	gt2_ford	unemployment	Qgdp_Maverage	gas_wt	gas_wot	gd_at	gt_crisisTh	gt_memTh	gt_debtTh	gt_cuts	gt_fir
99	2012-03	568	176	683	339	33	25	24	25	62	34	26	50	23%	14929	1795	774	729	19	30	15	29	7
100	2012-04	396	175	531	290	24	26	26	26	62	38	23	52	23%	15861	1844	814	700	11	18	8	0	19
101	2012-05	464	178	540	276	27	23	21	27	51	32	19	51	24%	15861	1756	744	525	16	34	37	7	17
102	2012-06	543	170	489	258	27	27	19	22	47	35	23	49	25%	15861	1706	703	611	11	30	8	0	11
103	2012-07	482	199	512	264	34	26	28	27	53	35	25	38	24%	16103	1706	703	599	15	6	20	8	22
104	2012-08	529	117	489	229	33	31	24	30	51	42	30	47	25%	16103	1794	773	647	12	6	12	48	15
105	2012-09	394	122	373	217	35	29	30	25	60	45	25	47	26%	16103	1821	795	739	13	9	23	21	10
106	2012-10	593	136	607	148	29	22	22	26	52	44	23	46	26%	15904	1758	745	801	21	13	32	39	12
107	2012-11	472	130	549	209	32	26	23	27	63	38	23	38	27%	15904	1685	686	809	12	42	16	26	75
108	2012-12	234	68	467	144	26	24	22	24	58	42	18	43	26%	15904	1682	684	908	12	11	12	20	19
109	2013-01	685	179	803	208	35	27	25	29	68	45	23	41	28%	14138	1700	698	987	11	11	8	24	11
110	2013-02	359	104	407	161	30	25	23	23	70	33	22	42	28%	14138	1752	740	1008	15	7	9	6	24
111	2013-03	462	133	433	143	27	27	32	26	64	38	33	46	27%	14138	1746	735	869	22	8	13	0	11
112	2013-04	697	163	421	180	28	23	20	22	52	34	22	43	28%	15154	1692	692	974	11	6	8	0	71
113	2013-05	642	175	408	211	24	26	22	25	52	37	19	40	28%	15154	1663	668	1015	16	5	8	0	23
114	2013-06	476	163	441	259	26	22	27	24	52	41	21	42	27%	15154	1688	688	848	6	5	2	0	39
115	2013-07	660	156	583	322	28	26	25	26	55	39	23	33	27%	15489	1701	699	885	7	6	8	6	88
116	2013-08	427	123	312	199	28	35	20	28	44	48	20	44	28%	15489	1712	708	900	4	4	11	0	44
117	2013-09	345	145	456	243	27	26	19	24	59	51	17	42	27%	15489	1703	701	1014	9	4	8	11	32
118	2013-10	660	197	418	321	28	33	22	26	54	50	18	44	27%	15091	1656	663	1188	9	2	14	0	31
119	2013-11	641	186	567	369	28	26	23	28	56	44	25	42	29%	15091	1636	647	1196	14	6	18	5	26
120	2013-12	476	139	320	227	29	27	21	28	48	56	23	43	28%	15091	1649	657	1163	7	4	4	0	24
121	2014-01	737	293	672	414	31	27	24	29	58	44	23	45	27%	13744	1650	658	1177	13	3	7	5	2
122	2014-02	539	157	427	203	28	25	20	25	48	46	20	42	28%	13744	1651	659	1310	14	4	2	6	11
123	2014-03	582	168	496	366	26	23	26	25	58	39	24	43	28%	13744	1663	669	1336	6	6	7	0	14
124	2014-04	770	174	537	303	26	25	22	25	50	44	19	40	27%	14723	1672	676	1232	10	5	8	0	3
125	2014-05	743	342	551	362	22	24	22	22	48	44	20	48	27%	14723	1678	680	1223	9	4	5	0	0
126	2014-06	810	227	740	341	28	24	18	23	47	38	14	33	26%	14723	1690	690	1214	7	4	4	6	16
127	2014-07	695	207	675	360	26	30	26	25	54	45	21	47	25%	15540	1707	704	1169	5	3	0	0	14
128	2014-08	427	150	318	211	31	31	24	30	47	51	24	40	25%	15540	1683	685	1162	2	1	7	6	3
129	2014-09	560	171	516	275	34	24	22	28	50	46	25	46	25%	15540	1666	671	1062	5	3	2	0	13
130	2014-10	662	217	575	301	28	28	24	31	44	43	20	44	25%	15109	1633	645	916	6	4	6	5	8
131	2014-11	774	149	411	227	29	28	24	32	42	43	22	48	26%	15109	1581	603	963	8	5	8	0	8
132	2014-12	722	178	563	319	29	22	24	29	55	38	23	40	27%	15109	1522	556	826	8	3	5	5	2
133	2015-01	870	223	553	392	31	27	26	29	49	48	20	44	26%	13664	1393	453	722	10	7	14	0	2
134	2015-02	649	159	394	268	30	29	25	32	50	51	21	47	26%	13664	1424	478	880	6	12	23	0	0
135	2015-03	629	310	410	349	32	26	26	34	54	47	25	52	27%	13664	1501	539	775	11	5	5	5	0
136	2015-04	818	370	724	349	32	27	25	28	51	45	25	51	26%	14689	1525	558	823	6	4	17	0	5
137	2015-05	942	282	701	337	30	28	22	29	56	48	19	51	24%	14689	1559	586	825	8	4	11	0	0
138	2015-06	655	225	720	242	24	25	19	25	45	45	20	36	25%	14689	1570	594	798	10	12	28	21	0
139	2015-07	641	216	313	141	27	23	20	22	43	39	16	33	25%	15162	1579	601	668	11	49	30	5	7
140	2015-08	588	130	314	263	37	31	27	31	52	46	23	41	23%	15162	1511	548	624	7	24	6	0	11
141	2015-09	456	206	404	261	35	31	26	30	54	49	24	41	25%	15162	1438	489	654	3	13	9	5	5
142	2015-10	531	147	473	297	31	26	27	31	47	38	20	38	24%	15189	1418	473	701	1	10	6	26	7
143	2015-11	982	197	555	287	29	30	23	29	40	40	25	41	24%	15189	1404	462	635	10	4	6	6	0
144	2015-12	1105	203	475	367	32	32	25	30	43	47	27	39	25%	15189	1402	460	631	8	3	4	6	0
145	2016-01	826	160	530	336	36	31	26	33	51	45	24	47	26%	13325	1360	427	553	11	6	15	29	21
146	2016-02	416	181	263	191	36	29	23	31	47	48	21	41	25%	13325	1318	394	517	8	7	4	10	24
147	2016-03	535	225	437	271	29	33	24	33	42	54	20	39	25%	13325	1318	393	577	12	6	0	0	7
148	2016-04	1102	417	979	290	29	31	25	31	46	38	23	48	24%	14581	1369	434	584	9	10	9	42	7
149	2016-05	1406	442	1273	614	32	31	26	35	48	48	24	50	23%	14581	1396	455	647	8	8	12	49	7
150	2016-06	1196	245	589	447	29	27	22	27	45	44	19	43	23%	14581	1430	474	542	7	5	11	75	12
151	2016-07	915	263	468	232	36	38	23	34	57	47	24	40	23%	15256	1418	465	571	3	2	0	26	8

#	date	TOYOTA	AUDI	OPEL	FORD	gt1_toyota	gt1_audi	gt1_opel	gt1_ford	gt2_toyota	gt2_audi	gt2_opel	gt2_ford	unemployment	Qgdp_Maverage	gas_wt	gas_wot	gd_at	gt_crisisTh	gt_memTh	gt_debtTh	gt_cuts	gt_fir
152	2016-08	699	118	344	215	37	37	24	35	53	51	26	45	23%	15256	1396	448	577	5	1	6	33	3
153	2016-09	539	144	508	179	37	31	26	33	58	49	21	40	22%	15256	1414	462	566	3	4	7	51	12
154	2016-10	707	144	458	281	34	39	28	34	48	55	24	41	22%	14917	1431	474	591	9	5	4	41	19
155	2016-11	757	188	619	391	34	33	24	38	54	47	25	55	24%	14917	1428	472	629	9	5	6	37	35
156	2016-12	398	91	586	289	47	39	27	39	52	46	27	51	24%	14917	1451	490	644	9	3	9	36	15
157	2017-01	985	310	438	246	53	37	26	40	66	46	25	46	24%	13506	1537	534	612	8	3	12	9	18
158	2017-02	644	226	479	358	50	36	28	36	67	49	25	44	24%	13506	1547	536	646	10	6	8	21	5
159	2017-03	888	254	726	397	45	36	27	35	54	56	25	49	23%	13506	1533	525	666	7	5	5	9	11
160	2017-04	831	385	599	276	40	33	24	35	50	51	20	51	21%	14767	1519	514	712	9	5	7	16	8
161	2017-05	1151	318	692	298	39	31	26	32	52	53	23	43	22%	14767	1516	512	775	6	16	16	39	9
162	2017-06	914	318	750	322	40	32	27	30	52	47	24	40	20%	14767	1485	488	824	7	5	19	16	7
163	2017-07	1031	181	734	289	49	38	24	36	54	58	29	41	20%	15839	1464	471	812	6	3	4	10	18
164	2017-08	738	173	439	253	46	37	27	39	51	53	25	51	21%	15839	1475	479	825	2	2	2	15	7
165	2017-09	680	248	514	324	48	31	28	37	54	49	26	50	20%	15839	1502	501	756	3	2	7	15	16
166	2017-10	784	262	567	368	48	34	28	38	57	49	26	50	20%	14939	1505	503	759	5	1	9	20	9
167	2017-11	793	209	403	408	43	32	28	34	53	53	26	44	22%	14939	1528	522	740	9	3	2	25	7
168	2017-12	587	201	303	364	48	37	32	40	57	47	30	52	21%	14939	1534	527	802	5	3	13	36	20
169	2018-01	1215	313	756	398	55	37	34	41	53	51	31	49	21%	13712	1543	534	879	11	3	10	19	16
170	2018-02	713	144	532	405	47	31	28	43	55	52	26	36	22%	13712	1550	539	836	6	4	7	41	12
171	2018-03	1167	206	636	529	45	30	31	36	58	69	26	48	21%	13712	1528	521	781	4	2	5	9	20
172	2018-04	1118	300	673	325	45	37	29	40	50	52	27	53	20%	14963	1557	544	858	7	1	7	10	20
173	2018-05	1054	276	837	403	46	30	27	37	57	47	22	48	19%	14963	1615	590	756	7	6	13	37	4
174	2018-06	1225	296	658	322	45	33	27	36	56	53	24	53	18%	14963	1646	615	758	4	4	6	18	9
175	2018-07	867	281	562	211	45	34	29	35	60	57	26	53	18%	16108	1639	610	761	3	2	8	13	13
176	2018-08	973	335	513	242	54	41	30	40	63	59	27	59	19%	16108	1644	613	730	3	16	21	13	6
177	2018-09	675	44	443	327	50	33	28	39	65	60	28	54	18%	16108	1644	614	692	3	3	11	31	2
178	2018-10	1073	195	550	399	56	37	34	43	55	52	25	55	18%	15126	1653	621	640	7	2	7	56	7
179	2018-11	1057	141	480	366	46	36	29	39	48	48	28	51	19%	15126	1593	574	630	9	3	5	72	9
180	2018-12	619	131	291	245	53	32	30	41	61	50	30	51	18%	15126	1519	515	613	2	3	7	20	12
181	2019-01	1149	370	834	286	61	35	37	45	70	45	30	53	20%	13850	1484	487	635	7	2	8	13	79
182	2019-02	761	240	560	329	50	38	29	44	62	55	28	50	20%	13850	1505	503	708	2	1	5	15	19
183	2019-03	829	219	595	360	48	36	33	45	57	47	31	49	18%	13850	1556	543	721	6	1	7	19	7
184	2019-04	1385	425	850	417	54	33	34	43	63	47	28	59	18%	15357	1620	594	773	7	3	5	0	12
185	2019-05	1452	451	1047	334	51	35	28	38	57	47	29	55	17%	15357	1646	615	830	4	4	9	18	25
186	2019-06	1045	386	918	311	48	40	27	36	64	53	25	55	16%	15357	1611	588	868	3	4	13	0	15
187	2019-07	1464	545	635	314	56	37	32	37	62	58	23	44	17%	16684	1616	592	900	2	3	9	8	6
188	2019-08	986	828	509	364	62	40	35	44	76	61	31	62	16%	16684	1610	587	868	3	1	6	0	29
189	2019-09	925	65	460	239	79	39	31	41	80	67	29	51	16%	16684	1599	578	868	2	2	3	8	22
190	2019-10	987	121	431	267	63	35	32	42	66	52	26	53	16%	15247	1593	573	883	2	1	12	0	14
191	2019-11	948	254	437	274	55	38	32	42	62	52	30	55	17%	15247	1588	569	902	6	1	11	0	14
192	2019-12	650	116	137	391	55	36	33	43	63	53	27	57	17%	15247	1594	574	917	4	2	2	9	49
193	2020-01	1475	358	475	240	67	36	34	53	77	51	30	65	17%	13520	1619	594	921	6	3	8	8	36
194	2020-02	956	190	446	210	61	35	33	45	61	55	26	54	17%	13520	1597	577	720	4	2	4	4	14
195	2020-03	711	108	175	117	36	26	18	29	39	32	18	43	15%	13520	1500	501	558	7	2	7	11	65
196	2020-04	457	175	176	53	31	22	17	25	33	34	13	39	16%	12710	1349	381	628	6	4	7	7	28
197	2020-05	729	214	413	130	49	33	28	39	53	40	28	52	17%	12710	1325	361	653	6	4	0	16	59
198	2020-06	1029	298	470	321	63	40	33	47	63	52	28	56	17%	12710	1387	410	639	5	3	8	8	54
199	2020-07	1237	326	564	296	68	39	39	48	73	54	35	56	16%	14864	1429	443	618	3	4	10	16	21
200	2020-08	904	216	351	269	67	44	33	47	68	56	30	57	16%	14864	1430	444	634	2	4	6	8	6
201	2020-09	879	165	511	280	72	32	33	45	76	51	29	51	15%	14864	1426	441	625	3	1	7	8	31
202	2020-10	592	229	501	326	63	36	32	43	66	51	26	50	16%	14182	1422	438	570	3	3	3	8	10
203	2020-11	663	263	378	316	44	27	18	30	49	38	17	44	17%	14182	1414	431	737	4	2	7	4	22
204	2020-12	631	155	590	236	44	28	20	36	53	43	19	47	16%	14182	1427	442	809	5	2	1	11	12

#	date	TOYOTA	AUDI	OPEL	FORD	gt1_toyota	gt1_audi	gt1_opel	gt1_ford	gt2_toyota	gt2_audi	gt2_opel	gt2_ford	unemployment	Qgdp_Maverage	gas_wt	gas_wot	gd_at	gt_crisisTh	gt_memTh	gt_debtTh	gt_cuts	gt_fir
205	2021-01	1269	235	476	313	56	37	30	44	63	49	20	49	17%	13066	1465	471	749	1	1	14	0	16
206	2021-02	932	242	560	290	60	37	27	43	62	48	24	46	17%	13066	1510	507	792	4	3	7	4	8
207	2021-03	1148	302	500	365	53	32	27	39	59	48	23	47	17%	13066	1585	567	865	3	1	8	4	13
208	2021-04	1262	369	736	296	57	33	26	40	58	40	21	50	17%		1600	579	910	3	2	6	8	17
209	2021-05	1238	326	695	331	60	32	31	42	63	44	27	47	15%		1609	586	895	3	1	3	4	6
210	2021-06	1348	490	922	379	59	38	27	42	62	49	27	41			1627	600	885	2	1	3	4	23
211	2021-07	1558	228	369	236	63	38	30	43	73	62	30	53			1676	639	888	0	1	6	12	13

Παράρτημα ΙΙ

Παρακάτω παρατίθεται όλος ο κώδικας με τον οποίο υλοποιήθηκε η διαδικασία των προβλέψεων (εκπαίδευση, διεξαγωγή προβλέψεων κλπ.) —αφού πρώτα είχε γίνει ο καθορισμός μοντέλων— για την περίοδο Σεπτέμβριος 2018 ως Αύγουστο 2021

```
#!/usr/bin/env python
# coding: utf-8
# # Predictions
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import pmdarima as pm

from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR

from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error, mean_absolute_percentage_error, mean_absolute_error

dataset=pd.read_excel('/Users/loukas/Desktop/dataset.xlsx',sheet_name=0,index_col='date')
toyota=dataset['TOYOTA']
toyota.index=pd.to_datetime(toyota.index)
toyota.index = pd.DatetimeIndex(toyota.index.values,freq=toyota.index.inferred_freq)
toyota=pd.DataFrame(toyota)
toyota["#"]=list(range(len(toyota)))
toyota_test=(toyota.iloc[176:211])
scores={}

# # TOYOTA Exponential smoothing
toyota=dataset['TOYOTA']
toyota.index=pd.to_datetime(toyota.index)
toyota.index = pd.DatetimeIndex(toyota.index.values,freq=toyota.index.inferred_freq)
toyota=pd.DataFrame(toyota)
toyota["#"]=list(range(len(toyota)))
test=(toyota.iloc[176:211])

f1=[]
for i in range(35):
    train=toyota.iloc[:165+i,0]
    fit1 = sm.tsa.ExponentialSmoothing(train,seasonal_periods=12,seasonal="add",use_boxcox=False, initialization_method="estimated").fit()
    fcast1 = fit1.forecast(12)
    f1.append(fcast1[11])

f=pd.DataFrame(f1,index=toyota_test.index,columns=["forecast"])
if np.inf in list(f["forecast"]):
    f.replace([np.inf, -np.inf], train.mean(), inplace=True)
```

```
f.fillna(train.mean(),inplace=True)

plt.plot(toyota_test['TOYOTA'])
plt.plot(f)
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 1")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model01.png",bbox_inches='tight')
plt.show()
print("RMSE = ",np.sqrt(mean_squared_error(toyota_test['TOYOTA'],f)))
print("MAPE = ",mean_absolute_percentage_error(toyota_test['TOYOTA'],f))
print('MAE = ',mean_absolute_error(toyota_test['TOYOTA'],f))
scores['model_1']=[np.sqrt(mean_squared_error(toyota_test['TOYOTA'],f)),mean_absolute_percentage_error(toyota_test['TOYOTA'],f),mean_absolut
e_error(toyota_test['TOYOTA'],f)]

# # TOYOTA Auto Exponential Smoothing
from sktime.forecasting.ets import AutoETS

toyota=pd.DataFrame(dataset['TOYOTA'])
toyota.index=idx =pd.period_range(start='2004-01', end='2021-07', freq='M')
toyota=pd.DataFrame(toyota,columns=["TOYOTA"])
toyota["TOYOTA"]=toyota["TOYOTA"].astype("float64")
test=toyota.iloc[-35:,0]
f1=[]
for i in range(35):
    train=toyota.iloc[:165+i,0]
    fit1 = AutoETS(auto=True, n_jobs=-1, sp=12).fit(train)
    fcast1 = fit1.predict(12)
    f1.append(fcast1[-1])
    print(fit1.get_params())
f=pd.DataFrame(f1,index=test.index,columns=["forecast"])
if np.inf in list(f["forecast"]):
    f.replace([np.inf, -np.inf], train.mean(), inplace=True)
f.fillna(train.mean(),inplace=True)
f.index=f.index.astype("string")
test=pd.DataFrame(test)
test.index=test.index.astype('string')

tir=pd.DatetimeIndex(test.index)
plt.plot(pd.DataFrame(list(test["TOYOTA"]),index=tir))
plt.plot(pd.DataFrame(list(f["forecast"]),index=tir))
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 2")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model02.png",bbox_inches='tight')
plt.show()

print("RMSE = ",np.sqrt(mean_squared_error(test,f["forecast"])))
print("MAPE = ",mean_absolute_percentage_error(test,f["forecast"]))
print('MAE = ',mean_absolute_error(test,f["forecast"]))
scores['model_2']=[np.sqrt(mean_squared_error(test,f["forecast"])),mean_absolute_percentage_error(test,f["forecast"]),mean_absolute_error(test,f["f
orecast"])]

toyota=dataset['TOYOTA']
```

```
toyota.index=pd.to_datetime(toyota.index)
toyota.index = pd.DatetimeIndex(toyota.index.values,freq=toyota.index.inferred_freq)
toyota=pd.DataFrame(toyota)
toyota["#"]=list(range(len(toyota)))
test=(toyota.iloc[176:211])

# # TOYOTA Auto ARIMA
toyota=dataset['TOYOTA']
toyota.index=pd.to_datetime(toyota.index)
toyota.index = pd.DatetimeIndex(toyota.index.values,freq=toyota.index.inferred_freq)
toyota=pd.DataFrame(toyota)
toyota["#"]=list(range(len(toyota)))
test=(toyota.iloc[176:211])

f1=[]
for i in range(35):

    train=toyota.iloc[:165+i,0]
    model = pm.auto_arima(train, seasonal=True, m=12)
    print(model.get_params)
    forecasts = model.predict(12)
    f1.append(forecasts[-1])

f=pd.DataFrame(f1,index=test.index,columns=["forecast"])
plt.plot(test['TOYOTA'])
plt.plot(f)
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 3")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model03.png",bbox_inches='tight')
plt.show()

print("RMSE = ",np.sqrt(mean_squared_error(test['TOYOTA'],f)))
print("MAPE = ",mean_absolute_percentage_error(test['TOYOTA'],f))
print('MAE = ',mean_absolute_error(test['TOYOTA'],f))
scores['model_3']=[np.sqrt(mean_squared_error(toyota_test['TOYOTA'],f)),mean_absolute_percentage_error(toyota_test['TOYOTA'],f),mean_absolut
e_error(toyota_test['TOYOTA'],f)]

# # TOYOTA AR model St=βo+β1*S(t-12)+β2*S(t-1)+εt
toyota=pd.DataFrame(dataset['TOYOTA'])
toyota1=toyota['TOYOTA'].shift(periods=1)
toyota12=toyota['TOYOTA'].shift(periods=12)
bltoyota=pd.DataFrame(toyota).join([pd.DataFrame(toyota1).rename(columns={'TOYOTA':'toyota-
1'}),pd.DataFrame(toyota12).rename(columns={'TOYOTA':'toyota-12'})])

y=bltoyota['TOYOTA'][12:].astype('int')
X=bltoyota[['toyota-1','toyota-12']][12:].astype('int')

X_train=X.iloc[:164]
X_test=X.iloc[164:]
y_train=y.iloc[:164]
y_test=y.iloc[164:]

reg = LinearRegression().fit(X_train, y_train)
```

```
y_pred=reg.predict(X_test)

plt.plot(pd.DataFrame(y_test,index=y_test.index))
plt.plot(pd.DataFrame(y_pred,index=y_test.index))
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 4")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model04.png",bbox_inches='tight')
plt.show()

print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
print('MAE = ', mean_absolute_error(y_test,y_pred))

scores['model_4']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_pre
d)]
# # Advanced Models ++ Gtrends
ds=dataset.loc[:,["TOYOTA","gt1_toyota","gt2_toyota"]+list(dataset.columns[-14:])]
ds["TOYOTA+12"]=ds["TOYOTA"].shift(-12)
ds1=ds.loc[:"2020-07",:]
ds1.tail(15)

index1=ds[-35:].index
ind=pd.DatetimeIndex(index1)

# ##  TOYOTA model_5 DecisionTreeRegressor
Y=ds1["TOYOTA+12"]
X=ds1[[  'TOYOTA','gt1_toyota','gt2_toyota','unemployment','Qgdp_Maverage', 'gt_fir']]

y_test=Y[-35:]
X_test=X[-35:]
y_train=Y[120:-35]
X_train=X[120:-35]

tree=DecisionTreeRegressor(random_state=0,max_depth=4)   ###### να θυμηθω να βρω αποτελέσματα και για χωρις google trends μεταβλητες
tree.fit(X_train,y_train)
y_pred=tree.predict(X_test)

plt.plot(pd.DataFrame(list(y_test),index=ind))
plt.plot(pd.DataFrame(y_pred,index=ind))
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 5")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model05.png",bbox_inches='tight')
plt.show()

print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
print('MAE = ', mean_absolute_error(y_test,y_pred))
scores['model_5']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_pre
d)]

# ## model_6 DecisionTreeRegressor
Y=ds1["TOYOTA+12"]
```

```
X=ds1[[  'TOYOTA','unemployment','Qgdp_Maverage']]

y_test=Y[-35:]
X_test=X[-35:]
y_train=Y[120:-35]
X_train=X[120:-35]

tree=DecisionTreeRegressor(random_state=0,max_depth=4)
tree.fit(X_train,y_train)

y_pred=tree.predict(X_test)

plt.plot(pd.DataFrame(list(y_test),index=ind))
plt.plot(pd.DataFrame(y_pred,index=ind))
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 6")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model06.png",bbox_inches='tight')
plt.show()

print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
print('MAE = ', mean_absolute_error(y_test,y_pred))
scores['model_6']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_pre
d)]

### TOYOTA model_7 RandomForestRegressor
Y=ds1["TOYOTA+12"]
X=ds1[[  'TOYOTA','gt1_toyota','gt2_toyota','unemployment','Qgdp_Maverage', 'gt_fir']]

y_test=Y[-35:]
X_test=X[-35:]

y_train=Y[120:-35]
X_train=X[120:-35]

forest=RandomForestRegressor(random_state=0,max_depth=4)   ###### να θυμηθω να βρω αποτελέσματα και για χωρις google trends μεταβλητες
forest.fit(X_train,y_train)
y_pred=forest.predict(X_test)

plt.plot(pd.DataFrame(list(y_test),index=ind))
plt.plot(pd.DataFrame(y_pred,index=ind))
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 7")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model07.png",bbox_inches='tight')
plt.show()

print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
print('MAE = ', mean_absolute_error(y_test,y_pred))
scores['model_7']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_pre
d)]
```

```
# ## TOYOTA model_8 RandomForestRegressor
Y=ds1["TOYOTA+12"]
X=ds1[[  'TOYOTA','unemployment','Qgdp_Maverage']]

y_test=Y[-35:]
X_test=X[-35:]
y_train=Y[120:-35]
X_train=X[120:-35]

forest=RandomForestRegressor(random_state=0,max_depth=4)   ###### να θυμηθω να βρω αποτελέσματα και για χωρις google trends μεταβλητες
forest.fit(X_train,y_train)
y_pred=forest.predict(X_test)

plt.plot(pd.DataFrame(list(y_test),index=ind))
plt.plot(pd.DataFrame(y_pred,index=ind))
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 8")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model08.png",bbox_inches='tight')
plt.show()

print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
print('MAE = ', mean_absolute_error(y_test,y_pred))
scores['model_8']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_pre
d)]

# ## TOYOTA model_9 LinearRegression
# Γραμμική παλινδρόμηση με πωλήσεις, gtrends, ΑΕΠ, ανεργία
Y=ds1["TOYOTA+12"]
X=ds1[['TOYOTA', 'gt1_toyota', 'Qgdp_Maverage', 'unemployment']]

y_test=Y[-35:]
X_test=X[-35:]
y_train=Y[120:-35]
X_train=X[120:-35]

reg = LinearRegression().fit(X_train, y_train)
y_pred=reg.predict(X_test)

plt.plot(pd.DataFrame(list(y_test),index=ind))
plt.plot(pd.DataFrame(y_pred,index=ind))
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 9")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model09.png",bbox_inches='tight')
plt.show()

print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
print('MAE = ', mean_absolute_error(y_test,y_pred))
scores['model_9']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_pre
d)]
```

```
# ## TOYOTA model_10 Linear Regression
# Γραμμική Παλινδρόμηση χωρίς gtrends
Y=ds1["TOYOTA+12"]
X=ds1[['TOYOTA','Qgdp_Maverage', 'unemployment']]
#####subtrain1
y_test=Y[-35:]
X_test=X[-35:]
y_train=Y[120:-35]
X_train=X[120:-35]

reg = LinearRegression().fit(X_train, y_train)
y_pred=reg.predict(X_test)

plt.plot(pd.DataFrame(list(y_test),index=ind))
plt.plot(pd.DataFrame(y_pred,index=ind))
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 10")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model10.png",bbox_inches='tight')
plt.show()

print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
print('MAE = ', mean_absolute_error(y_test,y_pred))
scores['model_10']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_pr
ed)]

# ## TOYOTA model_11 Support Vector Regressor
# Παλινδρόμηση με μηχανές διανυσματικής υποστήριξης με GTrends
Y=ds1["TOYOTA+12"]
X=ds1[['TOYOTA','gt1_toyota','Qgdp_Maverage', 'unemployment']]

y_test=Y[-35:]
X_test=X[-35:]
y_train=Y[120:-35]
X_train=X[120:-35]

regr = make_pipeline(StandardScaler(), SVR(C=15, epsilon=0.4,kernel="linear"))
regr.fit(X_train, y_train)
y_pred=regr.predict(X_test)

plt.plot(pd.DataFrame(list(y_test),index=ind))
plt.plot(pd.DataFrame(y_pred,index=ind))
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 11")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model11.png",bbox_inches='tight')
plt.show()

print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
print('MAE = ', mean_absolute_error(y_test,y_pred))
```

```
scores['model_11']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_pred)]

### TOYOTA model_12 Support Vector Regressor
# Παλινδρόμηση με μηχανές διανυσματικής υποστήριξης χωρίς GTrends
Y=ds1["TOYOTA+12"]
X=ds1[['TOYOTA','Qgdp_Maverage', 'unemployment']]

y_test=Y[-35:]
X_test=X[-35:]
y_train=Y[120:-35]
X_train=X[120:-35]

from sklearn.svm import SVR
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler, MinMaxScaler

regr = make_pipeline(StandardScaler(), SVR(C=15, epsilon=0.4,kernel="linear"))
regr.fit(X_train, y_train)
y_pred=regr.predict(X_test)

plt.plot(pd.DataFrame(list(y_test),index=ind))
plt.plot(pd.DataFrame(y_pred,index=ind))
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title("TOYOTA model 12")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/TOYOTA_model12.png",bbox_inches='tight')
plt.show()
print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
print('MAE = ', mean_absolute_error(y_test,y_pred))
scores['model_12']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_pred)]
scores_toyota=scores
# # Υπολοιπα Brands
sc=[]
sc.append(scores_toyota)
for k in ['AUDI','OPEL','FORD']:
    print(k,k,k,k,k,k,k,k,k,k,k,k,k)

    z=dataset[k]
    z.index=pd.to_datetime(z.index)
    z.index = pd.DatetimeIndex(z.index.values,freq=z.index.inferred_freq)
    z=pd.DataFrame(z)
    z["#"]=list(range(len(z)))
    test=(z.iloc[176:211])
    scores={}

    #### Exponential smoothing
    print(k,"EXPONENTIAL SMOOTHING HOLT WINTERS")
    z=dataset[k]
    z.index=pd.to_datetime(z.index)
    z.index = pd.DatetimeIndex(z.index.values,freq=z.index.inferred_freq)
    z=pd.DataFrame(z)
```

```
z["#"]=list(range(len(z)))
test=(z.iloc[176:211])

f1=[]
for i in range(35):
    train=z.iloc[:165+i,0]
    fit1 = sm.tsa.ExponentialSmoothing(train,seasonal_periods=12,seasonal="add",use_boxcox=False, initialization_method="estimated").fit()
    fcast1 = fit1.forecast(12)
    f1.append(fcast1[11])

f=pd.DataFrame(f1,index=test.index,columns=["forecast"])
if np.inf in list(f["forecast"]):
    f.replace([np.inf, -np.inf], train.mean(), inplace=True)
f.fillna(train.mean(),inplace=True)

plt.plot(test[k])
plt.plot(f)
plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
plt.title(k+" model 1")
plt.legend(["Actual values","Predicted values"])
plt.savefig("/Users/loukas/Desktop/par/"+k+"_model01.png",bbox_inches='tight')
plt.show()

print(k,'model_1')
print("RMSE = ",np.sqrt(mean_squared_error(test[k],f)))
print("MAPE = ",mean_absolute_percentage_error(test[k],f))
print('MAE = ',mean_absolute_error(test[k],f))
scores['model_1']=[np.sqrt(mean_squared_error(test[k],f)),mean_absolute_percentage_error(test[k],f),mean_absolute_error(test[k],f)]

######### AUTOETS
print(k,"AUTO EXPONENTIAL SMOOTHING")

from sktime.forecasting.ets import AutoETS
z=pd.DataFrame(dataset[k])
z.index=idx =pd.period_range(start='2004-01', end='2021-07', freq='M')
z=pd.DataFrame(z,columns=[k])
z[k]=z[k].astype("float64")
test=z.iloc[-35:,0]
f1=[]
for i in range(35):
    train=z.iloc[:165+i,0]
    fit1 = AutoETS(auto=True, n_jobs=-1, sp=12).fit(train)
    fcast1 = fit1.predict(12)
    f1.append(fcast1[-1])

f=pd.DataFrame(f1,index=test.index,columns=["forecast"])
if np.inf in list(f["forecast"]):
    f.replace([np.inf, -np.inf], train.mean(), inplace=True)
f.fillna(train.mean(),inplace=True)

f.index=f.index.astype("string")
test=pd.DataFrame(test)
test.index=test.index.astype('string')
```

```
    tir=pd.DatetimeIndex(test.index)
    plt.plot(pd.DataFrame(list(test[k]),index=tir))
    plt.plot(pd.DataFrame(list(f["forecast"]),index=tir))
    plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
    plt.title(k+" model 2")
    plt.legend(["Actual values","Predicted values"])
    plt.savefig("/Users/loukas/Desktop/par/"+k+"_model02.png",bbox_inches='tight')
    plt.show()

    print(k,'model_2')
    print("RMSE = ",np.sqrt(mean_squared_error(test,f["forecast"])))
    print("MAPE = ",mean_absolute_percentage_error(test,f["forecast"]))
    print('MAE = ',mean_absolute_error(test,f["forecast"]))
    scores['model_2']=[np.sqrt(mean_squared_error(test,f["forecast"])),mean_absolute_percentage_error(test,f["forecast"]),mean_absolute_error(test,f
["forecast"])]

    z=dataset[k]
    z.index=pd.to_datetime(z.index)
    z.index = pd.DatetimeIndex(z.index.values,freq=z.index.inferred_freq)
    z=pd.DataFrame(z)
    z["#"]=list(range(len(z)))
    test=(z.iloc[176:211])

    # ARIMA
    print(k,"AUTO ARIMA")

    z=dataset[k]
    z.index=pd.to_datetime(z.index)
    z.index = pd.DatetimeIndex(z.index.values,freq=z.index.inferred_freq)
    z=pd.DataFrame(z)
    z["#"]=list(range(len(z)))
    test=(z.iloc[176:211])

    f1=[]
    for i in range(35):

        train=z.iloc[:165+i,0]
        model = pm.auto_arima(train, seasonal=True, m=12)
        print(model.get_params)
        forecasts = model.predict(12)
        f1.append(forecasts[-1])

    f=pd.DataFrame(f1,index=test.index,columns=["forecast"])
    plt.plot(test[k])
    plt.plot(f)
    plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
    plt.title(k+" model 3")
    plt.legend(["Actual values","Predicted values"])
    plt.savefig("/Users/loukas/Desktop/par/"+k+"_model03.png",bbox_inches='tight')
    plt.show()

    print(k,'model_3')
    print("RMSE = ",np.sqrt(mean_squared_error(test[k],f)))
    print("MAPE = ",mean_absolute_percentage_error(test[k],f))
```

```
    print('MAE = ',mean_absolute_error(test[k],f))
    scores['model_3']=[np.sqrt(mean_squared_error(test[k],f)),mean_absolute_percentage_error(test[k],f),mean_absolute_error(test[k],f)]

    # AR model St=βo+β1*S(t-12)+β2*S(t-1)+εt
    print(k,"AR Baseline Model")

    z=pd.DataFrame(dataset[k])
    z1=z[k].shift(periods=1)
    z12=z[k].shift(periods=12)

    blz=pd.DataFrame(z).join([pd.DataFrame(z1).rename(columns={k:k.lower()+'-1'}),pd.DataFrame(z12).rename(columns={k:k.lower()+'-12'})])
    y=blz[k][12:].astype('int')
    X=blz[[k.lower()+'-1',k.lower()+'-12']][12:].astype('int')

    X_train=X.iloc[:164]
    X_test=X.iloc[164:]
    y_train=y.iloc[:164]
    y_test=y.iloc[164:]

    reg = LinearRegression().fit(X_train, y_train)
    y_pred=reg.predict(X_test)

    plt.plot(pd.DataFrame(y_test,index=y_test.index))
    plt.plot(pd.DataFrame(y_pred,index=y_test.index))
    plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
    plt.title(k+" model 4")
    plt.legend(["Actual values","Predicted values"])
    plt.savefig("/Users/loukas/Desktop/par/"+k+"_model04.png",bbox_inches='tight')
    plt.show()

    print(k,'model_4')
    print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
    print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
    print('MAE = ', mean_absolute_error(y_test,y_pred))
    scores['model_4']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_p
red)]

    # Advanced Models ++ Gtrends
    ds=dataset.loc[:,[k,'gt1_'+k.lower(),'gt2_'+k.lower()]+list(dataset.columns[-14:])]
    ds[k+'+12']=ds[k].shift(-12)
    ds1=ds.loc[:"2020-07",:]

    print(k,"Decision Tree + gtrends")

    Y=ds1[k+'+12']
    X=ds1[[ k,'gt1_'+k.lower(),'gt2_'+k.lower(),'unemployment','Qgdp_Maverage', 'gt_fir']]

    y_test=Y[-35:]
    X_test=X[-35:]
    y_train=Y[120:-35]
    X_train=X[120:-35]

    tree=DecisionTreeRegressor(random_state=0,max_depth=4)   ###### να θυμηθω να βρω αποτελέσματα και για χωρις google trends μεταβλητες
    tree.fit(X_train,y_train)
```

```
    y_pred=tree.predict(X_test)

    plt.plot(pd.DataFrame(list(y_test),index=ind))
    plt.plot(pd.DataFrame(y_pred,index=ind))
    plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
    plt.title(k+" model 5")
    plt.legend(["Actual values","Predicted values"])
    plt.savefig("/Users/loukas/Desktop/par/"+k+"_model05.png",bbox_inches='tight')
    plt.show()

    print(k,'model_5')
    print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
    print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
    print('MAE = ', mean_absolute_error(y_test,y_pred))
    scores['model_5']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_p
red)]

    print(k,"Decision Tree without gtrends")
    Y=ds1[k+'+12']
    X=ds1[[  k,'unemployment','Qgdp_Maverage']]

    y_test=Y[-35:]
    X_test=X[-35:]
    y_train=Y[120:-35]
    X_train=X[120:-35]

    tree=DecisionTreeRegressor(random_state=0,max_depth=4)   ###### να θυμηθω να βρω αποτελέσματα και για χωρις google trends μεταβλητες
    tree.fit(X_train,y_train)
    y_pred=tree.predict(X_test)

    plt.plot(pd.DataFrame(list(y_test),index=ind))
    plt.plot(pd.DataFrame(y_pred,index=ind))
    plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
    plt.title(k+" model 6")
    plt.legend(["Actual values","Predicted values"])
    plt.savefig("/Users/loukas/Desktop/par/"+k+"_model06.png",bbox_inches='tight')
    plt.show()

    print(k,'model_6')
    print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
    print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
    print('MAE = ', mean_absolute_error(y_test,y_pred))
    scores['model_6']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_p
red)]

    print(k,"Random Forest + gtrends")
    Y=ds1[k+'+12']
    X=ds1[[  k,'gt1_'+k.lower(),'gt2_'+k.lower(),'unemployment','Qgdp_Maverage', 'gt_fir']]

    y_test=Y[-35:]
    X_test=X[-35:]
    y_train=Y[120:-35]
    X_train=X[120:-35]
```

```
    forest=RandomForestRegressor(random_state=0,max_depth=4)   ###### να θυμηθω να βρω αποτελέσματα και για χωρις google trends μεταβλητες
    forest.fit(X_train,y_train)
    y_pred=forest.predict(X_test)

    plt.plot(pd.DataFrame(list(y_test),index=ind))
    plt.plot(pd.DataFrame(y_pred,index=ind))
    plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
    plt.title(k+" model 7")
    plt.legend(["Actual values","Predicted values"])
    plt.savefig("/Users/loukas/Desktop/par/"+k+"_model07.png",bbox_inches='tight')
    plt.show()

    print(k,'model_7')
    print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
    print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
    print('MAE = ', mean_absolute_error(y_test,y_pred))
    scores['model_7']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_pred)]

    print(k,"Random Forest without gtrends")
    Y=ds1[k+'+12']
    X=ds1[[  k,'unemployment','Qgdp_Maverage']]

    y_test=Y[-35:]
    X_test=X[-35:]
    y_train=Y[120:-35]
    X_train=X[120:-35]

    forest=RandomForestRegressor(random_state=0,max_depth=4)   ###### να θυμηθω να βρω αποτελέσματα και για χωρις google trends μεταβλητες
    forest.fit(X_train,y_train)
    y_pred=forest.predict(X_test)

    plt.plot(pd.DataFrame(list(y_test),index=ind))
    plt.plot(pd.DataFrame(y_pred,index=ind))
    plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
    plt.title(k+" model 8")
    plt.legend(["Actual values","Predicted values"])
    plt.savefig("/Users/loukas/Desktop/par/"+k+"_model08.png",bbox_inches='tight')
    plt.show()

    print(k,'model_8')
    print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
    print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
    print('MAE = ', mean_absolute_error(y_test,y_pred))
    scores['model_8']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_pred)]

    print(k,"Linear Regression with gtrends")
    Y=ds1[k+'+12']
    X=ds1[[k, 'gt1_'+k.lower(), 'Qgdp_Maverage', 'unemployment']]

    y_test=Y[-35:]
```

```
    X_test=X[-35:]
    y_train=Y[120:-35]
    X_train=X[120:-35]

    reg = LinearRegression().fit(X_train, y_train)
    y_pred=reg.predict(X_test)

    plt.plot(pd.DataFrame(list(y_test),index=ind))
    plt.plot(pd.DataFrame(y_pred,index=ind))
    plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
    plt.title(k+" model 9")
    plt.legend(["Actual values","Predicted values"])
    plt.savefig("/Users/loukas/Desktop/par/"+k+"_model09.png",bbox_inches='tight')
    plt.show()

    print(k,'model_9')
    print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
    print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
    print('MAE = ', mean_absolute_error(y_test,y_pred))
    scores['model_9']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_p
red)]

    print(k,"Linear Regression without gtrends")
    Y=ds1[k+'+12']
    X=ds1[[k,'Qgdp_Maverage', 'unemployment']]

    y_test=Y[-35:]
    X_test=X[-35:]
    y_train=Y[120:-35]
    X_train=X[120:-35]

    reg = LinearRegression().fit(X_train, y_train)
    y_pred=reg.predict(X_test)

    plt.plot(pd.DataFrame(list(y_test),index=ind))
    plt.plot(pd.DataFrame(y_pred,index=ind))
    plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
    plt.title(k+" model 10")
    plt.legend(["Actual values","Predicted values"])
    plt.savefig("/Users/loukas/Desktop/par/"+k+"_model10.png",bbox_inches='tight')
    plt.show()

    print(k,'model_10')
    print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
    print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
    print('MAE = ', mean_absolute_error(y_test,y_pred))
    scores['model_10']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_
pred)]

    print(k,"SVR with gtrends")
    Y=ds1[k+'+12']
    X=ds1[[k,'gt1_'+k.lower(),'Qgdp_Maverage', 'unemployment']]

    y_test=Y[-35:]
```

```
    X_test=X[-35:]
    y_train=Y[120:-35]
    X_train=X[120:-35]

    from sklearn.svm import SVR
    from sklearn.pipeline import make_pipeline
    from sklearn.preprocessing import StandardScaler, MinMaxScaler

    regr = make_pipeline(StandardScaler(), SVR(C=15, epsilon=0.4,kernel="linear"))
    regr.fit(X_train, y_train)
    y_pred=regr.predict(X_test)

    plt.plot(pd.DataFrame(list(y_test),index=ind))
    plt.plot(pd.DataFrame(y_pred,index=ind))
    plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
    plt.title(k+" model 11")
    plt.legend(["Actual values","Predicted values"])
    plt.savefig("/Users/loukas/Desktop/par/"+k+"_model11.png",bbox_inches='tight')
    plt.show()

    print(k,'model_11')
    print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
    print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
    print('MAE = ', mean_absolute_error(y_test,y_pred))
    scores['model_11']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_
pred)]

    print(k,"SVR without gtrends")
    Y=ds1[k+'+12']
    X=ds1[[k,'Qgdp_Maverage', 'unemployment']]

    y_test=Y[-35:]
    X_test=X[-35:]
    y_train=Y[120:-35]
    X_train=X[120:-35]

    from sklearn.svm import SVR
    from sklearn.pipeline import make_pipeline
    from sklearn.preprocessing import StandardScaler, MinMaxScaler

    regr = make_pipeline(StandardScaler(), SVR(C=15, epsilon=0.4,kernel="linear"))
    regr.fit(X_train, y_train)
    y_pred=regr.predict(X_test)

    plt.plot(pd.DataFrame(list(y_test),index=ind))
    plt.plot(pd.DataFrame(y_pred,index=ind))
    plt.tick_params(axis='x', which='major',labelsize=10,labelrotation=45)
    plt.title(k+" model 12")
    plt.legend(["Actual values","Predicted values"])
    plt.savefig("/Users/loukas/Desktop/par/"+k+"_model12.png",bbox_inches='tight')
    plt.show()

    print(k,'model_12')
    print('RMSE = ',np.sqrt(mean_squared_error(y_test,y_pred)))
```

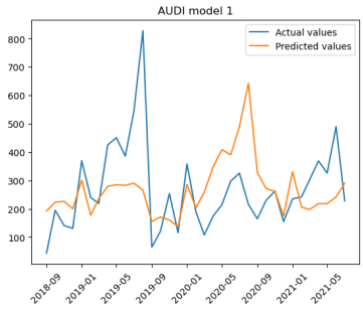
```
    print('MAPE = ',mean_absolute_percentage_error(y_test,y_pred))
    print('MAE = ', mean_absolute_error(y_test,y_pred))
    scores['model_12']=[np.sqrt(mean_squared_error(y_test,y_pred)),mean_absolute_percentage_error(y_test,y_pred),mean_absolute_error(y_test,y_
pred)]

    sc.append(scores)

#### Συγκέντρωση τα αποτελέσματα πινακοποιημένα για κάθε μάρκα
pd.DataFrame(sc[0],index=["RMSE","MAPE","MAE"])####TOYOTA
pd.DataFrame(sc[1],index=["RMSE","MAPE","MAE"])####AUDI
pd.DataFrame(sc[2],index=["RMSE","MAPE","MAE"])####OPEL
pd.DataFrame(sc[3],index=["RMSE","MAPE","MAE"])####FORD

with pd.ExcelWriter("/Users/loukas/Desktop/scores.xlsx") as writer:
    pd.DataFrame(sc[0],index=["RMSE","MAPE","MAE"]).to_excel(writer,sheet_name='Scores',startrow=5,startcol=5)####TOYOTA
    pd.DataFrame(sc[1],index=["RMSE","MAPE","MAE"]).to_excel(writer,sheet_name='Scores',startrow=10,startcol=5)####AUDI
    pd.DataFrame(sc[2],index=["RMSE","MAPE","MAE"]).to_excel(writer,sheet_name='Scores',startrow=15,startcol=5)####OPEL
    pd.DataFrame(sc[3],index=["RMSE","MAPE","MAE"]).to_excel(writer,sheet_name='Scores',startrow=20,startcol=5)####FORD
```

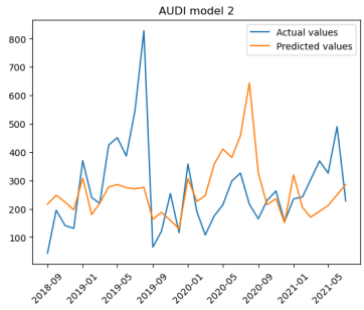
Παράρτημα ΙΙΙ

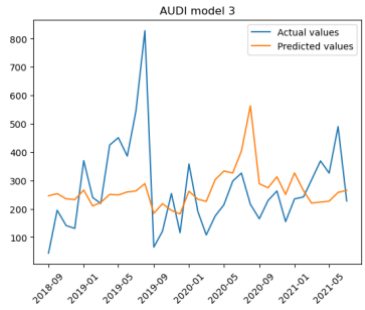
Στο συγκεκριμένο Παράρτημα της εργασίας για κάθε μάρκα, παρουσιάζονται σε διαγράμματα οι προβλέψεις που παράχθηκαν από κάθε μοντέλο για την περίοδο Σεπτέμβριος 2018 ως Ιούλιος 2021 σε αντιστοιχία με τις πραγματικές τιμές των πωλήσεων αυτοκινήτων την ίδια περίοδο.

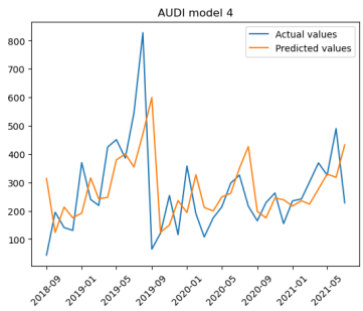
Μάρκα **TOYOTA**

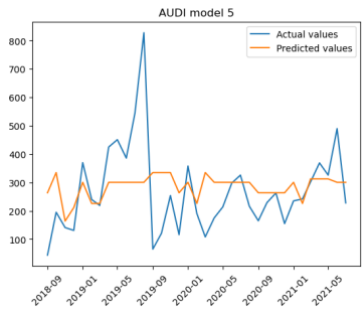
Μάρκα **AUDI**

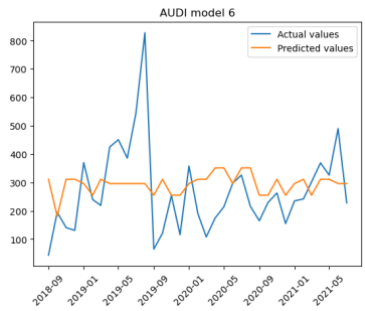

AUDI model 1
Actual values
Predicted values

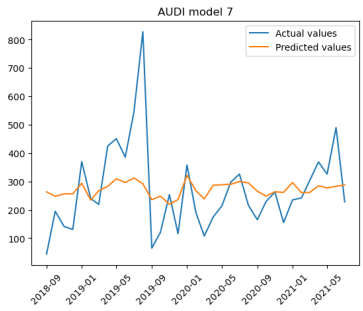

AUDI model 2
Actual values
Predicted values

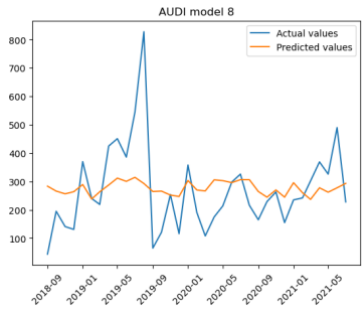

AUDI model 3
Actual values
Predicted values

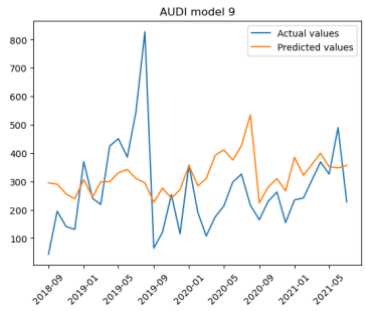

AUDI model 4
Actual values
Predicted values

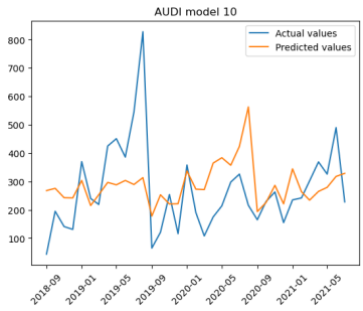

AUDI model 5
Actual values
Predicted values

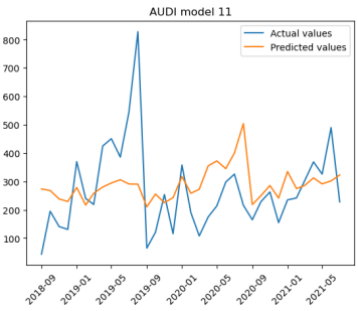

AUDI model 6
Actual values
Predicted values

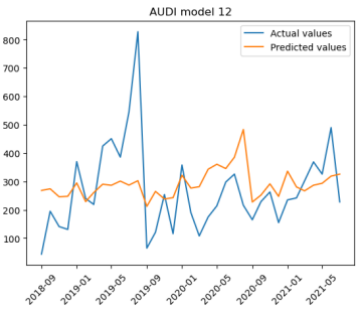

AUDI model 7
Actual values
Predicted values


AUDI model 8
Actual values
Predicted values

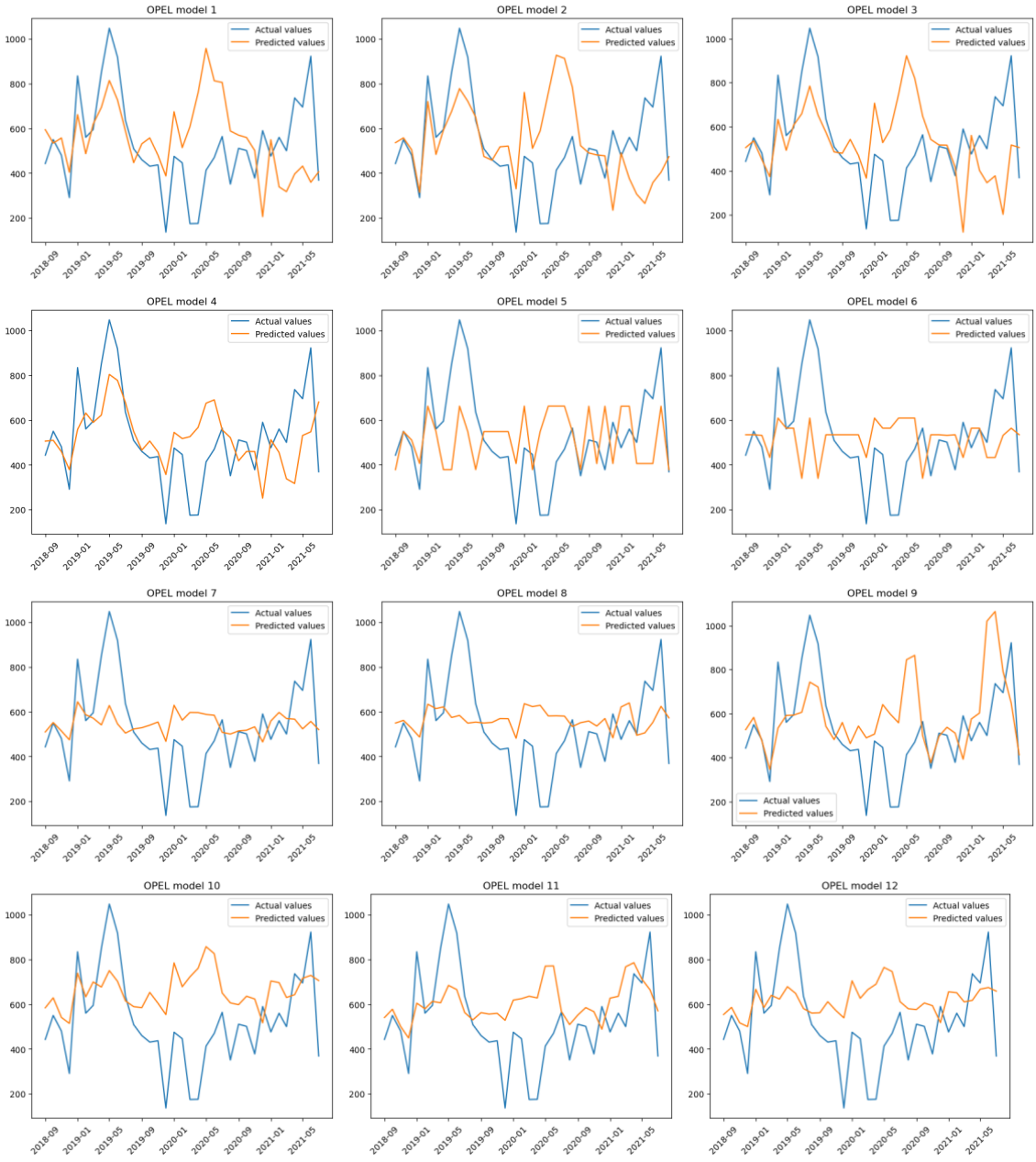

AUDI model 9
Actual values
Predicted values


AUDI model 10
Actual values
Predicted values

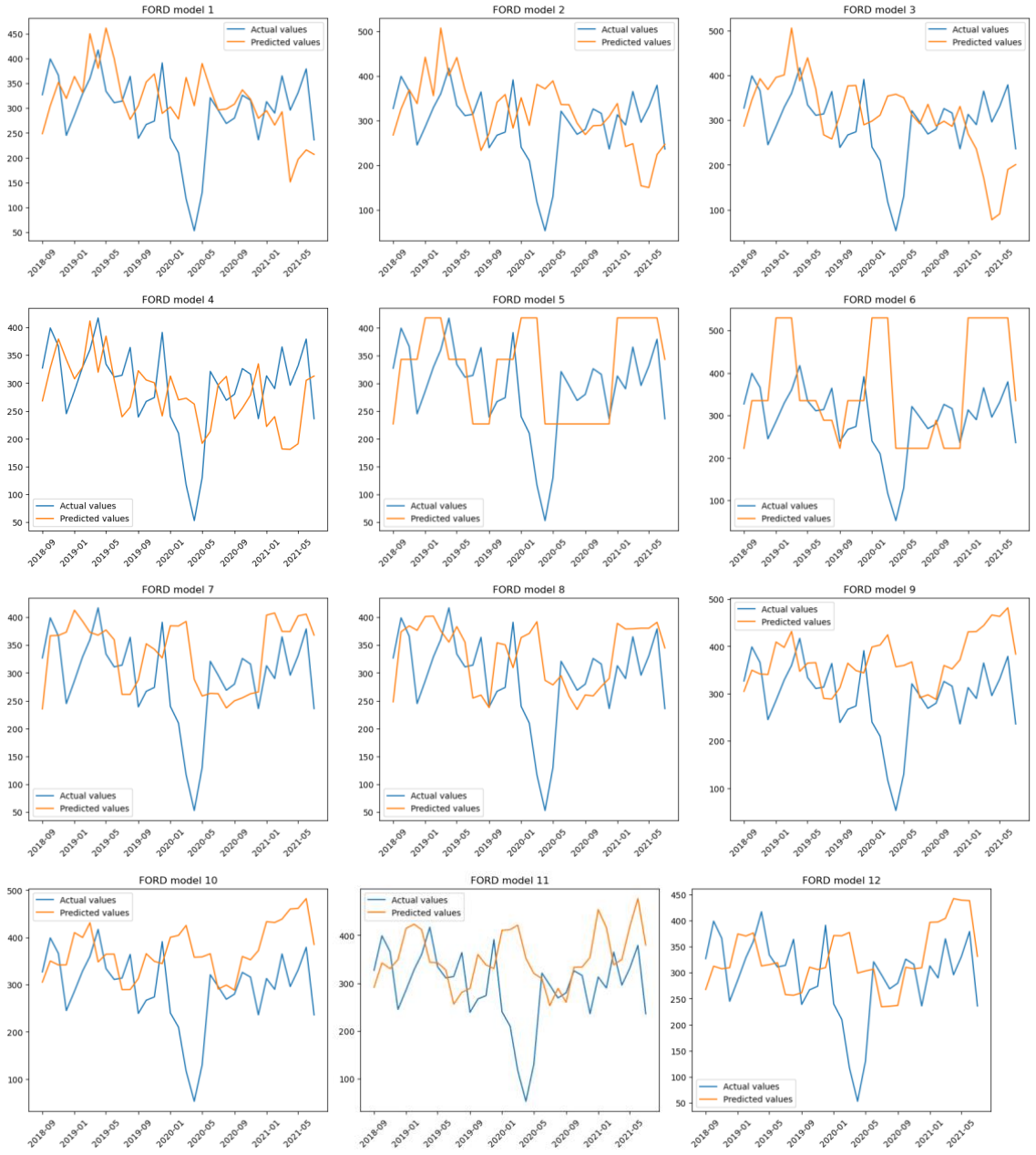

AUDI model 11
Actual values
Predicted values


AUDI model 12
Actual values
Predicted values

Μάρκα **OPEL**


OPEL model 1
OPEL model 2
OPEL model 3
OPEL model 4
OPEL model 5
OPEL model 6
OPEL model 7
OPEL model 8
OPEL model 9
OPEL model 10
OPEL model 11
OPEL model 12
Actual values
Predicted values

Μάρκα **FORD**


FORD model 1
FORD model 2
FORD model 3
FORD model 4
FORD model 5
FORD model 6
FORD model 7
FORD model 8
FORD model 9
FORD model 10
FORD model 11
FORD model 12
Actual values
Predicted values