Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

Ανάπτυξη Πλατφόρμας Συλλογής, Δημιουργίας, Διαμοιρασμού και Οπτικοποίησης Δεδομένων Χρονοσειρών

Διπλωματικη Εργασια

του

ΚΟΥΛΟΥΜΟΥ ΑΝΔΡΕΑ

Επιβλέπων: Ασημακόπουλος Βασίλειος
Ομότιμος Καθηγητής Ε.Μ.Π

Υπεύθυνος: Ευάγγελος Σπηλιώτης
Διδάκτωρ Ε.Μ.Π

Αθήνα, Οκτώβριος 2023

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

Ανάπτυξη Πλατφόρμας Συλλογής, Δημιουργίας, Διαμοιρασμού και Οπτικοποίησης Δεδομένων Χρονοσειρών

Διπλωματικη Εργασια

του

ΚΟΥΛΟΥΜΟΥ ΑΝΔΡΕΑ

Επιβλέπων: Ασημακόπουλος Βασίλειος
Ομότιμος Καθηγητής Ε.Μ.Π

Υπεύθυνος: Ευάγγελος Σπηλιώτης
Διδάκτωρ Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Οκτωβρίου 2023.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ασημακόπουλος Βασίλειος	Ψαρράς Ιωάννης	Ασκούνης Δημήτριος
Ομότιμος Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβριος 2023

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

(Υπογραφή)

.....
Κούλουμος Ανδρέας
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π

Οκτώβριος 2023

Περίληψη

Στην σημερινή εποχή, οι ραγδαίες εξελίξεις σε κάθε τομέα δημιουργούν ένα περιβάλλον με υψηλό βαθμό αβεβαιότητας. Οι τεχνικές προβλέψεων προσφέρουν το πλαίσιο για την αξιοποίηση των διαθέσιμων δεδομένων, επιτρέποντας την πρόβλεψη τάσεων και την προετοιμασία για μελλοντικές προκλήσεις και ευκαιρίες. Η ανάγκη για αξιόπιστες προβλέψεις είναι πιο κρίσιμη από ποτέ.

Ωστόσο, η αξιολόγηση των διαφόρων μεθόδων πρόβλεψης αντιμετωπίζει προκλήσεις. Παρά την ευρεία αποδοχή δημοφιλών συνόλων δεδομένων ως ¨σημεία αναφοράς¨ (benchmarks), η ερευνητική κοινότητα εκφράζει ανησυχίες σχετικά με την περιορισμένη ποικιλομορφία και ποσότητα τους. Επιπλέον, αυτή η προσέγγιση αρχίζει να εμφανίζει δυσκολίες όταν ο ερευνητής επιθυμεί να αξιολογήσει μια μέθοδο που αφορά χρονοσειρές με πολύ συγκεκριμένα χαρακτηριστικά ή όταν χρειάζεται μεγάλο πλήθος χρονοσειρών.

Η παρούσα διπλωματική εργασία αντιμετωπίζει αυτά τα ζητήματα, προτείνοντας το FOREcasting & Strategy Unit Data Collection, ή αλλιώς FOREDeCk, μια διαδικτυακή πλατφόρμα σχεδιασμένη για να παρέχει εύκολη πρόσβαση σε πάνω από ένα εκατομμύριο πραγματικές χρονοσειρές με ποικίλα χαρακτηριστικά που καλύπτουν κάθε τομέα, από την βιομηχανία και τον τουρισμό μέχρι την υγεία και το περιβάλλον. Επιπλέον, για πιο εξειδικευμένες ανάγκες, επιτρέπει τη δημιουργία τεχνητών χρονοσειρών με ελεγχόμενα χαρακτηριστικά.

Ο ευρύτερος στόχος του FOREDeCk είναι η υποστήριξη της ερευνητικής κοινότητας των τεχνικών προβλέψεων με την παροχή νέων υψηλής ποιότητας συνόλων δεδομένων και τα κατάλληλα εργαλεία για την αποτελεσματική αξιοποίησή τους. Σε αυτό το πλαίσιο, το σύστημα έχει αξιοποιηθεί για την υποστήριξη του διαγωνισμού προβλέψεων M4, παρέχοντας τις 100.000 χρονοσειρές που χρησιμοποιήθηκαν σε αυτόν.

Λέξεις Κλειδιά

Χρονοσειρές, Διαδικτυακή Εφαρμογή, Τεχνικές Προβλέψεων, Γεννήτρια Τεχνητών Χρονοσειρών

Abstract

In today's era, rapid developments in every field create an environment characterized by a high degree of uncertainty. Forecasting techniques provide the framework for utilizing available data, allowing the prediction of trends and preparation for future challenges and opportunities. The need for reliable forecasts is more critical than ever.

However, evaluating the various forecasting methods faces challenges. Despite the widespread acceptance of popular datasets as "benchmarks", the research community expresses concerns about their limited diversity and quantity. Additionally, this approach begins to pose difficulties when a researcher wishes to evaluate a method that pertains to time series with very specific characteristics or when a large number of time series is required.

This diploma thesis addresses these issues by proposing the FOREcasting & Strategy Unit Data Collection, named FOREDeCk, an online platform designed to provide easy access to over a million real-world time series with diverse characteristics covering every sector, from industry and tourism to health and the environment. Furthermore, for more specialized needs, it allows the generation of artificial time series with controllable characteristics.

The broader goal of FOREDeCk is to support the forecasting research community by providing new high-quality datasets and the appropriate tools for their effective utilization. In this context, the system has been utilized to support the M4 forecasting competition, contributing the 100,000 time series used in it.

Keywords

Time-Series, Web-based application, Forecasting Techniques, Time-Series generation

Στους γονείς μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ.Βασίλειο Ασημακόπουλο για την ευκαιρία που μου έδωσε μερικά χρόνια πριν, να γίνω μέλος της Μονάδας Προβλέψεων και Στρατηγικής και να ασχοληθώ με τον συναρπαστικό τομέα των προβλέψεων στα πλαίσια της ερευνητικής δραστηριότητας της Μονάδας.

Επίσης θα ήθελα να ευχαριστήσω ιδιαίτερα και τα υπόλοιπα μέλη της Μονάδας Προβλέψεων και Στρατηγικής. Συγκεκριμένα την Διδάκτορα κ.Νικολέττα Ζαμπέτα Λεγάκη, η οποία ήταν η πρώτη μου επαφή με την Μονάδα και ο κύριος λόγος που ο τομέας των προβλέψεων αποτέλεσε βασικό κομμάτι τον προπτυχιακόν μου σπουδών. Κυρίως όμως, θα ήθελα να ευχαριστήσω τον Διδάκτορα κ.Βαγγέλη Σπηλιώτη για την πολύτιμη καθοδήγηση που μου παρείχε καθ' όλη τη διάρκεια της συμμετοχής μου στη Μονάδα και που μετά από τόσα χρόνια, συνεχίζει ακόμα να απαντά στα emails μου.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, που πιστεύουν σε μένα και με στηρίζουν όλα αυτά τα χρόνια. Χωρίς αυτούς δεν θα ήμουν αυτός που είμαι σήμερα.

Αθήνα, Οκτώβριος 2023

Κούλουμος Ανδρέας

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο της Εργασίας

Στη σύγχρονη εποχή, η ανάγκη για λήψη σωστών αποφάσεων για το μέλλον είναι πιο κρίσιμη από ποτέ. Οι ραγδαίες εξελίξεις σε κάθε τομέα, από την οικονομία και την τεχνολογία μέχρι την υγεία και το περιβάλλον, δημιουργούν ένα περιβάλλον υψηλής αβεβαιότητας και καθιστούν την προνοητική πρόβλεψη αναγκαία. Οι τεχνικές προβλέψεων προσφέρουν τη δυνατότητα να αντληθούν πολύτιμες πληροφορίες από τα διαθέσιμα δεδομένα, επιτρέποντας στους οργανισμούς και τις επιχειρήσεις να προβλέπουν τις τάσεις, να προετοιμάζονται για προκλήσεις και ευκαιρίες, και να λαμβάνουν αποφάσεις που βασίζονται σε δεδομένα και ανάλυση, προσφέροντας έτσι μια αξιόπιστη προοπτική για το μέλλον.

Για το λόγο αυτό, είναι ζωτικής σημασίας η αξιολόγηση των διαφόρων μεθόδων πρόβλεψης. Αυτό μας επιτρέπει να αναγνωρίσουμε τις δυνατότητες και τα όρια της κάθε μεθόδου, καθώς και τις συνθήκες υπό τις οποίες η κάθε μία ξεπερνά τις άλλες. Έτσι, μπορούμε να προσαρμόσουμε την προσέγγισή μας σύμφωνα με τον τύπο των δεδομένων που αντιμετωπίζουμε, ενισχύοντας την ακρίβεια και την αξιοπιστία των προβλέψεών μας.

Για να αξιολογηθεί η αποτελεσματικότητα των μεθόδων πρόβλεψης, απαραίτητη προϋπόθεση είναι η δοκιμή τους σε μεγάλο πλήθος δεδομένων με στόχο τη σύγκριση των αποτελεσμάτων που προκύπτουν από αυτές. Αυτή η πρακτική εδραιώνεται στη βιβλιογραφία μέσω των διαγωνισμών πρόβλεψης [6], όπου διάφορες μεθόδοι εφαρμόζονται σε συγκεκριμένα σύνολα δεδομένων προερχόμενα από ένα ευρή φάσμα εφαρμογών και τα οποία διαθέτουν διαφορετικά χαρακτηριστικά και ιδιαιτερότητες. Στη συνέχεια, τα αποτελέσματα αξιολογούνται συγκριτικά προκειμένου να διαπιστωθεί ποια μέθοδος παρουσιάζει καλύτερη επίδοση και υπό ποιες συνθήκες. Αυτές οι διαδικασίες αξιολόγησης διαδραματίζουν καθοριστικό ρόλο στην καθιέρωση δημοφιλών συνόλων δεδομένων και μεθόδων ως ¨σημεία αναφοράς¨ (benchmarks) στον τομέα των προβλέψεων. Ως αποτέλεσμα, η αξιολόγηση νέων μεθόδων επικεντρώνεται σε μεγάλο βαθμό στην απόδοσή τους σε αυτά τα δημοφιλή σύνολα δεδομένων.

Ωστόσο, παρά την ευρεία αποδοχή των διαγωνισμών πρόβλεψης ως μέσου αξιολόγησης, η ερευνητική κοινότητα εκφράζει ανησυχίες σχετικά με την επάρκεια των δεδομένων που χρησιμοποιούνται [7] [8] [9]. Κάποιες από τις αδυναμίες που έχουν διερευνηθεί συνδέονται κυρίως με την περιορισμένη ποικιλομορφία των δεδομένων και το σχετικά μικρό τους μέγεθος. Αυτό μπορεί να περιορίσει την ικανότητά τους να αντικατοπτρίσουν πλήρως τις συνθήκες και προκλήσεις που αντιμετωπίζουν οι οργανισμοί και οι επιχειρήσεις στον πραγ-

ματικό κόσμο [10].

Ταυτόχρονα, αυτή η προσέγγιση αντιμετωπίζει προκλήσεις όταν ο ερευνητής επιθυμεί να αναπτύξει και να αξιολογήσει μια μέθοδο πρόβλεψης που αφορά χρονοσειρές με πολύ συγκεκριμένα χαρακτηριστικά για πιο εξειδικευμένα προβλήματα ή εφαρμογές. Η εύρεση ικανοποιητικής, για τις ανάγκες επικύρωσης ενός μοντέλου, ποσότητας χρονοσειρών με τις ιδιότητες που επιθυμεί, μπορεί να είναι δύσκολη έως και μη εφικτή. Αν και υπάρχει πληθώρα πηγών δεδομένων στο διαδίκτυο, αυτές σπανίως παρέχουν αρκετές πληροφορίες για τα συγκεκριμένα χαρακτηριστικά των χρονοσειρών που ο ερευνητής ενδεχομένως να επιθυμεί να αξιολογήσει. Αυτό μπορεί να οδηγήσει τους ερευνητές σε συλλογή δεδομένων με δικές τους μεθόδους, προκειμένου να καλύψουν τα συγκεκριμένα χαρακτηριστικά που τους ενδιαφέρουν. Αυτό έχει ως αποτέλεσμα τη δημιουργία διαφορετικών συνόλων δεδομένων, τα οποία, τελικά, δεν θα έχουν υποβληθεί σε ίδιο επίπεδο δοκιμασίας με τα πιο ευρέως αναγνωρισμένα σύνολα δεδομένων που χρησιμοποιούνται σε διαγωνισμούς. Επιπλέον, η επαλήθευση των αποτελεσμάτων τους μπορεί να απαιτεί περισσότερο χρόνο και προσπάθεια λόγω της πιο περίπλοκης διαδικασίας συγκέντρωσης και επεξεργασίας των δεδομένων [11].

Έχοντας υπόψη τους περιορισμούς και τις αδυναμίες που έχουν προαναφερθεί, η παρούσα διπλωματική εργασία έχει ως ευρύτερο στόχο την παροχή νέων συνόλων δεδομένων στην κοινότητα των τεχνικών προβλέψεων. Το σύστημα που έχει υλοποιηθεί προσπαθεί να παρέχει τα εργαλεία και την υποδομή ώστε να καθιερωθεί ως ένας ποιοτικός και αξιόπιστος πόρος προς υποστήριξη του σχετικού ερευνητικού πεδίου.

Το FOREcasting & Strategy Unit Data Collection, ή αλλιώς FOREDeCk, είναι μια διαδικτυακή πλατφόρμα σχεδιασμένη για να παρέχει εύκολη πρόσβαση σε ένα ευρύ φάσμα δεδομένων. Το σύστημα προσφέρει πάνω από ένα εκατομμύριο χρονοσειρές από διάφορους τομείς και με ποικιλία χαρακτηριστικών. Οι χρονοσειρές αυτές έχουν συλλεχθεί από πολλές και διαφορετικές αξιόπιστες πηγές, καλύπτοντας τομείς όπως βιομηχανία, υπηρεσίες, τουρισμός, εισαγωγές και εξαγωγές, δημογραφικά στοιχεία, εκπαίδευση, εργασία, κυβέρνηση, νοικοκυριά, ομόλογα, μετοχές, ασφάλειες, δάνεια, μεταφορές, φυσικοί πόροι, περιβάλλον, κ.λπ. Προτού ενσωματωθούν στην πλατφόρμα, οι χρονοσειρές αυτές υποβλήθηκαν σε αυστηρή διαδικασία φιλτραρίσματος για τη διασφάλιση της ποιότητας και της αξιοπιστίας του συστήματος. Το σύστημα παρέχει πολλαπλά φίλτρα για την εύρεση και ανάκτηση χρονοσειρών, είτε βάση της συχνότητας ή της κατηγορίας των δεδομένων, είτε βάση συγκεκριμένων χαρακτηριστικών. Αυτό επιτρέπει την επιλεκτική πρόσβαση σε σύνολα δεδομένων που ανταποκρίνονται στις ανάγκες και τις προδιαγραφές του εκάστοτε χρήστη. Επιπλέον, παρέχεται η δυνατότητα λήψης των χρονοσειρών, συμπεριλαμβανομένων όλων των σχετικών μεταδεδομένων και χαρακτηριστικών τους.

Παράλληλα, σε περίπτωση που οι ανάγκες του χρήστη δεν καλύπτονται από τα διαθέσιμα δεδομένα, το σύστημα του δίνει την δυνατότητα να δημιουργήσει προσαρμοσμένες χρονοσειρές βάση των χαρακτηριστικών που αυτός επιθυμεί. Αυτό γίνεται με την ενσωμάτωση στο σύστημα της γεννήτριας παραγωγής χρονοσειρών GRATIS [5], η οποία επιτρέπει την παρα-

γωγή μεγάλου πλήθους τεχνητών χρονοσειρών με ελεγχόμενα χαρακτηριστικά.

Ταυτόχρονα, το FOREDeCk παρέχει ένα περιβάλλον οπτικοποίησης των χρονοσειρών στον δισδιάστατο χώρο που έχει οριστεί στο πλαίσιο της μελέτης των Kang et al., 2017 [9] μέσω της τεχνικής Principal Component Analysis (PCA) και τον σχετικών χαρακτηριστικών. Αυτή η προσέγγιση διευκολύνει τη σύγκριση και την καλύτερη κατανόηση της ποικιλομορφίας των διαφορετικών χρονοσειρών (πραγματικών και τεχνητών), ενισχύοντας τη δυνατότητα των χρηστών να επιλέξουν τα κατάλληλα δείγματα για τις αναλύσεις τους και προσφέροντας έναν ολοκληρωμένο και διαφανή τρόπο ανάκτησης και εξερεύνησης των δεδομένων.

Αξίζει να σημειωθεί πως στα πλαίσια της υποστήριξης της ερευνητικής κοινότητας, το σύστημα έχει αξιοποιηθεί για την υποστήριξη του διαγωνισμού προβλέψεων M4 [12], καθώς το σύνολο των 100.000 χρονοσειρών του διαγωνισμού προήλθαν μέσω τυχαίας δειγματοληψία στο σύνολο των δεδομένων του FOREDeCk.

1.2 Οργάνωση της Εργασίας

Το δεύτερο κεφάλαιο αναλύει το βασικό υπόβαθρο του κλάδου των τεχνικών προβλέψεων, προσφέροντας τις απαραίτητες θεωρητικές βάσεις για την κατανόηση του αντικειμένου της παρούσας διπλωματικής εργασίας. Επικεντρώνεται στην ανάλυση των χρονοσειρών, παρουσιάζοντας τα βασικά τους χαρακτηριστικά και εννοιολογικό πλαίσιο. Επιπλέον, γίνεται αναφορά, μέσα από την βιβλιογραφία, σε συγκεκριμένες μεθόδους κατηγοριοποίησης χρονοσειρών και παραγωγής τεχνητών χρονοσειρών, τα οποία αποτελούν βασικό κομμάτι του συστήματος που αναπτύχθηκε στα πλαίσια της παρούσας εργασίας.

Στο τρίτο κεφάλαιο εξετάζεται η πρώτη και βασικότερη φάση της ανάπτυξης του συστήματος - η διαδικασία συλλογής δεδομένων. Αναλύονται οι προδιαγραφές που καθορίστηκαν για την αξιολόγηση και επιλογή των πηγών δεδομένων, με στόχο την διασφάλιση της ποιότητας και της ευελιξίας του συστήματος. Στη συνέχεια, εξετάζονται τα βασικά κριτήρια και οι μεθόδοι για το αποτελεσματικό φιλτράρισμα και την επεξεργασία των δεδομένων. Τέλος, παρουσιάζεται η διαδικασία που ακολουθήθηκε για την τελική ομαδοποίηση των δεδομένων ώστε να εξασφαλιστεί η αποδοτικότητα του συστήματος.

Στο τέταρτο κεφάλαιο, παρουσιάζεται η μελέτη που έγινε για την υλοποίηση του συστήματος. Αναλύονται οι απαιτήσεις του συστήματος και παρουσιάζεται η αρχιτεκτονική του, καθώς και οι τεχνολογίες που χρησιμοποιήθηκαν για την υλοποίησή του. Επιπλέον, αναλύονται οι διάφορες λειτουργίες του συστήματος, όπως η συλλογή, η ανάλυση και η οπτικοποίηση των δεδομένων χρονοσειρών. Τέλος, παρουσιάζονται οι διάφορες δοκιμές που πραγματοποιήθηκαν για την επαλήθευση της λειτουργικότητας του συστήματος.

Στο πέμπτο και τελευταίο κεφάλαιο, γίνεται μια σύντομη ανακεφαλαίωση και αναφέρονται ενδεχόμενες επεκτάσεις και βελτιώσεις της παρούσας εργασίας στο μέλλον.

Κεφάλαιο 2

Ποιοτικά Χαρακτηριστικά και Κατηγοριοποίηση Χρονοσειρών

2.1 Εισαγωγή στις χρονοσειρές

Οι χρονοσειρές (time series) είναι μια ακολουθία διαχρονικών παρατηρήσεων που εκφράζουν την εξέλιξη ενός μεγέθους στον χρόνο και συνήθως λαμβάνονται με σταθερό βήμα παρατήρησης. Συναντώνται σε πολλές επιστήμες, όπως η οικονομία, η μετεωρολογία, η ενέργεια, η ιατρική, και χρησιμοποιούνται ευρέως για την περιγραφή μεγεθών τους οποίους η ιστορική πληροφορία μπορεί να έχει κάποια χρήση. Από την ημερήσια τιμή μιας μετοχής στο χρηματιστήριο και τον ωριαίο αριθμό πελατών σε ένα κατάστημα, μέχρι την παραγωγή ηλεκτρικής ενέργειας από αιολικούς σταθμούς, οι χρονοσειρές είναι σημαντικό εργαλείο για τη λήψη αποφάσεων σε πολλές επιστημονικές περιοχές.

Η ανάλυση χρονοσειρών περιλαμβάνει την ανάλυση αυτών των δεδομένων με στόχο την εξαγωγή σημαντικών πληροφοριών, τάσεων και χαρακτηριστικών που εμφανίζονται. Μέσω της ανάλυσής τους, μπορούμε να ανιχνεύσουμε τάσεις, κύκλους, ακρότατα και άλλα χαρακτηριστικά που αποκαλύπτουν πληροφορίες για τη φύση και τη συμπεριφορά του μεγέθους που μελετούμε. Για παράδειγμα, στην οικονομία, οι χρονοσειρές μπορούν να χρησιμοποιηθούν για να αναλύσουν τις τάσεις της αγοράς, την πρόβλεψη των οικονομικών επιδόσεων ενός κράτους ή την παρακολούθηση των αναδιατάξεων στην παγκόσμια οικονομία. Στη μετεωρολογία, αναλύονται για την πρόβλεψη του καιρού, την ανίχνευση κλιματικών μοτίβων και τη μελέτη των αλλαγών του κλίματος. Στον τομέα της ενέργειας, οι χρονοσειρές χρησιμοποιούνται για την ανάλυση της κατανάλωσης ενέργειας, την πρόβλεψη της ζήτησης, την αξιολόγηση της απόδοσης των ενεργειακών συστημάτων, καθώς και την ανίχνευση ανωμαλιών στην παραγωγή και διανομή ενέργειας.

2.2 Συχνότητα Χρονοσειρών

Η περίοδος ή συχνότητα μιας χρονοσειράς αναφέρεται στην απόσταση μεταξύ των διαδοχικών παρατηρήσεων της. Συχνά, η περιοδικότητα παρέχεται κατευθείαν από την πηγή δεδομένων και γιαυτό χρησιμοποιείται για τον βασικό διαχωρισμό των χρονοσειρών σε κατηγορίες:

- **Ετήσιες** χρονοσειρές: μία παρατήρηση ανά έτος

- **Τριμηνιαίες** χρονοσειρές: μία παρατήρηση ανά τρίμηνο
- **Μηνιαίες** χρονοσειρές: μία παρατήρηση ανά μήνα
- **Εβδομαδιαίες** χρονοσειρές: μία παρατήρηση ανά εβδομάδα
- **Ημερήσιες** χρονοσειρές: μία παρατήρηση ανά ημέρα
- **Ωριαίες** χρονοσειρές: μία παρατήρηση ανά ώρα

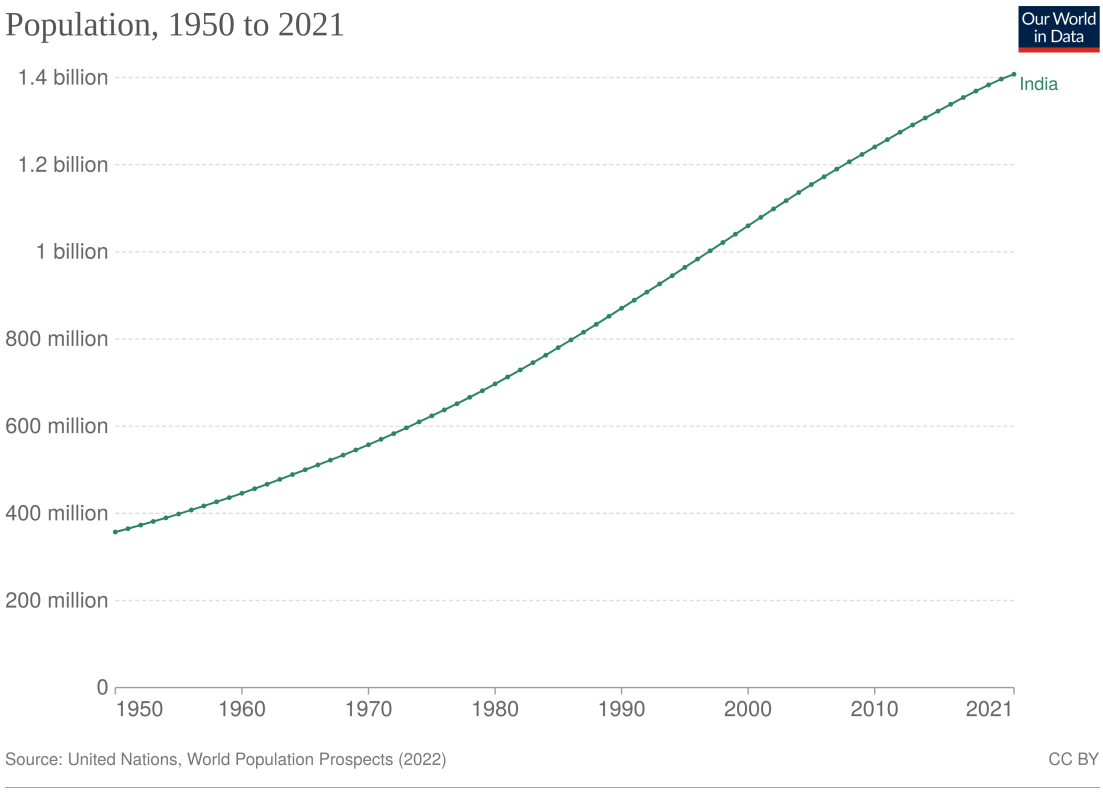
Η κατηγοριοποίηση των χρονοσειρών με βάση τη διαφορετική συχνότητα παρατήρησης αποτελεί έναν σημαντικό διαχωρισμό. Πέραν της διαφοράς στη συχνότητα, οι χρονοσειρές που ανήκουν σε διαφορετικές κατηγορίες επιδεικνύουν επίσης διαφορές σε άλλα σημαντικά χαρακτηριστικά. Για παράδειγμα, οι μηνιαίες χρονοσειρές μπορεί να εμφανίζουν εποχιακή συμπεριφορά, ενώ οι χρονιαίες χρονοσειρές μπορεί να μην την έχουν. Αυτές οι διαφορές πηγάζουν από τη φύση των χρονοσειρών, καθώς η εξάρτηση μεταξύ των παρατηρήσεων αποκτά διαφορετική έννοια ανάλογα με τη συχνότητα καταγραφής. Επομένως, προτού προχωρήσουμε στην ανάλυση μιας χρονοσειράς, είναι απαραίτητο να αναγνωρίσουμε την κατηγορία στην οποία ανήκει.

Αν και η εποχιακότητα της χρονοσειράς είναι ένα από τα χαρακτηριστικά που συνήθως μας δίνεται από την πηγή δεδομένων μας, υπάρχει μια πληθώρα άλλων χαρακτηριστικών που μπορούμε να υπολογίσουμε και να αξιοποιήσουμε ανάλογα με το αντικείμενο της έρευνας μας.

2.3 Ποιοτικά χαρακτηριστικά χρονοσειρών

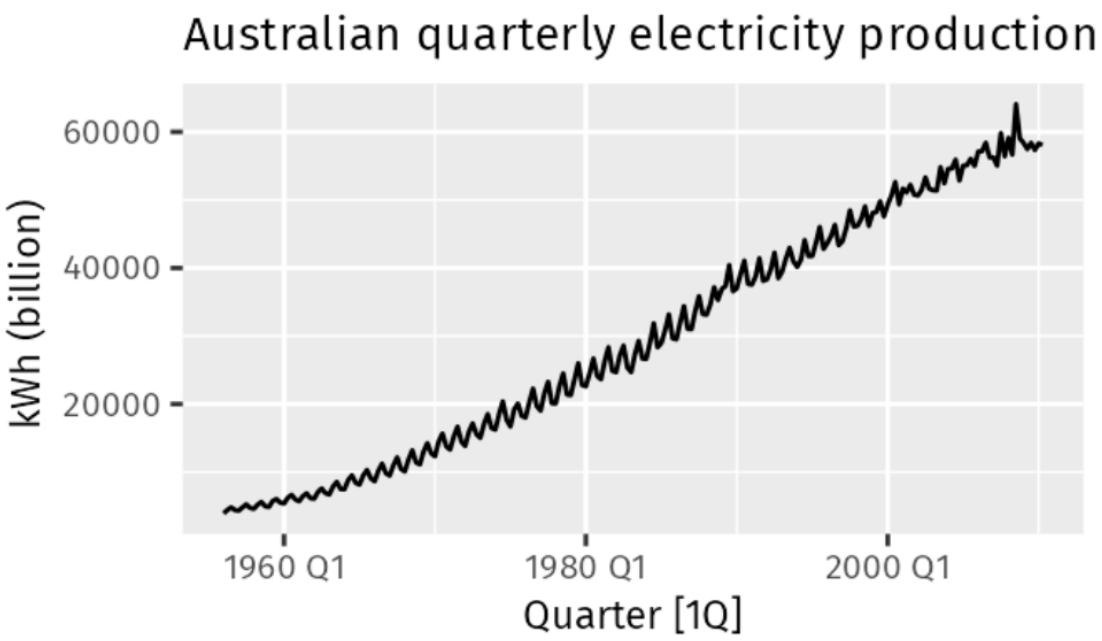
Η μελέτη μίας χρονοσειράς ξεκινάει με την επισκόπηση του γραφήματος της στο πεδίο του χρόνου. Τα βασικά ποιοτικά χαρακτηριστικά που βλέπουμε κατά την οπτική μελέτη μίας χρονοσειράς είναι η τάση, η εποχικότητα, η κυκλικότητα (κύκλος) και η τυχαιότητα (θόρυβος/διακύμανση). Παρακάτω δίνεται μία σύντομη περιγραφή τους.

Τάση: Η τάση (trend) σε μια χρονοσειρά αναφέρεται στο μακροπρόθεσμο μοτίβο μεταβολής των τιμών της χρονοσειράς σε σχέση με τον χρόνο. Αυτή μπορεί να είναι θετική (upward trend), αρνητική (downward trend) ή σταθερή/μηδενική, ανάλογα με το αν η χρονοσειρά παρουσιάζει αυξανόμενο, φθίνον ή σταθερό μακροπρόθεσμο μοτίβο. Για να αποδοθούν έγκυρα συμπεράσματα για την ύπαρξη τάσης, απαιτείται ένας επαρκής αριθμός παρατηρήσεων και κατάλληλο χρονικό διάστημα για ανάλυση. Στο Σχήμα 2.1 απεικονίζεται ο πληθυσμός ανά έτος για την Ινδία, όπου μπορούμε να παρατηρήσουμε την έντονη ανοδική τάση.



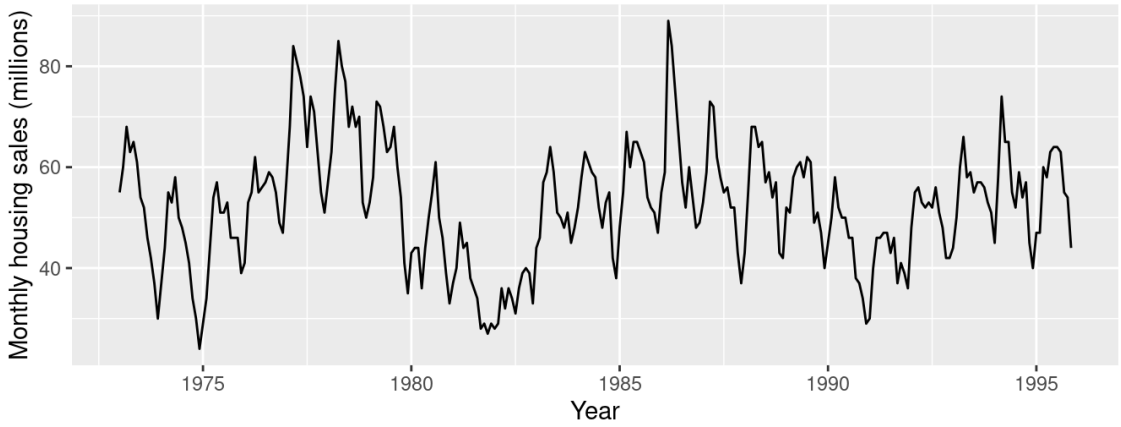
Σχήμα 2.1: *Ο πληθυσμός ανά έτος για την Ινδία [1]*

Εποχικότητα: Η εποχικότητα (seasonality) είναι ένα μοτίβο περιοδικών διακυμάνσεων στις τιμές μιας χρονοσειράς. Αναδεικνύει τη συσχέτιση ανάμεσα στις παρατηρήσεις που ανήκουν σε αντίστοιχες χρονικές περιόδους και μπορεί να εκφράσει την επίδραση των εποχικών παραγόντων στη χρονοσειρά. Στο Σχήμα 2.2 απεικονίζεται η τριμηνιαία παραγωγή ηλεκτρικής ενέργειας στην Αυστραλία, όπου μπορούμε να παρατηρήσουμε έντονη θετική τάση και έντονη εποχικότητα.



Σχήμα 2.2: *Η τριμηνιαία παραγωγή ηλεκτρικής ενέργειας στην Αυστραλία [2]*

Κυκλικότητα: Η κυκλικότητα (cyclical) αναφέρεται στη μακροχρόνια και κυματοειδή μεταβολή που παρουσιάζεται σε μια χρονοσειρά λόγω εξωτερικών παραγόντων. Αυτή η μεταβολή εμφανίζει ανόδους και πτώσεις σε μη σταθερή συχνότητα και δεν εμφανίζεται συστηματικά σε συγκεκριμένα χρονικά διαστήματα. Ο κύκλος δεν είναι περιοδικός, καθώς η διάρκεια και η έντασή του μπορεί να διαφέρουν από περίοδο σε περίοδο. Παράλληλα, η κυκλικότητα είναι παρατηρήσιμη αλλά δεν μπορεί να προβλεφθεί εύκολα. Συνήθη παραδείγματα κυκλικότητας παρουσιάζονται σε οικονομικές χρονοσειρές, όπως το Ακαθάριστο Εγχώριο Προϊόν (ΑΕΠ) και την χρηματιστηριακή αγορά, καθώς αναδεικνύουν τις αυξομειώσεις της οικονομίας μεταξύ περιόδων κρίσης. Στο Σχήμα 2.3 απεικονίζονται οι μηνιαίες πωλήσεις νέων μονοκατοικιών στις ΗΠΑ, όπου μπορούμε να παρατηρήσουμε έντονη εποχικότητα μέσα σε κάθε χρόνο, καθώς και έντονη κυκλική συμπεριφορά με περίοδο περίπου 6-10 ετών.



Σχήμα 2.3: *Μηνιαίες πωλήσεις νέων μονοκατοικιών στις ΗΠΑ [2]*

Τυχαιότητα: Η τυχαιότητα αναφέρεται στις μεταβολές που δεν μπορούν να ερμηνευθούν μέσω των υπολοίπων ποιοτικών χαρακτηριστικών της χρονοσειράς. Περιλαμβάνει τις μη κανονικές διακυμάνσεις που που μπορεί να παρουσιάζονται είτε ως μια εντελώς τυχαία μεταβλητή (με τη στατιστική έννοια), είτε ως ασυνέχειες που συνδέονται με κάποιο εξαιρετικό γεγονός (special event). Οι ασυνέχειες μπορεί να έχουν περιοδικό (ακραίες τιμές) ή μόνιμο (αλλαγές επιπέδου) χαρακτήρα και εμφανίζονται ως απότομες αλλαγές στο πρότυπο συμπεριφοράς της χρονοσειράς. Οι ακραίες τιμές (outliers) είναι απρόβλεπτες, απομονωμένες παρατηρήσεις που έχουν μικρή χρονική διάρκεια και μπορεί να οφείλονται, για παράδειγμα, φυσικές καταστροφές ή απρόβλεπτες αλλαγές στις συνθήκες της αγοράς, με αποτέλεσμα την επαναφορά της χρονοσειράς μετά το πέρας τους στην \κανονικότητα" της. Από την άλλη. οι αλλαγές επιπέδου (level-shifts) εμφανίζονται ως μόνιμες αλλαγές στο μέσο επίπεδο των τιμών της χρονοσειράς που μπορεί να οφείλονται, για παράδειγμα, σε αλλαγές στην οικονομική κατάσταση, τις συνθήκες αγοράς ή την πολιτική της εταιρείας.

2.4 Αποσύνθεση Χρονοσειρών

Η αποσύνθεση είναι η διαδικασία ανάλυσης μιας χρονοσειράς στις επιμέρους συνιστώσες που την αποτελούν. Μέσω αυτής της διαδικασίας, αναδεικνύονται τα χαρακτηριστικά που συμβάλλουν στην κατανόηση της ιστορικής συμπεριφοράς της χρονοσειράς και επιτρέπουν την πρόβλεψη μελλοντικών τιμών. Για την απομόνωση και τον ποσοτικό προσδιορισμό των συνιστωσών, χρησιμοποιούμε μια αναπαράσταση της χρονοσειράς ως μια συνάρτηση των βασικών χαρακτηριστικών της:

$$Y_t = f(S_t + T_t + C_t + R_t) \tag{2.1}$$

Y_t : παρατήρηση την χρονική στιγμή t

S_t : συνιστώσα της εποχικότητας την χρονική στιγμή t

T_t : συνιστώσα της τάσης την χρονική στιγμή t

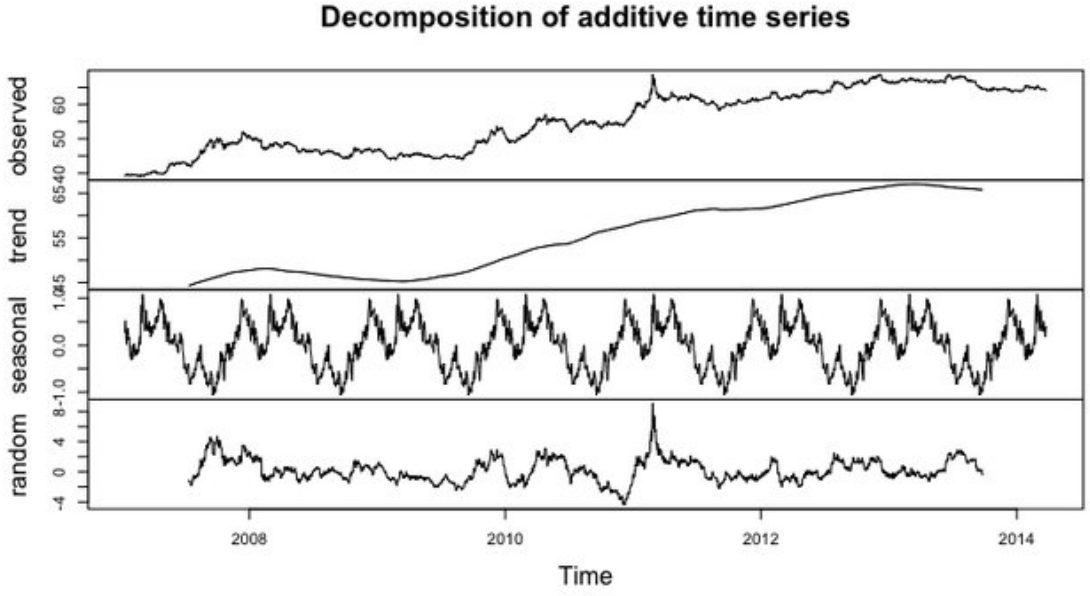
C_t : συνιστώσα του κύκλου την χρονική στιγμή t

R_t : συνιστώσα της τυχαιότητας την χρονική στιγμή t

Η συναρτησιακή σχέση f φανερώνει τον τρόπο με τον οποίο οι παρατηρήσεις της χρονοσειράς προσδιορίζονται από τις συνιστώσες και οι συνηθέστερες μορφές που έχει είναι η προσθετική (2.2) και η πολλαπλασιαστική (2.3). Στο προσθετικό μοντέλο όλες οι συνιστώσες είναι ανεξάρτητες μεταξύ τους, ενώ στο πολλαπλασιαστικό τα τέσσερα αυτά χαρακτηριστικά δεν είναι αναγκαστικά ανεξάρτητα και μπορούν να επηρεάσουν το ένα το άλλο.

$$Y_t = S_t + T_t + C_t + R_t \tag{2.2}$$

$$Y_t = S_t * T_t * C_t * R_t \tag{2.3}$$



Σχήμα 2.4: *Αποσύνθεση χρονοσειράς στα βασικά της χαρακτηριστικά. Στο πρώτο διάγραμμα παρουσιάζεται η χρονοσειρά με βάση τις τιμές παρατήρησης, πριν την εφαρμογή της μεθόδου της αποσύνθεσης. Το δεύτερο διάγραμμα αναπαριστά την τάση της χρονοσειράς, το τρίτο της εποχικότητα της και το τελευταίο την συνιστώσα της τυχαιότητας [3]*

2.5 Στατιστική ανάλυση χρονοσειρών

Πέραν της γραφικής απεικόνισης της χρονοσειράς και της εξαγωγής των πρώτων συμπερασμάτων επί αυτής, είναι χρήσιμο να γίνει και μία βασική στατιστική ανάλυση. Οι στατιστικοί δείκτες που μπορούν να υπολογισθούν σε δεδομένη χρονοσειρά Y μεγέθους n παρατηρήσεων είναι [13]:

Μέση τιμή: Η μέση τιμή (average) ορίζεται ως ο γραμμικός μέσος όρος των τιμών όλων των παρατηρήσεων της χρονοσειράς και μας παρέχει μια εκτίμηση του επιπέδου γύρω από το οποίο κυμαίνονται οι πραγματικές τιμές της χρονοσειράς.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{2.4}$$

Μέγιστη και ελάχιστη τιμή: Η μέγιστη και ελάχιστη τιμή (maximum and minimum) αποτελούν τις ακραίες παρατηρήσεις της χρονοσειράς. Αυτές οι τιμές παρέχουν πληροφορίες σχετικά με τις ακραίες αλλαγές που μπορεί να παρουσιάζει η χρονοσειρά, και μας δίνουν μια εκτίμηση της διακύμανσης των δεδομένων, καθώς αντικατοπτρίζουν την έκταση των διαφορών μεταξύ των τιμών της χρονοσειράς. Επιπλέον, η ύπαρξη ακραίων τιμών μπορεί να υποδηλώνει την παρουσία τυχαιότητας στα δεδομένα, ειδικά όταν αποτελούν απρόβλεπτες αποκλίσεις από τον τυπικό μέσο όρο.

Τυπική απόκλιση: Η τυπική απόκλιση (standard deviation) εκφράζει το βαθμό διασποράς των παρατηρήσεων γύρω από την μέση τιμή. Μια χαμηλότερη τιμή δείχνει πως οι παρατηρήσεις τείνουν να συγκεντρώνονται πιο κοντά στην μέση τιμή.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \mu)^2}{n}} \tag{2.5}$$

Διακύμανση: Η διακύμανση (variance) ορίζεται ως το τετράγωνο της τυπικής απόκλισης και δείχνει πόσο απομακρύνονται οι τιμές της χρονοσειράς από τη μέση τιμή της.

Συνδιακύμανση: Η συνδιακύμανση (covariance) ορίζεται ως το μέτρο συσχέτισης μεταξύ δύο διακριτών τυχαίων μεταβλητών X και Y που εξετάζει πώς αυτές μεταβάλλονται ανάλογα (θετική συνδιακύμανση), αντιστρόφως ανάλογα (αρνητική συνδιακύμανση) ή είναι ανεξάρτητες (μηδενική συνδιακύμανση). Κατά τη μελέτη των χρονοσειρών, υπολογίζουμε την συνδιακύμανση των δεδομένων ανάλογα με τον αύξοντα αριθμό της χρονικής περιόδου.

$$Cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}[(X_i - \mu_X)(Y_i - \mu_Y)] \tag{2.6}$$

Συντελεστής αυτοσυσχέτισης: Ο συντελεστής αυτοσυσχέτισης (autocorrelation coefficient) εκφράζει τη συσχέτιση μεταξύ παρατηρήσεων της ίδιας μεταβλητής με χρονική υστέρηση k περιόδου, δηλαδή μας δείχνει κατά πόσο η τιμή της χρονοσειράς σε μία περίοδο εξαρτάται από την τιμή της παρατήρησης k περιόδων πίσω. Λαμβάνει τιμές στο διάστημα [0, 1], με τιμή κοντά στο μηδέν να δηλώνει μηδενική συσχέτιση, ενώ τιμή κοντά στη μονάδα δηλώνει μεγάλη συσχέτιση.

$$ACF_k = \frac{\sum_{i=1+k}^{n}[(Y_i - \mu)(Y_{i-k} - \mu)]}{\sum_{i=1}^{n}(Y_i - \mu)^2} \tag{2.7}$$

2.6 Κατηγοριοποίηση χρονοσειρών

Ανάλογα με τους στόχους και το ερευνητικό πεδίο της εκάστοτε έρευνας έχουν αναπτυχθεί ποικίλα χαρακτηριστικά για την περιγραφή, κατηγοριοποίηση και οπτικοποίηση χρονοσειρών. Δεν υπάρχει μοναδική ¨καλύτερη¨ αναπαράσταση των χαρακτηριστικών μιας χρονοσειράς που να θεωρείται η βέλτιστη για κάθε περίπτωση, το πιο χρήσιμο μέτρο εξαρτάται από τα δεδομένα και τις ερωτήσεις που τίθενται για αυτά[14]. Ωστόσο, η ανάλυση ενός ποικιλόμορφου συνόλου χρονοσειρών παραμένει ένα πολύπλοκο έργο, κυρίως λόγω της άμεσης συσχέτισης των δεδομένων με την πάροδο του χρόνου και των διαφορών στο μήκος τους. Μια έξυπνη προσέγγιση για την αντιμετώπιση αυτού του προβλήματος είναι να οριστούν ενδεικτικές και ανεξάρτητες στατιστικές για τη μέτρηση των βασικών τους χαρακτηριστικών, επιτρέποντας σε κάθε σειρά να αναπαρασταθεί ως ένα στατικό σημείο σε έναν χώρο χαρακτηριστικών υψηλής διάστασης (high dimensional feature space) [15]. Αυτή η πρακτική έχει χρησιμοποιηθεί ευρέως για την αποτελεσματική κατηγοριοποίηση και ομαδοποίηση χρονοσειρών, με τον αριθμό και τον τύπο των χαρακτηριστικών να εξαρτώνται από την εφαρμογή που εξετάζεται.

Στην προσπάθεια τους να αναλύσουν και να βγάλουν καλύτερα συμπεράσματα για ένα από τα πλέον χρησιμοποιούμενα σετ δεδομένων για την αξιολόγηση μεθόδων και τεχνικών πρόβλεψης, τον διαγωνισμό προβλέψεων M3 [16], οι Kang et al., 2017 [9] όρισαν ένα σύνολο έξι χαρακτηριστικών τα οποία όταν ομαδοποιηθούν με χρήση της τεχνικής Principal Components Analysis [17] μπορούν να χρησιμοποιηθούν για την απεικόνιση των χρονοσειρών σε ένα δισδιάστατο χώρο (2-dimensional instance space) χωρίς σημαντική απώλεια πληροφορίας.

2.6.1 Τα χαρακτηριστικά των χρονοσειρών

Στην έρευνα τους, οι Kang et al., 2017 [9] ποσοτικοποίησαν τα ποιοτικά χαρακτηριστικά των χρονοσειρών ως προς την έντασή τους και επιπλέον χρησιμοποίησαν την αυτοσυσχέτιση των παρατηρήσεων, την περιοδικότητα της χρονοσειράς και την στασιμότητα, ή αλλιώς την σταθερότητα της διακύμανσης. Τα χαρακτηριστικά αυτά δεν εξαρτώνται από την κλίμακα της χρονοσειράς, οπότε είναι κατάλληλα για εφαρμογή σε ένα μεγάλο και ποικίλο σύνολο χρονοσειρών. Πιο συγκεκριμένα, σε χρονοσειρά X_t που αποσυντίθεται με χρήση της προσθετικής αποσύνθεσης STL στις συνιστώσες τάσης (T_t), εποχιακότητας (S_t) και τυχαιότητας (R_t), τα μεγέθη που αναφέρθηκαν μπορούν να καθοριστούν ως εξής:

Ένταση τυχαιότητας F_1

Χρησιμοποιείται για τη μέτρηση της «προβλεψιμότητας» (τυχαιότητας) της χρονοσειράς. Μια σχετικά μικρή τιμή υποδηλώνει ότι η χρονοσειρά περιέχει περισσότερο σήμα και είναι πιο προβλέψιμη. Από την άλλη, μια σχετικά μεγάλη τιμή υποδηλώνει μεγαλύτερη αβεβαιότητα για το μέλλον, και ως εκ τούτου ότι η χρονοσειρά είναι πιο δύσκολο να προβλεφθεί. Η ένταση τυχαιότητας ισοδυναμεί με την εντροπία της χρονοσειράς (spectral entropy) η οποία και υπολογίζεται ως:

$$F_1 = \int_{-\pi}^{\pi} f_x(\lambda) log f_x(\lambda) d\lambda \tag{2.8}$$

όπου $f_x(\lambda)$ είναι η πυκνότητα εντροπίας (spectral density).

Ένταση τάσης F_2

Χρησιμοποιείται για τη μέτρηση μακροπρόθεσμων μεταβολών στο μέσο επίπεδο (mean level) της χρονοσειράς. Η ένταση τάσης (strength of trend) υπολογίζεται ως ο λόγος της διακύμανσης του τυχαίου παράγοντα προς τη διακύμανση του παράγοντα τάσης:

$$F_2 = 1 - \frac{var(R_t)}{var(X_t - S_t)} \tag{2.9}$$

Ένταση εποχιακότητας F_3

Χρησιμοποιείται για τη μέτρηση της επίδρασης εποχιακών παραγόντων, όπως το τρίμηνο ή ο μήνας του έτους. Η ένταση εποχιακότητας (strength of seasonality) υπολογίζεται ως ο λόγος της διακύμανσης του τυχαίου παράγοντα προς τη διακύμανση του εποχιακού παράγοντα:

$$F_3 = 1 - \frac{var(R_t)}{var(X_t - T_t)} \tag{2.10}$$

Συχνότητα F_4

Υποδεικνύει το μήκος της περιοδικότητας της χρονοσειράς (seasonal period). Είναι $F_4 = 24$ για ωριαία δεδομένα, $F_4 = 7$ για ημερήσια, $F_4 = 52$ για εβδομαδιαία, $F_4 = 4$ για τριμηνιαία, $F_4 = 12$ για μηνιαία, και $F_4 = 1$ σε περίπτωση ετήσιων δεδομένων. Η περιοδικότητα της χρονοσειράς είναι ένα από τα χαρακτηριστικά που συνήθως μας δίνεται από την πηγή δεδομένων μας.

Ένταση αυτοσυσχέτισης F_5

Μετρά τη γραμμική σχέση μεταξύ μιας χρονοσειράς x_t και της σειράς με καθυστέρηση ενός βήματος x_{t-1}. Η ένταση αυτοσυσχέτισης (first order autocorrelation) ταυτίζεται με τον συντελεστή αυτοσυσχέτισης των παρατηρήσεων πρώτου βαθμού $AR(1)$. Μια υψηλότερη απόλυτη τιμή υποδηλώνει ότι οι μελλοντικές τιμές του x_t εξαρτώνται περισσότερο από την προηγούμενη τιμή, το οποίο σε κάποιο βαθμό, δείχνει την προβλεψιμότητα μιας χρονοσειράς.

Στασιμότητα F_6

Υπολογίζεται ο βέλτιστος συντελεστής λ μετασχηματισμού Box-Cox (Optimal Box-Cox transformation parameter) στο διάστημα $[0,1]$ μέσω μεγιστοποίησης της πιθανοφάνειας και αυτός χρησιμοποιείται ως το μέτρο της στασιμότητας.

Αυτά τα έξι χαρακτηριστικά επιτρέπουν σε οποιαδήποτε χρονοσειρά, οποιουδήποτε μήκους, να συνοψιστεί ως ένα διάνυσμα χαρακτηριστικών $F = (F_1, F_2, F_3, F_4, F_5, F_6)$.

2.6.2 Principal Component Analysis (PCA)

Η Ανάλυση Κύριων Συνιστωσών (αγγλ. Principal Component Analysis - PCA) [17] είναι μια ευρέως χρησιμοποιούμενη τεχνική που χρησιμοποιείται για τη μείωση των διαστάσεων (dimensionality reduction) ενός συνόλου δεδομένων. Με την PCA, τα δεδομένα απλοποιο-

ύνται και αντιστοιχίζονται σε ένα χώρο λιγότερων διαστάσεων, διατηρώντας όμως τις πιο σημαντικές πληροφορίες.

Αυτό μπορεί να αποδειχθεί εξαιρετικά χρήσιμο κατά την εξέταση των χαρακτηριστικών ενός συνόλου δεδομένων. Παρόλο που οι σειρές εξακολουθούν να αναπαρίστανται από ένα διάνυσμα με n χαρακτηριστικά - αρχικά οπτικοποιημένο σε έναν n-διάστατο χώρο - πλέον γίνεται εφικτή η άμεση παρατήρηση τους στον παρατηρήσιμο χώρο, δηλαδή, έως τρεις διαστάσεις, στις οποίες περιορίζεται η ανθρώπινη αντίληψη.

Αυτό επιτυγχάνεται μετασχηματίζοντας τις αρχικές μεταβλητές σε ένα νέο σύνολο ανεξάρτητων μεταβλητών, που ονομάζονται κύριες συνιστώσες. Αυτές οι συνιστώσες είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών και κατατάσσονται με βάση το ποσοστό της διακύμανσης που εξηγούν στα δεδομένα.

Η PC1, ή πρώτη κύρια συνιστώσα, είναι ο γραμμικός συνδυασμός των μεταβλητών που αποτυπώνουν το μέγιστο ποσοστό διακύμανσης στα δεδομένα. Αντιπροσωπεύει την κατεύθυνση στον αρχικό χώρο χαρακτηριστικών (feature space) με τη μεγαλύτερη διακύμανση.

Η PC2, ή δεύτερη κύρια συνιστώσα, είναι ο γραμμικός συνδυασμός των μεταβλητών που αποτυπώνουν το δεύτερο υψηλότερο ποσοστό διακύμανσης στα δεδομένα. Είναι ορθογώνια (ασυσχέτιστη) προς την PC1 και αντιπροσωπεύει την κατεύθυνση που εξηγεί την υπόλοιπη διακύμανση των δεδομένων αφού έχουμε λάβει υπόψη την PC1.

Κάθε επόμενη κύρια συνιστώσα (PC3, PC4, κλπ.) εξηγεί μειούμενο ποσοστό διακύμανσης στα δεδομένα. Ο συνολικός αριθμός κύριων συνιστωσών είναι ίσος με τον αριθμό των αρχικών μεταβλητών στο σύνολο δεδομένων.

Οι PC1 και PC2, καθώς είναι οι πιο σημαντικές και ερμηνεύσιμες συνιστώσες, χρησιμοποιούνται συχνά για την οπτικοποίηση και την κατανόηση της υποκείμενης δομής ή μοτίβων στα δεδομένα.

2.6.3 Οπτικοποίηση χρονοσειρών με χρήση της τεχνικής PCA

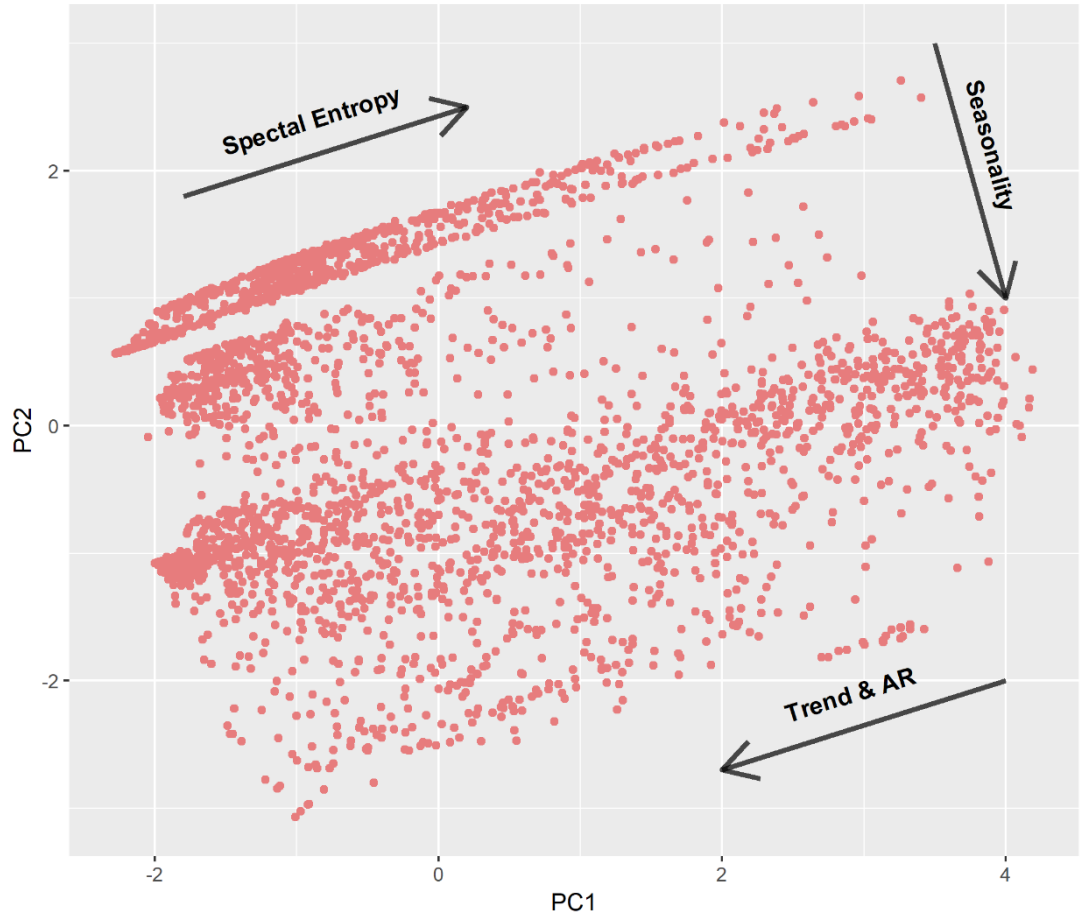
Για να κατανοήσουμε πως λειτουργεί αυτή η διαδικασία στην πράξη, μπορούμε να εξετάσουμε τα αποτελέσματα της τεχνικής Principal Component Analysis (PCA) στο πλαίσιο της μελέτης των Kang et al., 2017 [9], και να δούμε πως οι έξι μεταβλητές που περιγράψαμε προηγουμένως μπορούν να χρησιμοποιηθούν για την ανάλυση των χρονοσειρών. Σε αυτή την περίπτωση, θα εξετάσουμε μια πρακτική εφαρμογή της τεχνικής στις χρονοσειρές του διαγωνισμού προβλέψεων M3 ώστε να διαπιστώσουμε πως η PCA μπορεί να εφαρμοστεί σε πραγματικά δεδομένα και πως μπορεί να συμβάλει στην κατηγοριοποίηση και οπτικοποίηση των χρονοσειρών.

Για να μπορέσουμε να εφαρμόσουμε την ανάλυση κύριων συνιστωσών στα δεδομένα, πρέπει πρώτα να υπολογίσουμε τις τιμές όλων των χαρακτηριστικών για όλες τις χρονοσειρές του συνόλου δεδομένων. Αφού το κάνουμε αυτό, οι έξι μεταβλητές ομαδοποιούνται χρησιμοποιώντας μία γραμμική σχέση των επιμέρους μεταβλητών, η οποία βασίζεται στη συσχέτισή τους. Μέσω αυτής της διαδικασίας, εξάγονται ισάριθμες τεχνητές μεταβλητές που προκύπτουν ως γραμμικό άθροισμα των αρχικών και μπορούν να χρησιμοποιηθούν για να περιγράψουν τις επιμέρους χρονοσειρές. Για τις χρονοσειρές του M3, οι πρώτες δύο τεχνητές μεταβλητές επεξηγούν σχεδόν το 70% της συνολικής διακύμανσης των δεδομένων και

μπορούν να εκφραστούν αλγεβρικά ως:

$$\begin{bmatrix} PC1 \\ PC2 \end{bmatrix} = \begin{bmatrix} 0.530 & -0.537 & 0.228 & 0.234 & -0.555 & -0.122 \\ 0.252 & -0.160 & -0.656 & -0.655 & -0.180 & 0.136 \end{bmatrix} F \tag{2.11}$$

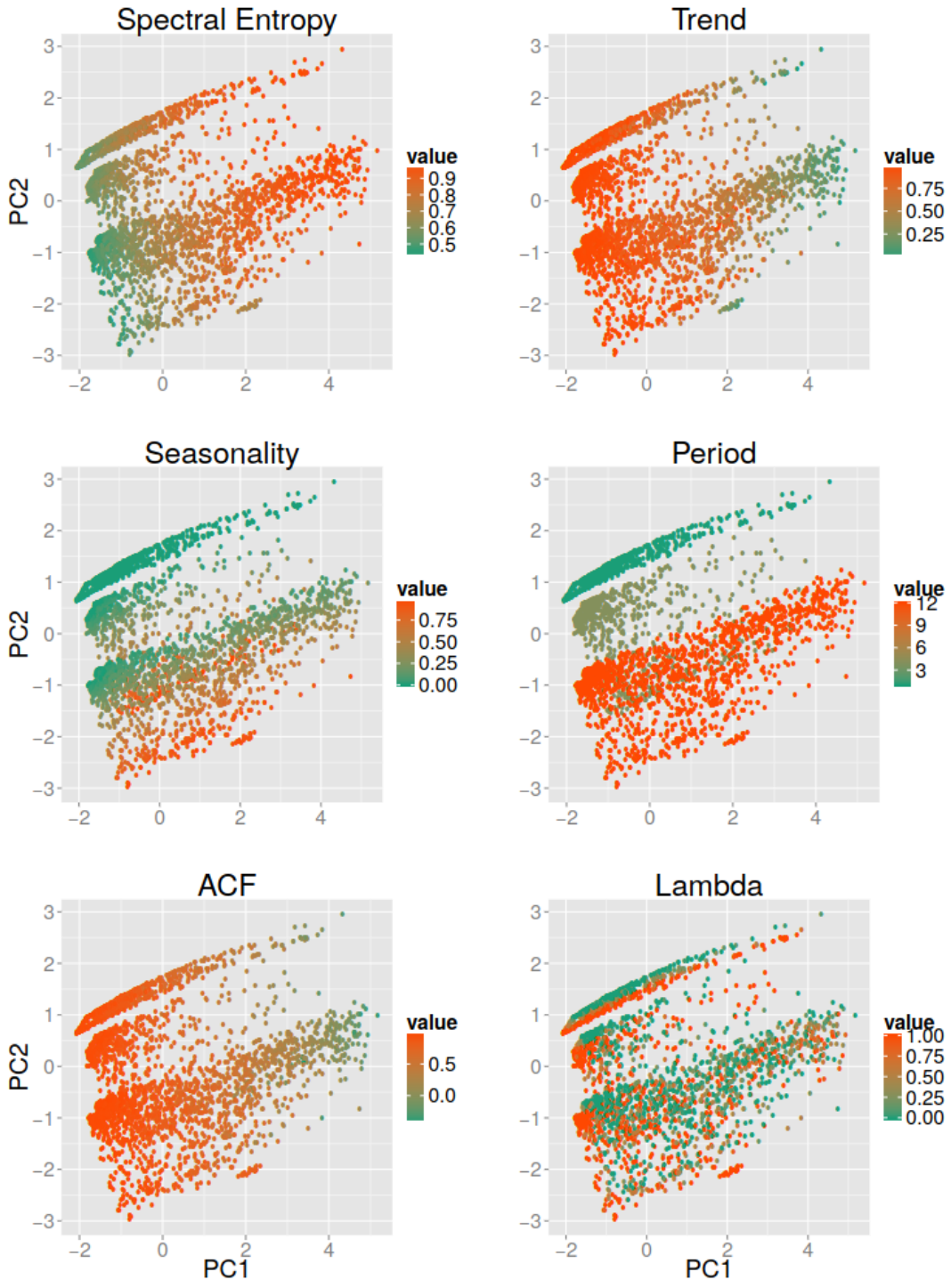
Δεδομένου ότι οι δύο τεχνητές μεταβλητές εξηγούν ένα σημαντικό ποσοστό της συνολικής διακύμανσης, μπορούν να χρησιμοποιηθούν για την οπτικοποίηση του συνόλου δεδομένων σε ένα δισδιάστατο χώρο, χωρίς σημαντική απώλεια πληροφορίας. Στο Σχήμα 2.5 φαίνεται η απεικόνιση των χρονοσειρών του M3 στον χώρο που ορίστηκε μέσω της PCA και των μεταβλητών των Kang et al., 2017.



Σχήμα 2.5: *Δυσδιάστατη παρουσίαση των χρονοσειρών του M3 στον χώρο που ορίστηκε μέσω της PCA και των μεταβλητών των Kang et al., 2017*

Στο Σχήμα 2.5, η ένταση εποχιακότητας (strength of seasonality) και περιοδικότητα των δεδομένων αυξάνεται κατά μήκος του y-άξονα καθώς κινούμαστε από πάνω προς τα κάτω. Επιλέον, κατά μήκος του x-άξονα, η ένταση τυχαιότητας (spectral entropy) αυξάνεται καθώς κινούμαστε από αριστερά στα δεξιά και η ένταση τάσης trend και ένταση αυτοσυσχέτισης (first order autocorrelation - AR) αυξάνονται από δεξιά στα αριστερά. Επομένως, χρονοσειρές που βρίσκονται κοντά μεταξύ τους σε αυτόν τον χώρο έχουν παρόμοιες τιμές στα έξι χαρακτηριστικά που εξετάστηκαν. Αυτό γίνεται ακόμα πιο εμφανές στο Σχήμα 2.6 όπου παρουσιάζονται οι κατανομές των χαρακτηριστικών για το σύνολο δεδομένων του M3 στον

χώρο που ορίστηκε.



Σχήμα 2.6: *Κατανομή χαρακτηριστικών των χρονοσειρών του M3 στον χώρο που ορίστηκε μέσω της PCA και των μεταβλητών των Kang et al., 2017*

Η αναπαράσταση των δεδομένων στον δισδιάστατο χώρο προσφέρει μια διαφορετική οπτική για την ανάλυση και την περιγραφή ενός συνόλου δεδομένων. Αυτό μπορεί να αποδειχθεί εξαιρετικά χρήσιμο, καθώς μπορεί να επιτρέψει την αναγνώριση διαφορετικών μοτίβων ή χα-

ρακτηριστικών που ίσως δεν ήταν εμφανή σε άλλες αναλύσεις.

2.7 Γεννήτριες Παραγωγής Χρονοσειρών

Για να αξιολογηθεί η αποτελεσματικότητα των διάφορων μεθόδων πρόβλεψης, απαραίτητη προϋπόθεση είναι η δοκιμή τους σε μεγάλο πλήθος δεδομένων με στόχο τη σύγκριση των αποτελεσμάτων που προκύπτουν από αυτές. Αυτή η διαδικασία αξιολόγησης είναι ευρέως διαδεδομένη στη βιβλιογραφία μέσω των διαγωνισμών πρόβλεψης [6], στους οποίους διάφορες μεθόδοι εφαρμόζονται σε συγκεκριμένα σύνολα δεδομένων προερχόμενα από ένα ευρή φάσμα εφαρμογών και τα οποία διαθέτουν διαφορετικά χαρακτηριστικά και ιδιαιτερότητες. Στη συνέχεια, τα αποτελέσματα αξιολογούνται συγκριτικά προκειμένου να διαπιστωθεί ποια μέθοδος παρουσιάζει καλύτερη επίδοση και υπό ποιες συνθήκες.

Ωστόσο, για την αντικειμενικότερη αξιολόγηση των μεθόδων πρόβλεψης, έχει παρατηρηθεί ότι απαιτείται η ύπαρξη χρονοσειρών ομοιόμορφα κατανεμημένων ως προς τα χαρακτηριστικά τους και τις εφαρμογές στις οποίες αναφέρονται [4]. Αυτό το πρόβλημα δεν είναι εύκολο να αντιμετωπιστεί, καθώς ακόμα και στην περίπτωση που διαθέτουμε έναν μεγάλο όγκο δεδομένων, αφενός δεν μπορούμε να είμαστε βέβαιοι ότι αντικατοπτρίζει πλήρως την πραγματικότητα και αφετέρου δεν γνωρίζουμε αν η πραγματικότητα θα έχει τελικά την ομοιόμορφη κατανομή που ψάχνουμε.

Αυτή η ανομοιομορφία στην κατανομή χρονοσειρών ως προς τα χαρακτηριστικά τους επισημάνθηκε από τους Hyndman et al., 2016 [18] για τον διαγωνισμό προβλέψεων M3 [16]. Επιπλέον, με χρήση της προαναφερθείσας τεχνικής *Principal Component Analysis (PCA)* έδειξαν (Σχήμα 2.5) πως στο δυσδιάστατο χώρο που δημιουργείται, οι χρονοσειρές του M3 συγκεντρώνονται σε συγκεκριμένες θέσεις, ενώ κάποιες άλλες παραμένουν εντελώς άδειες ή αραιοκατοικημένες, με αποτέλεσμα μεγάλο μέρους αυτού να παραμένει κενό [9].

Εδώ εκδηλώνεται η ανάγκη για επιπλέον χρονοσειρές με ποικίλα χαρακτηριστικά διαφορετικής έντασης ώστε να καλύψουν αυτό το κενό. Δεδομένης της έλλειψης πραγματικών χρονοσειρών, δημιουργείται η ανάγκη για χρήση γεννητριών που μπορούν να παράγουν μεγάλο πλήθος τεχνητών χρονοσειρών με ελεγχόμενα χαρακτηριστικά. Ακολούθως, θα εξετάσουμε δύο διαφορετικές προσεγγίσεις που μπορούν να χρησιμοποιηθούν για τη δημιουργία τεχνητών χρονοσειρών.

2.7.1 Γεννήτρια με βάση πραγματικά δεδομένα

Η γεννήτρια που αναπτύχθηκε από τον Σπηλιώτης, 2017 [4] στηρίζεται αποκλειστικά στη λογική της κλασικής πολλαπλασιαστικής μεθόδου αποσύνθεσης, με εξαίρεση όμως του χαρακτηριστικού της κυκλικότητας. Έτσι, για κάθε χρονοσειρά παράγονται δείκτες εποχιακότητας και ένα μοτίβο τάσης, και στο γινόμενο αυτών εφαρμόζεται λευκός θόρυβος:

$$Y_t = S_t * T_t * R_t \tag{2.12}$$

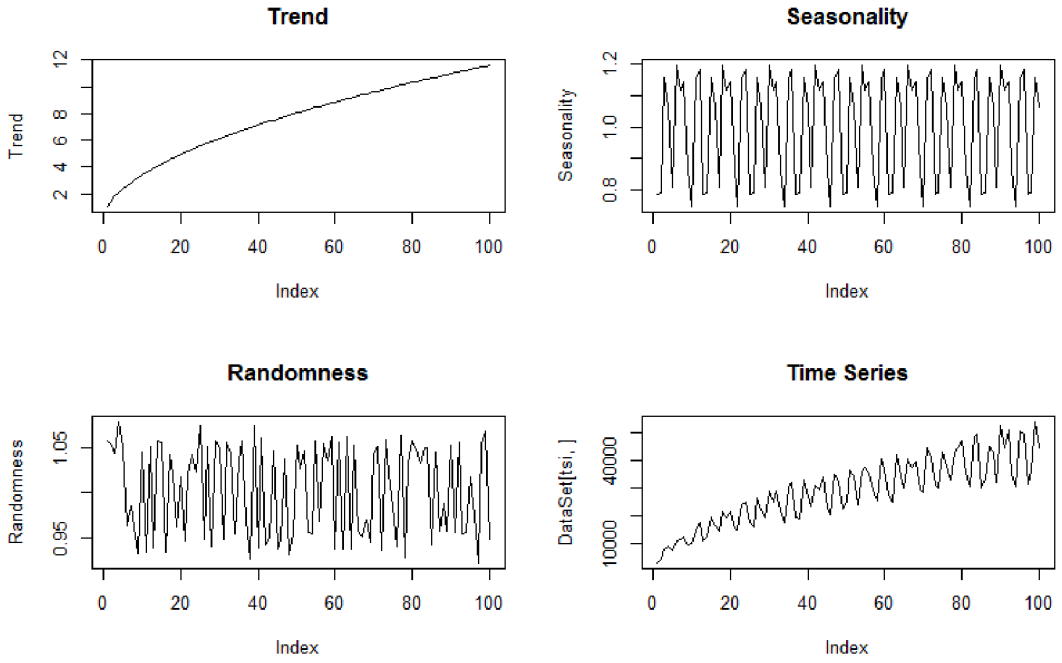
Το αποτέλεσμα είναι μία νέα τυχαία χρονοσειρά που μπορεί να είναι Χαμηλής (Low), Μεσαίας (Medium) ή Υψηλής (High) έντασης ως προς την εποχιακότητα, την τάση και την τυχαιότητα της.

Βασισμένοι σε αυτόν το διαχωρισμό, μπορούμε να κατηγοριοποιήσουμε οποιαδήποτε χρονοσειρά ανάλογα με το επίπεδο έντασης των συνιστωσών της χρησιμοποιώντας τον συμβολισμό Trend(Z,Z,Z), όπου ο όρος Trend αντιστοιχεί στον τύπο τάσης της χρονοσειράς, με τη γραμμική τάση που συναντάμε στην μεθοδολογία της κλασικής αποσύνθεσης να αντικαθίσταται με εναλλακτικές καμπύλες:

- Γραμμική τάση (**LIN**): $T_t = b + at$
- Εκθετική τάση (**EXP**): $T_t = be^{at}$
- Λογαριθμική τάση (**LOG**): $T_t = b + alog(t)$
- Αντίστροφη τάση (**INV**): $T_t = b + a/t$
- Τάση σε μορφή δύναμης (**POW**): $T_t = bt^a$

Τα ορίσματα Z περιγράφουν την ένταση των συνιστωσών Τάσης, Εποχιακότητας και Τυχαιότητας αντίστοιχα. Έτσι, ο συμβολισμός LOG(M,L,H) (Σχήμα 2.7) αναπαριστά μία χρονοσειρά με τάση σε μορφή δύναμης μεσαίας έντασης, χαμηλής έντασης εποχιακότητας και υψηλού επιπέδου τυχαιότητας.

Ο τρόπος με τον οποίο μετράμε το κάθε χαρακτηριστικό και τα όρια κάθε κατηγορίας έντασης χρησιμοποιούν ως βασική αναφορά πραγματικά δεδομένα από τον διαγωνισμό M3, με την ακριβή μεθοδολογία να περιγράφεται μέσα στην μελέτη.



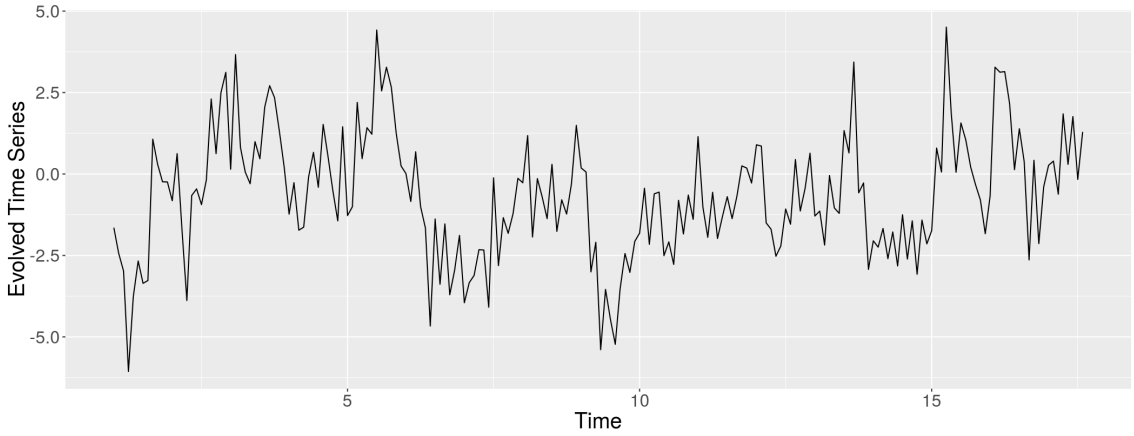
Σχήμα 2.7: *Παραγωγή μίας τυχαίας χρονοσειράς POW(M,L,H) μέσω της γεννήτριας του Σπηλιώτης, 2017 [4]*

2.7.2 Γεννήτρια με βάση mixture autoregressive (MAR) models

Οι Kang et al., 2020 [5] ανέπτυξαν το GRATIS (GeneRAting TIme Series with diverse and controllable characteristics, μια γεννήτρια για την παραγωγή τεχνητών χρονοσειρών που βασίζεται στην χρήση μεικτών αυτοπαλίνδρομων μοντέλων (mixture autoregressive (MAR) models). Τα μοντέλα MAR είναι αρκετά ευέλικτα ώστε να συλλαμβάνουν ένα ευρύ φάσμα εξαρτήσεων, συμπεριλαμβανομένων γραμμικών και μη γραμμικών σχέσεων μεταξύ παρελθοντικών και μελλοντικών τιμών μιας χρονοσειράς. Αυτή η ευελιξία επιτρέπει τη δημιουργία χρονοσειρών με ποικίλα μοτίβα τάσης, εποχικότητας και αυτοσυσχέτισης.

Ένα από τα κυριότερα χαρακτηριστικά του GRATIS είναι η ικανότητά του, ρυθμίζοντας τις παραμέτρους των μοντέλων MAR, να δημιουργεί αποτελεσματικά νέες χρονοσειρές με ελεγχόμενα χαρακτηριστικά. Αυτό είναι ιδιαίτερα χρήσιμο σε περιπτώσεις όπου μας ενδιαφέρουν χρονοσειρές με συγκεκριμένα χαρακτηριστικά, όπως για παράδειγμα το πρόβλημα της ανομοιομορφίας στην κατανομή χρονοσειρών που επισημάνθηκε στην εισαγωγή αυτού το κεφαλαίου.

Επιπλεόν, οι συγγραφείς δείχνουν ότι η μέθοδός τους για τη δημιουργία νέων χρονοσειρών από συγκεκριμένα χαρακτηριστικά μπορεί να επιταχύνει τον υπολογιστικό χρόνο, σε σχέση με προηγούμενες μεθόδους, έως και 40 φορές, καθιστώντας εφικτές αναλύσεις χρονοσειρών που επικεντρώνονται στα χαρακτηριστικά.



Σχήμα 2.8: *Παραγωγή μίας τυχαίας χρονοσειράς με (trend, entropy, seasonal_strength) = (0.59, 0.85, 0.29) μέσω της γεννήτριας των Kang et al., 2020 [5]*

Κεφάλαιο 3

Συλλογή Δεδομένων

Η συλλογή και ο επαρκής χειρισμός των δεδομένων αποτελεί τη βάση ολόκληρου του προς ανάπτυξη συστήματος. Στο πλαίσιο αυτό, παρουσιάζουμε τη διαδικασία συλλογής δεδομένων που ακολουθήθηκε κατά τη δημιουργία του συστήματος FOREDeCk.

3.1 Αναγκαιότητα

Στο πλαίσιο της ανάλυσης ή της πρόβλεψης χρονοσειρών, είναι σύνηθες να αξιολογείται η αποτελεσματικότητα των εξεταζόμενων μεθόδων δοκιμάζοντας την απόδοση τους σε γνωστά σύνολα δεδομένων. Ωστόσο, αυτή η προσέγγιση αρχίζει να εμφανίζει δυσκολίες όταν ο ερευνητής επιθυμεί να αξιολογήσει μια μέθοδο που αφορά χρονοσειρές με πολύ συγκεκριμένα χαρακτηριστικά ή όταν χρειάζεται μεγάλο πλήθος χρονοσειρών. Η εύρεση ικανοποιητικής, για της ανάγκες επικύρωσης ενός μοντέλου, ποσότητας χρονοσειρών με τα συγκεκριμένα χαρακτηριστικά που εμείς θέλουμε, μπορεί να είναι δύσκολη έως και μη εφικτή.

Η ανεπάρκεια ποιοτικών δεδομένων με συγκεκριμένα χαρακτηριστικά μπορεί να οφείλεται τόσο στην έλλειψη τέτοιων δεδομένων, όσο και στην δυσκολία αναζήτησης τους, καθώς τα χαρακτηριστικά που ενδεχομένως ενδιαφέρουν ένα ερευνητή σπανίως προσφέρονται από την εκάστοτε πηγή δεδομένων. Αν και υπάρχει πληθώρα πηγών δεδομένων διαθέσιμη στο διαδίκτυο, αυτές οι πηγές συχνά δεν παρέχουν αρκετές πληροφορίες για τις συγκεκριμένες ιδιότητες των χρονοσειρών που θέλουμε να αξιολογήσουμε. Αυτό υποχρεώνει τους ερευνητές να εξετάζουν κάθε χρονοσειρά ξεχωριστά, προκειμένου να διαπιστώσουν αν ανήκει στο επιθυμητό σύνολο δεδομένων. Αυτή η διαδικασία είναι χρονοβόρα και μπορεί να εμποδίζει την αποτελεσματική αξιολόγηση μιας μεθόδου.

Αυτή την ανάγκη προσπαθούμε να καλύψουμε με το FOREDeCk. Δημιουργώντας μια δεξαμενή δεδομένων με πραγματικές χρονοσειρές και υποστηρίζοντας πολλαπλό φιλτράρισμα με βάση τον τύπο και τα χαρακτηριστικά των δεδομένων, το σύστημα παρέχει ευέλικτες δυνατότητες πρόσβασης σε επιθυμητά δεδομένα, εξοικονομώντας χρόνο και απλοποιώντας την διαδικασία αξιολόγησης και ανάπτυξης μοντέλων.

3.2 Προδιαγραφές

Οι προδιαγραφές που καθορίζονται στην ενότητα αυτή αποτελούν τον οδηγό για την ανάπτυξη μιας αποτελεσματικής διαδικασίας συλλογής δεδομένων. Μέσα από τον καθορισμό

βασικών αρχών, ορίζεται η διασφάλιση της ποιότητας και της αξιοπιστίας του τελικού συνόλου δεδομένων.

Παρακάτω αναλύονται οι βασικές προδιαγραφές που καθορίστηκαν για τη διαδικασία συλλογής δεδομένων:

- **Διαφορετικά πεδία πρόβλεψης**: Η διαδικασία συλλογής πρέπει να μπορεί να συλλέγει χρονοσειρές από πολλά διαφορετικά πεδία πρόβλεψης (forecasting fields). Αυτό επιτρέπει στους χρήστες να αξιοποιούν τα δεδομένα για διαφορετικές εφαρμογές και ανάγκες πρόβλεψης σε διάφορους τομείς.

- **Ποικιλία ποιοτικών χαρακτηριστικών**: Η διαδικασία συλλογής πρέπει να εξασφαλίζει την συγκέντρωση χρονοσειρών με ποικιλία χαρακτηριστικών, όπως συχνότητα δειγματοληψίας, τάση, εποχικότητα, κτλ. Αυτό επιτρέπει στους χρήστες να επιλέγουν τα χαρακτηριστικά που ταιριάζουν στις ανάγκες τους.

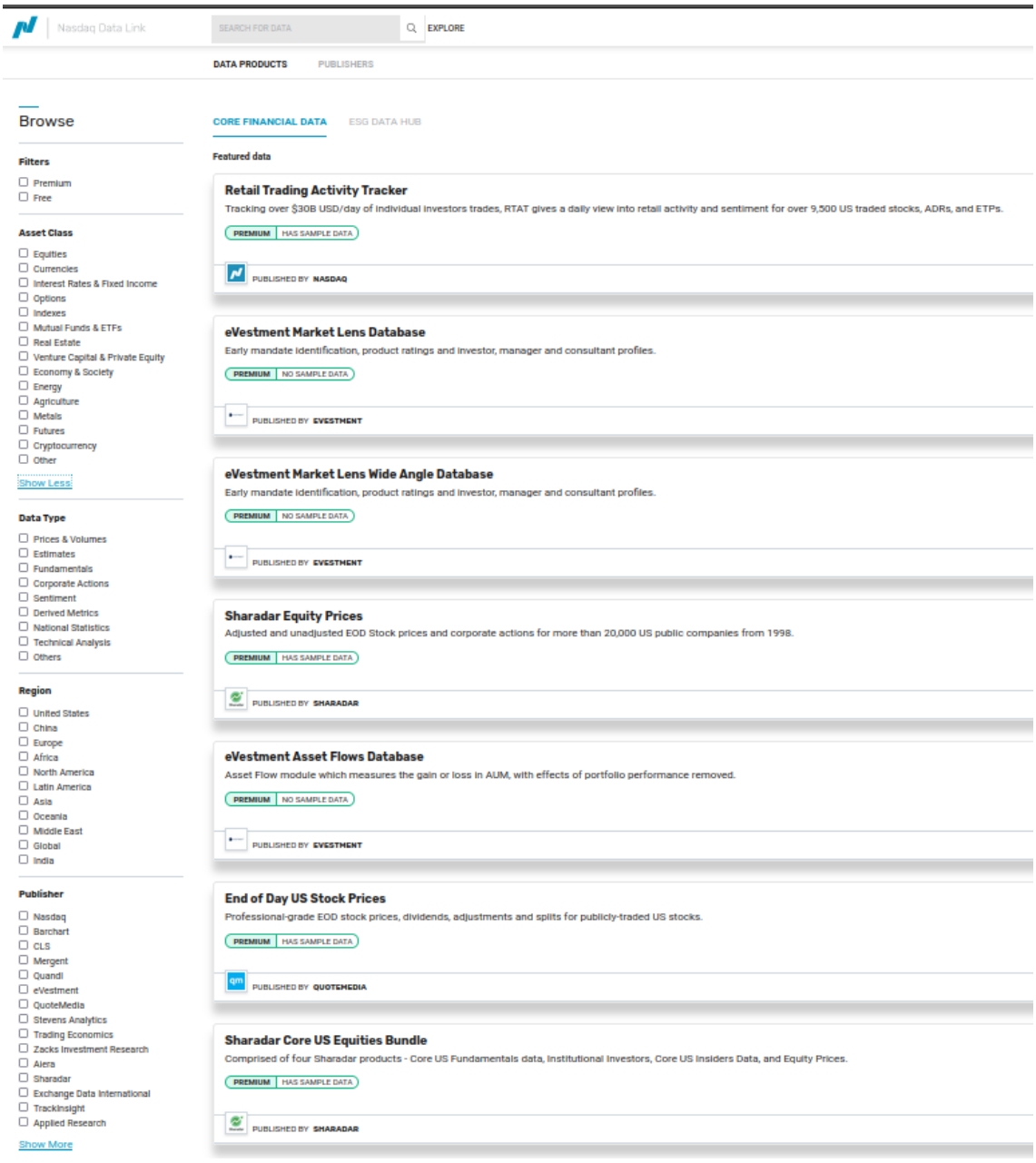
- **Αυτοματοποίηση**: Η διαδικασία συλλογής πρέπει να είναι αυτοματοποιημένη και να μειώνει την ανθρώπινη παρέμβαση στο μέτρο του δυνατού. Αυτό εξασφαλίζει την εξοικονόμηση χρόνου και τη μείωση του σφάλματος. Για να επιτευχθεί αυτό, οι πηγές δεδομένων πρέπει να παρέχουν μια μορφή API που θα επιτρέπει την αυτόματη ανάκτηση των χρονοσειρών.

- **Διαφοροποίηση Πηγών**: Η συλλογή δεδομένων πρέπει να επιτρέπεται από διάφορες πηγές, συμπεριλαμβανομένων αρχείων τύπου CSV, στην περίπτωση που δεν υπάρχει δυνατότητα πρόσβασης μέσω API. Αυτή η επιλογή επιτρέπει τη μεγαλύτερη ευελιξία στη συλλογή δεδομένων.

- **Επεκτασιμότητα**: Πρέπει να υπάρχει η δυνατότητα εύκολης προσθήκης νέων πηγών δεδομένων και χαρακτηριστικών στο μέλλον, καθώς το σύστημα εξελίσσεται.

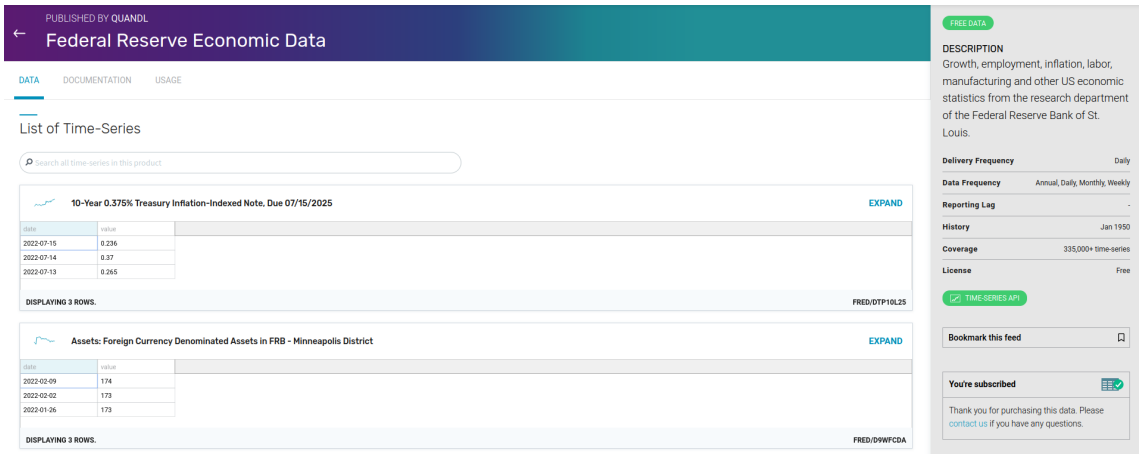
3.3 Πηγές δεδομένων

Η συλλογή δεδομένων πραγματοποιήθηκε κυρίως μέσω της πλατφόρμας Quandl (πλέον Nasdaq Data Link) [19], η οποία αποτελεί μία από τις κορυφαίες και αξιόπιστες πηγές δεδομένων για χρηματοοικονομικά, οικονομικά και εναλλακτικά σύνολα δεδομένων. Η επιλογή της πλατφόρμας Quandl βασίστηκε στην εξαιρετική ποιότητα και ποικιλία χρονοσειρών που προσφέρει, επιτρέποντας μας να έχουμε ευρεία κάλυψη δεδομένων και να απευθυνόμαστε σε διάφορα πεδία ενδιαφέροντος.



Σχήμα 3.1: *Η πλατφόρμα Quandl (Nasdaq Data Link)*

Στην πλατφόρμα Quandl, τα δεδομένα οργανώνονται σε βάσεις δεδομένων (databases) με βάση το αντικείμενο τους (Σχήμα 3.1). Αυτά τα databases χωρίζονται σε δωρέαν (free) και επί πληρωμή (premium) και κάθε database συνήθως περιέχει περισσότερα από ένα σύνολα δεδομένων (datasets), τα οποία, με τη σειρά τους, περιλαμβάνουν περισσότερες από μία χρονοσειρές. Για παράδειγμα, η βάση δεδομένων "Federal Reserve Economic Data (FRED)" περιέχει περισσότερα από 335χιλιάδες datasets τα οποία περιέχουν χρονοσειρές διαφόρων μεγεθών και συχνοτήτων (Σχήμα 3.2). Ενδεικτικά, το Quandl παρέχει εκατοντάδες βάσεις δεδομένων, με κάθε μία να περιλαμβάνει εκατοντάδες εώς και χιλιάδες σύνολα δεδομένων, προσφέροντας συνολικά εκατομμύρια χρονοσειρές.



Σχήμα 3.2: *Ομαδοποίηση δεδομένων στην πλατφόρμα Quandl*

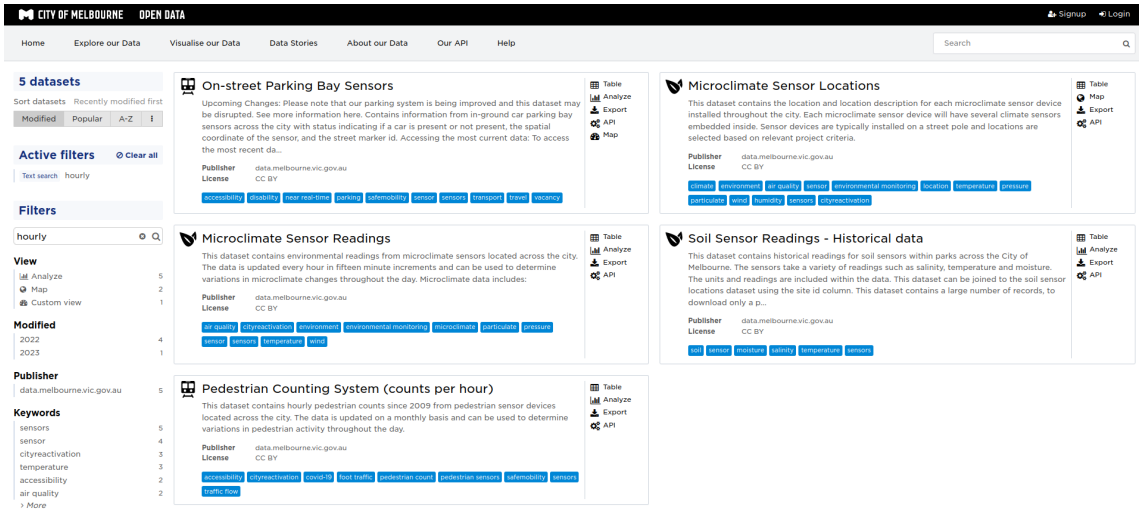
Για την αποτελεσματική αναζήτηση και συλλογή των δεδομένων χρησιμοποιήθηκε το API του Quandl. Αρχικά, πραγματοποιήθηκε αναζήτηση και καταγραφή όλων των διαθέσιμων δωρεάν βάσεων δεδομένων (databases) ώστε να έχουμε μια πρώτη εικόνα της ποσότητας και των βασικών χαρακτηριστικών των δεδομένων που έχουμε στην διάθεση μας. Για την κάθε βάση καταγράψαμε το πλήθος των συνόλων δεδομένων που περιέχει και το top category, δηλαδή την κατηγορία στην οποία ανήκουν τα περισσότερα από τα datasets της κάθε βάσης. Οι κατηγορίες αυτές περιλαμβάνουν δεδομένα Stock, Interest Rate, Economic, Index, Industry, Currency, Futures, Commodity και Asset Management.

Με βάση τις προδιαγραφές που ορίσαμε για τη συλλογή δεδομένων, ο στόχος μας ήταν να δημιουργήσουμε ένα εκτενές τελικό σύνολο δεδομένων που θα κάλυπτε μια ευρεία γκάμα πεδίων πρόβλεψης. Για να το επιτύχουμε αυτό, επιλέξαμε βάσεις από διαφορετικές κατηγορίες, οι οποίες περιλάμβαναν αρκετά datasets. Στον Πίνακα 3.1 παρατίθενται οι τελικές βάσεις δεδομένων που επιλέχθηκαν για τη συλλογή δεδομένων, ανάμεσα σε περισσότερες από 200 διαθέσιμες στο Quandl.

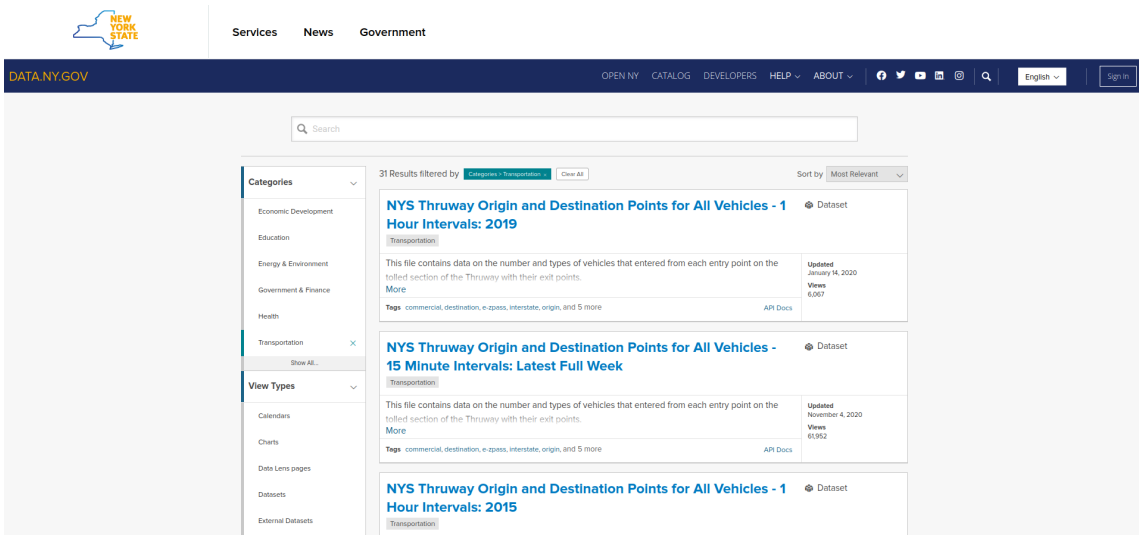
Επιπλέον, για τη συμπλήρωση του συνόλου δεδομένων μας, αξιοποιήθηκαν και άλλες πηγές δεδομένων που παρείχαν χρονοσειρές με συγκεκριμένα χαρακτηριστικά, τα οποία δεν ήταν διαθέσιμα μέσω του Quandl. Πιο συγκεκριμένα, το Quandl δεν περιλαμβάνει ωριαίες χρονοσειρές (high-frequency datasets), τις οποίες χρειαζόμασταν για την πληρότητα του συνόλου δεδομένων μας. Για να καλύψουμε αυτό το κενό, ανατρέξαμε σε ανοιχτές πλατφόρμες δεδομένων (open data platforms) από διάφορες πόλεις παγκοσμίως. Ειδικότερα, αξιοποιήσαμε δεδομένα αισθητήρων από την πόλη της Μελβούρνης [20] και το σύστημα συγκοινωνιών της Νέας Υόρκης [21].

Κωδικός	Όνομα βάσης δεδομένων
ADB	Asian Development Bank
ADP	ADP Research Institute
AMECO	European Commission Annual Macro-Economic Database
AUSBS	Australian Bureau of Statistics
BANKISRAEL	Bank of Israel
BCB	Central Bank of Brazil Statistical Database
BDM	Bank of Mexico
BEA	Bureau of Economic Analysis
BOE	Bank of England Official Statistics
BOJ	Bank of Japan
BUNDESBANK	Deutsche Bundesbank Data Repository
CEPEA	Center for Applied Studies on Applied Economics (Brazil)
CFTC	Commodity Futures Trading Commission Reports
CME	Chicago Mercantile Exchange
EIA	U.S. Energy Information Administration Data
FED	US Federal Reserve Data Releases
FRED	Federal Reserve Economic Data
HKEX	Hong Kong Exchange
JODI	JODI Oil World Database
LME	London Metal Exchange
NASDAQOMX	NASDAQ OMX Global Index Data
NBSC	National Bureau of Statistics of China
OECD	Organisation for Economic Co-operation and Development
RBA	Reserve Bank of Australia
SNB	Swiss National Bank
UFAO	United Nations Food and Agriculture
UGID	United Nations Global Indicators
UKONS	United Kingdom Office of National Statistics
ZILLOW	Zillow Real Estate Research

Πίνακας 3.1: *Οι βάσεις δεδομένων που επιλέχθηκαν από το Quandl για την συλλογή δεδομένων*



Σχήμα 3.3: *Melbourne's Open Data Platform*



Σχήμα 3.4: *New York's Open Data Portal*

Συνολικά, η ποικιλία των πηγών δεδομένων που αξιοποιήσαμε μας επιτρέπει να καλύψουμε μια ευρεία γκάμα αναγκών και εφαρμογών, εξασφαλίζοντας την ποικιλία και αξιοπιστία των δεδομένων του συστήματος.

3.4 Φιλτράρισμα και Επεξεργασία Δεδομένων

Η πληθώρα των διαθέσιμων δεδομένων στο διαδίκτυο παρέχει μια ποικιλία πληροφοριών, αλλά όχι πάντα και την απαραίτητη ποιότητα. Η ποιότητα των δεδομένων είναι κρίσιμης σημασίας για την αξιοποίηση τους σε αναλύσεις και προβλέψεις. Για τον λόγο αυτό, η διαδικασία του φιλτραρίσματος αποτελεί σημαντικό βήμα για τη δημιουργία ενός αξιόπιστου συνόλου δεδομένων που θα διευκολύνει την περαιτέρω ανάλυση και επεξεργασία.

Τα βασικά κριτήρια που καθορίστηκαν για την διαδικασία φιλταρίσματος των αρχικών δεδομένων είναι τα ακόλουθα:

1. **Μελλοντικές παρατηρήσεις**: Ελέγχουμε εάν τα δεδομένα περιέχουν μελλοντικά γεγονότα, αφαιρώντας χρονοσειρές που αναφέρονται σε προβλέψεις για το μέλλον.

2. **Μηδενικές παρατηρήσεις**: Εξασφαλίζουμε ότι δεν υπάρχουν μηδενικές παρατηρήσεις ή απουσιάζουσες τιμές (missing values) στις χρονοσειρές.

3. **Ελάχιστος αριθμός παρατηρήσεων**: Βεβαιωνόμαστε ότι η χρονοσειρά έχει επαρκή αριθμό παρατηρήσεων σύμφωνα με την συχνότητα της ώστε να μπορεί να γίνει αξιόπιστη ανάλυση και πρόβλεψη. Γενικότερα, απαιτούμε τουλάχιστον 5 φορές τον αριθμό παρατηρήσεων σε σχέση με τη συχνότητα.

4. **Ελάχιστη τιμή παρατήρησης**: Αποκλείουμε τις χρονοσειρές με παρατηρήσεις με τιμή μικρότερη του 10.

5. **Θόρυβος και ακραίες τιμές**: Ελέγχουμε για πιθανές λανθασμένες ή ακραίες τιμές (outliers) στο σύνολο δεδομένων που θα μπορούσαν να επηρεάσουν την αξιοπιστία των αναλύσεών μας.

Από το σύνολο των χρονοσειρών που συλλέχθηκαν από κάθε βάση, περίπου ποσοστό 20%-30% πληρούσε τα κριτήρια που τέθηκαν. Οπότε, για τις περίπου 1 εκατομμύριο χρονοσειρές που καταλήξαμε, χρειάστηκε να συλλέξουμε και να περάσουμε από διαδικασία φιλτραρίσματος περίπου 5 εκατομμύρια χρονοσειρές. Στον Πίνακα 3.2 παρουσιάζονται ενδεικτικά τα ακριβή αριθμητικά αποτελέσματα της διαδικασίας φιλταρίσματος για μερικές από τις βάσεις δεδομένων.

Βάσεις δεδομένων	BCB	FED	CFTC	UFAO	UGID
Σύνολο datasets σε βάση	13,700	48,078	31,953	33,460	64,920
Σύνολο χρονοσειρών σε datasets	13,700	48,300	352,820	114,213	96,374
Αφαιρεθηκαν με Φίλτρο 1	93	584	539	3,957	4,202
Αφαιρεθηκαν με Φίλτρο 2	5,429	29,290	70,844	32,591	24,795
Αφαιρεθηκαν με Φίλτρο 3	2,872	3,977	228,473	21,196	15,246
Αφαιρεθηκαν με Φίλτρο 4	0	0	10,517	2,586	10,705
Αφαιρεθηκαν με Φίλτρο 5	1,270	5,793	6,380	6,529	2,603
Χρονοσειρές που τελικά προστέθηκαν στην βάση	4,036	8,656	36,067	47,354	38,823
Ποσοστό που πληρούσε τα κριτήρια	29%	18%	10%	41%	40%

Πίνακας 3.2: *Ενδεικτικά αριθμητικά αποτελέσματα διαδικασίας φιλταρίσματος για το Quandl*

Ολοκληρώνοντας το φιλτράρισμα των δεδομένων, ακολουθεί η επεξεργασία τους για την απόκτηση επιπρόσθετων πληροφοριών. Σε αυτό το στάδιο, στοχεύουμε στην εξαγωγή βασικών μεταδεδομένων (metadata) από τις χρονοσειρές μας, καθώς και στον υπολογισμό των χαρακτηριστικών τους.

Τα μεταδεδομένα περιλαμβάνουν πληροφορίες που ανακτώνται άμεσα από την πηγή δεδομένων, όπως η πηγή των δεδομένων, το όνομα της χρονοσειράς, μια σύντομη περιγραφή, κλπ. Επιπλέον, δημιουργούμε και προσθέτουμε μεταδεδομένα που μπορεί να αποδειχθούν χρήσιμα στο μέλλον. Για παράδειγμα, προσθέτουμε την ημερομηνία προσθήκης της χρονοσειράς στη βάση δεδομένων, καθώς και τις ημερομηνίες της πρώτης και τελευταίας παρατήρησης της χρονοσειράς. Αυτές οι πληροφορίες μπορούν να αποδειχθούν πολύτιμες για την περαιτέρω ανάλυση, διαχείριση και επεκτασιμότητα των δεδομένων μας στο μέλλον.

Πέρα από τα μεταδεδομένα, σε αυτό το σημείο, κάνουμε μια βασική στατιστική ανάλυση υπολογίζοντας τα μεγέθοι που παρουσιάζονται στο Κεφάλαιο 2.5 και προχωρούμε στον υπολογισμό των χαρακτηριστικών που περιγράψαμε στο Κεφάλαιο 2.6.1.

Το επόμενο βήμα στην επεξεργασία δεδομένων είναι η χρήση των ήδη υπολογισμένων χαρακτηριστικών για την πραγματοποίηση της Principal Component Analysis (PCA) όπως αυτή έχει οριστεί στο Κεφάλαιο 2.6.2. Η τεχνική PCA εξαρτάται από το σύνολο των χρονοσειρών, γιαυτό και υπολογίστηκε στο τέλος της διαδικασίας συλλογής δεδομένων και αφού είχαμε στα χέρια μας όλες τις χρονοσειρές που αποτελούν το σύνολο δεδομένων του συστήματος μας. Αυτό σημαίνει πως σε περίπτωση προσθήκης επιπλέον χρονοσειρών, η PCA πρέπει να υπολογιστεί εκ νέου.

3.5 Ομαδοποίηση Δεδομένων

Η ομαδοποίηση των δεδομένων καθορίζει σε μεγάλο βαθμό την αποδοτικότητα του συστήματος, καθώς διευκολύνει την αποτελεσματική οργάνωση, επεξεργασία και ανάκτηση των δεδομένων. Αυτή η διαδικασία βασίστηκε σε δύο βασικούς πυλώνες: τη συχνότητα των χρονοσειρών και τις σχετιζόμενες με αυτές κατηγορίες.

Η συχνότητα (ετήσια, τριμηνιαία, μηνιαία, εβδομαδιαία, ημερήσια, ανά ώρα) ανακτήθηκε κατευθείαν από τις πηγές δεδομένων και η κατηγοριοποίηση έγινε με τεχνικές Επεξεργασίας Φυσικής Γλώσσας (αγγλ. Natural Language Processing - NLP) χρησιμοποιώντας ως βάση τις κατηγορίες που χρησιμοποιήθηκαν στον διαγωνισμό πρόβλεψης χρονοσειρών M3 (Micro, Industry, Macro, Finance, Demographics, Others) [16]. Αυτές οι κατηγορίες παρέχουν ένα κοινό πλαίσιο αναφοράς για τις χρονοσειρές μας και διευκολύνουν τη σύγκριση με άλλες μελέτες και διαγωνισμούς.

Αρχικά, ακολουθήσαμε μια προσέγγιση που επέτρεπε την κατηγοριοποίηση σε ένα ευρύτερο φάσμα κατηγοριών με βάση το όνομα της κάθε χρονοσειράς το οποίο και δίνει την περισσότερη πληροφορία για το περιεχόμενο της. Για να το επιτύχουμε αυτό, καθαρίσαμε τα ονόματα από σημεία στίξης, εφαρμόσαμε μετατροπές όπως η μετατροπή σε πεζά γράμματα και αφαιρέσαμε ειδικούς χαρακτήρες. Μετά διαμερίσαμε τα ονόματα σε λέξεις (tokenization) και στη συνέχεια αφαιρέσαμε τις λεγόμενες "stop words", που είναι λέξεις χωρίς σημασία όπως ¨και¨, ¨από¨, κλπ. Έπειτα, εφαρμόσαμε την τεχνική του stemming [22] προκειμένου να ενοποιήσουμε λέξεις με κοινή ρίζα ώστε να καταλήξουμε σε μια λίστα με όλες τις μοναδικές λέξεις που προέκυψαν από τα ονόματα των χρονοσειρών.

Στη συνέχεια, για να βελτιστοποιήσουμε την ομαδοποίηση, χρησιμοποιήσαμε το WordNet [23], μια λεξική βάση δεδομένων που περιλαμβάνει σημασιολογικές σχέσεις μεταξύ λέξεων, για να εξετάσουμε ποιες από τις λέξεις που είχαμε ήδη στη λίστα μας ήταν συνώνυμες μεταξύ τους. Αυτή η διαδικασία μας επέτρεψε να καταλήξουμε σε μια λίστα λέξεων-κλειδιών για την κάθε κατηγορία.

Government	323844	gdp	prices	price	gnp									
Energy	202597	current	energy	electricity	cloud	resonance	conduction	pressure	pressures	flotations				
Health	161537	care	work	market	recreation	assistance	support	enterprises	service	search	procedure	performing	enterprise	logging
Labor and Wage	137572	services	jobs	occupation	occupation	catering	craft	bookings	coaching					
Natural Resources	135526	fuel	coal	food	petroleum	fuels	gasoline	propane	biomass	coke	kerosene	lubricants	foods	butane
Industry	128826	industry	industries	business	collective	cooperative								
Parity and currency	121892	gross	tens	million	billions	hundreds	grand	digit	thousand	billion	quadrillion	hundred	twelve	twenty
Imports and Exports	111540	product	products	exports	commodity	merchandis	clothing	freight	durables	import	shipments	commoditi	cargo	foodst
Imports and Exports	111151	trade	financing	retail	investment	finance	agriculture	publishing	sale	investment	operations	manufactur	manufactur	renting
Households	108023	consumption	sex	eating	tanning	feeding	intake	dining	uptake					
Labor and Wage	78217	expenditure	expenditures	payments	payment	benefits	premium	benefit	portage	premiums	loss	expenses	overhead	remitta
Industry	73043	manufacturing	construction	creation	metalworki	baking	constructio	forgings	moldings	forging	shipbuilding	molding		
Government	72155	government	organizations	authorities	organizatio	porte	governmen	governmen	regime					
Transportation	62565	transportation	base	facilities	station	museums	depository	facility	terminals	plumbing	airport	substation	terminal	museu
Transportation	59547	equipment	vehicle	transport	vehicles	telecommu	key	medium	instrument	appliance	media	machine	machines	appara
Inventory	54349	goods	asset	good	plus	profitability	lead							
Labor and Wage	53683	labour	labor	people										
Services	53299	building	hospitals	housing	dwelling	buildings	plant	structures	dwellings	houses	hotels	house	restaurants	hospita
Households	51898	domestic	maids											
Energy	49098	gas	exchange	gases	liquidity	exchanges	gas							
Labor and Wage	47611	employees	labourers	salespersons	earner	hire	hires	servers	gypsy	miner	sawyer	collier	laborers	sorters
Insurances	45016	insurance	assets	reinsurance										
Labor and Wage	43674	compensation	wage	costs	salary	pension	salaries	pay	pensions	remunerati	fine	annuities	fines	annuit
Labor and Wage	42218	hours	periods	workweek	go									

Σχήμα 3.5: *Μερικές από τις αρχικές κατηγορίες και τις αντίστοιχες λέξεις-κλειδιά που προέκυψαν κατά την ομαδοποίηση δεδομένων*

Έχοντας διαθέσιμη τη λίστα με τις κατηγορίες και τις αντίστοιχες λέξεις, προχωρήσαμε στο βήμα της αρχικής κατηγοριοποίησης των χρονοσειρών. Σε αυτό το σημείο, εφαρμόσαμε

την τεχνική TF-IDF (Term Frequency - Inverse Document Frequency). Αυτή η τεχνική μας επέτρεψε να αξιολογήσουμε κάθε χρονοσειρά ξεχωριστά, λαμβάνοντας υπόψη τη συχνότητα εμφάνισης κάθε λέξης-κλειδιού στο όνομά της σε σχέση με το σύνολο των δεδομένων μας. Αυτό μας επέτρεψε να αναθέσουμε κατηγορίες με βάση τις σημαντικές λέξεις που εμφανίστηκαν στα ονόματα των χρονοσειρών, δίνοντας περισσότερη βάση στις λέξεις που είχαν μεγαλύτερη σημασία για το περιεχόμενό τους. Αυτό το σύστημα αξιολόγησης της κάθε χρονοσειράς μας βοήθησε να την κατηγοριοποιήσουμε ακόμη πιο ακριβώς και αποτελεσματικά, μειώνοντας την πιθανότητα ανάμειξης σε πολλές κατηγορίες και αυξάνοντας την ακρίβεια της ομαδοποίησης.

Στη συνέχεια, επιλέξαμε να συμπυκνώσουμε τη λίστα αυτή σε ένα πιο περιορισμένο σύνολο κατηγοριών, συνοψίζοντας και συνδυάζοντας τις ευρύτερες κατηγορίες σε μια μικρότερη, αλλά πιο πυκνή λίστα με τις τελικές 6 κατηγορίες: Micro, Industry, Macro, Finance, Demographics, και Other.

Αυτή η διαδικασία ομαδοποίησης διευκόλυνε τη διαχείριση των χρονοσειρών, επέτρεψε την αποδοτική οργάνωση, ανάκτηση και επεξεργασία των δεδομένων μας, και συνέβαλε στην επιτυχημένη ανάπτυξη του συστήματος μας.

Συχνότητα	Micro	Industry	Macro	Finance	Demographics	Other	Σύνολο
Ετήσιες	9,681	482,034	126,388	33,461	72,383	2,647	813,734
Τριμηνιαίες	10,746	14,401	24,704	37,083	10,989	911	98,834
Μηνιαίες	16,553	43,037	20,310	33,497	21,819	648	135,864
Εβδομαδιαίες	1,547	3,640	449	9,854	247	12	15,749
Ημερήσιες	2,201	21,164	302	18,758	21	3,091	45,537
Ωριαίες	0	0	0	0	0	414	414
Σύνολο	127,868	564,276	172,153	132,653	105,459	7,723	1,110,132

Πίνακας 3.3: *Πλήθος χρονοσειρών ανά συχνότητα και κατηγορία μετά την συλλογή δεδομένων*

Κεφάλαιο 4

Παρουσίαση Συστήματος

Στο κεφάλαιο αυτό παρουσιάζεται η μελέτη που έγινε για την υλοποίηση του συστήματος. Για τις απαιτήσεις της λύσης και σύμφωνα με την μεθοδολογία και τις ανάγκες που περιγράφονται στα προηγούμενα κεφάλαια, υλοποιήθηκε ένα σύστημα που επιτρέπει στον χρήστη να κατεβάσει αλλά και να δημιουργήσει χρονοσειρές σύμφωνα με τις ανάγκες που μπορεί να έχει. Η μορφή και λειτουργία του συστήματος παρουσιάζονται σε αυτή την ενότητα.

4.1 Απαιτήσεις Συστήματος

Η διαδικασία ανάπτυξης ενός συστήματος απαιτεί την τήρηση συγκεκριμένων κανόνων σχεδίασης και ανάπτυξης προκειμένου να επιτευχθεί η απρόσκοπτη λειτουργία του και η ικανοποίηση των χρηστών του. Οι απαιτήσεις του συστήματος περιλαμβάνουν τόσο τεχνικές απαιτήσεις, που σχετίζονται με την εμφάνιση και τη διεπαφή χρήστη (user interface), όσο και επιχειρηματικές απαιτήσεις, που καθορίζονται από τους τελικούς χρήστες και την εργασία που θα πραγματοποιήσουν μέσω του συστήματος.

4.1.1 Τεχνικές απαιτήσεις συστήματος

Οι τεχνικές απαιτήσεις συστήματος αφορούν θέματα που σχετίζονται με την εμφάνιση και τη διεπαφή του συστήματος με τον χρήστη, καθώς και τη δυνατότητά του να εκτελεί αποδοτικά τις απαιτούμενες εργασίες. Παρακάτω παραθέτονται οι τεχνικές απαιτήσεις πάνω στις οποίες βασίστηκε η ανάπτυξη του συστήματος:

- **Ευκολία και ταχύτητα εξοικείωσης με την διεπαφή χρήστη**: Το σύστημα πρέπει να σχεδιαστεί έτσι ώστε οι χρήστες να μπορούν να εκτελούν τις απαιτούμενες λειτουργίες με ευκολία και να εξοικειώνονται γρήγορα με τη χρήση του. Αυτό περιλαμβάνει μια κατανοητή δομή μενού και διεπαφή χρήστη που να είναι απλή και ευανάγνωστη. Επίσης, πρέπει να λαμβάνονται υπόψη οι βέλτιστες πρακτικές σχεδίασης διεπαφής χρήστη για να βοηθηθούν οι χρήστες να κατανοήσουν και να χειριστούν την εφαρμογή με ευκολία.

- **Απόδοση και ταχύτητα απόκρισης**: Το σύστημα πρέπει να είναι αποδοτικό και να παρέχει γρήγορη απόκριση στις εντολές των χρηστών. Αυτό περιλαμβάνει την αποτελεσματική χρήση των πόρων του συστήματος και τη βελτιστοποίηση του κώδικα των επί

μέρους στοιχείων του συστήματος ώστε οι χρόνοι φόρτωσης των δεδομένων να ελαχιστοποιούνται για να παρέχεται μια ομαλή και άριστη εμπειρία χρήσης.

- **Επεκτασιμότητα**: Η εφαρμογή πρέπει να είναι ευέλικτη και επεκτάσιμη, έτσι ώστε να μπορεί να προσαρμόζεται σε μελλοντικές ανάγκες και απαιτήσεις. Αυτό περιλαμβάνει την καλή δομή και σχεδίαση του κώδικα, τη χρήση σύγχρονων τεχνολογιών και προτύπων, καθώς και τη δυνατότητα επέκτασης λειτουργικοτήτων με ευκολία.

4.1.2 Επιχειρηματικές απαιτήσεις συστήματος (Business requirements)

Οι επιχειρηματικές απαιτήσεις αποτελούν το θεμέλιο για την ανάπτυξη μιας εφαρμογής που θα είναι χρήσιμη, αξιόπιστη και θα προσφέρει όντως αξία στους χρήστες. Παρακάτω παραθέτονται οι επιχειρηματικές απαιτήσεις πάνω στις οποίες βασίστηκε η ανάπτυξη του συστήματος:

- **Διαθεσιμότητα χρονοσειρών από πολλά διαφορετικά domains**: Το σύστημα πρέπει να παρέχει χρονοσειρές από πολλά διαφορετικά πεδία πρόβλεψης (forecasting fields). Αυτό επιτρέπει στους χρήστες να αξιοποιούν το σύστημα για διαφορετικές εφαρμογές και ανάγκες πρόβλεψης σε διάφορους τομείς.

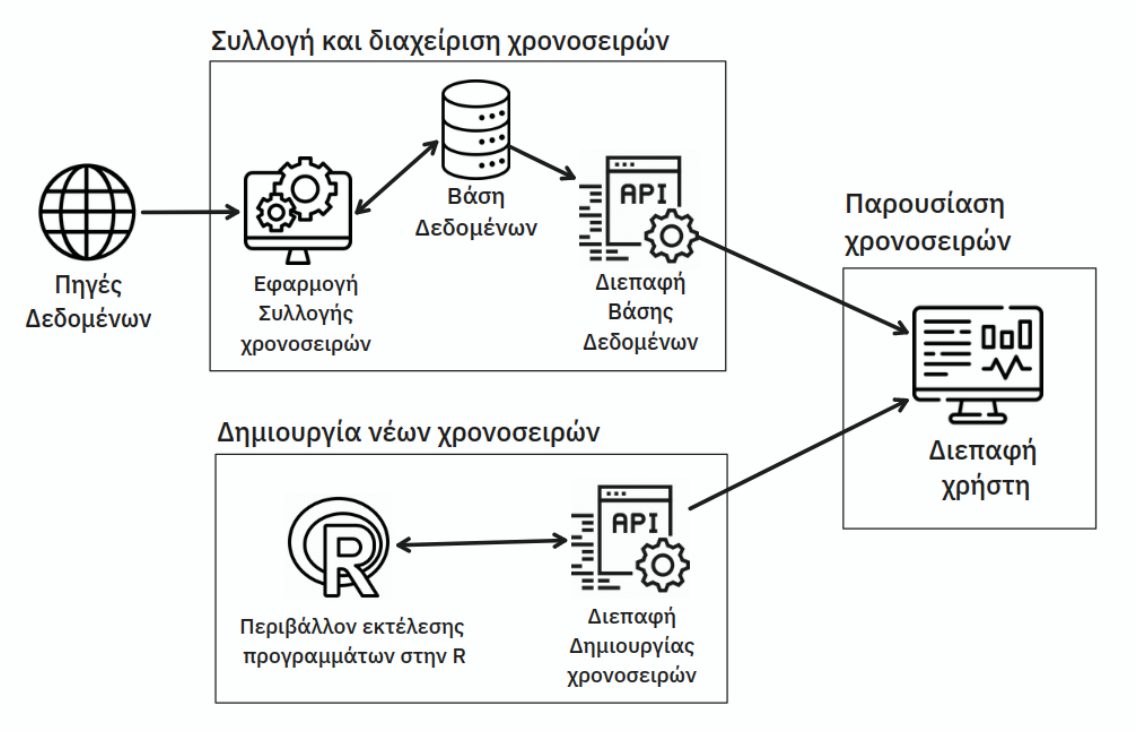
- **Ποιοτικά δεδομένα χρονοσειρών και ορθά υπολογισμένα χαρακτηριστικά**: Τα χαρακτηριστικά των χρονοσειρών είναι το βασικό κριτήριο με το οποίο ο χρήστης θα ζητάει χρονοσειρές από το σύστημα. Άρα είναι πολύ βασικό τα δεδομένα αυτά να είναι σωστά υπολογισμένα, αλλιώς η αξιοπιστία και η ακρίβεια των αναλύσεων μπορεί να επηρεαστούν αρνητικά. Εάν τα δεδομένα δεν είναι υπολογισμένα σωστά, οι χρήστες μπορεί να λάβουν λανθασμένες αποφάσεις ή να διαπιστώσουν ανακρίβειες στην ανάλυση τους.

- **Δυνατότητα λήψης χρονοσειρών**: Οι χρήστες πρέπει να έχουν τη δυνατότητα να κατεβάσουν τις χρονοσειρές που τους ενδιαφέρουν, μαζί με τα αντίστοιχα μεταδεδομένα που τις συνοδεύουν. Αυτό δίνει τη δυνατότητα στους χρήστες να διερευνήσουν περαιτέρω τα δεδομένα, να δημιουργήσουν δικές τους αναλύσεις και να διατηρήσουν ανεξάρτητα αντίγραφα των δεδομένων για αρχειοθέτηση ή επαναχρησιμοποίηση.

- **Ξεκάθαρη οπτικοποίηση χρονοσειρών**: Το σύστημα πρέπει να παρουσιάζει τις χρονοσειρές με τρόπο που να είναι εύκολα αναγνώσιμος και κατανοητός από τον χρήστη ώστε να μπορεί να συγκρίνει τα διαφορετικά αποτελέσματα που λαμβάνει από το σύστημα.

4.2 Αρχιτεκτονική Συστήματος

Το σύστημα έχει τρεις βασικούς πυλώνες: Συλλογή και διαχείριση, Δημιουργία και Παρουσίαση χρονοσειρών.



Σχήμα 4.1: *Αλληλεπίδραση επί μέρους στοιχείων του συστήματος*

Η **συλλογή και διαχείριση χρονοσειρών** βασίζεται σε μεγάλο βαθμό στο οικοσύστημα ανάπτυξης λογισμικού της Microsoft, καθώς οι επί μέρους εφαρμογές έχουν υλοποιηθεί στο Visual Studio, βασίζονται στο .NET framework για τις λειτουργίες τους και χρησιμοποιούν τον Microsoft SQL Server για την αποθήκευση των δεδομένων.

Συγκεκριμένα, τα επί μέρους κομμάτια της συλλογής και διαχείρισης χρονοσειρών είναι:

- **Εφαρμογή Συλλογής χρονοσειρών**: Αυτή η εφαρμογή συλλέγει τις χρονοσειρές από τις πηγές δεδομένων, τις φιλτράρει σύμφωνα με τα κριτήρια που έχουμε θέσει και τις αναλύει προκειμένου να υπολογίσει τα επιθυμητά χαρακτηριστικά σύμφωνα με την διαδικασία που περιγράφηκε στο Κεφάλαιο 3. Στη συνέχεια, αποθηκεύει αυτές τις χρονοσειρές στη βάση δεδομένων. Για την υλοποίηση της χρησιμοποιήθηκε η γλώσσα προγραμματισμού C# .NET και το Entity Framework για την αλληλεπίδραση με την βάση δεδομένων.

- **Βάση Δεδομένων**: Η βάση δεδομένων αποτελεί την καρδιά του συστήματος. Περιέχει όλα τα δεδομένα που χρειαζόμαστε για τις απαιτήσεις της λύσης μας. Χρησιμοποιήθηκε ο Microsoft SQL Server για την βάση δεδομένων, και το γραφικό περιβάλλον SQL Server Management Studio για την διαχείριση της.

- **Διεπαφή Βάσης Δεδομένων**: Επιτρέπει την πρόσβαση στις χρονοσειρές που υπάρχουν στην βάση δεδομένων. Για την υλοποίηση της διεπαφής χρησιμοποιήθηκε η γλώσσα προγραμματισμού C# .NET σε συνδυασμό με το ASP.NET framework για την δημιουργία της διαδικτυακής Διεπαφής Προγραμματισμού Εφαρμογών (αγγλ. Web API από το Application Programming Interface) και το Entity Framework για την αλληλεπίδραση με την βάση δεδομένων.

Η **δημιουργία χρονοσειρών** είναι εντελώς ανεξάρτητη από το υπόλοιπο σύστημα και αποτελείται από την **Διεπαφή Δημιουργίας χρονοσειρών** η οποία αποτελείται από έναν Node.js διακομιστή (server) του οποίο η βασική λειτουργία είναι η αμφίδρομη επικοινωνία με ένα περιβάλλον εκτέλεσης προγραμμάτων στην γλώσσα προγραμματισμού R [24] και συγκεκριμένα η αξιοποίηση της βιβλιοθήκης gratis για την δημιουργία μοναδικών χρονοσειρών σύμφωνα με τα χαρακτηριστικά που εμείς επιθυμούμε και ζητάμε μέσω τις διεπαφής.

Η **παρουσίαση χρονοσειρών** γίνεται μέσω της **Διεπαφής χρήστη (User Interface)**, μιας διαδικτυακής εφαρμογής (web app) υλοποιημένης με την χρήση της javascript βιβλιοθήκης React, η οποία χρησιμοποιεί τις προαναφερθείσες διεπαφές για να παρουσιάσει στον χρήστη χρονοσειρές και άλλα σχετικά δεδομένα σύμφωνα με τα κριτήρια που αυτός θα θέσει.

Στις πιο κάτω υποενότητες γίνεται μια σύντομη αναφορά στα επί μέρους εργαλεία που χρησιμοποιήθηκαν για την ανάπτυξη του συστήματος.

4.2.1 .NET framework και Visual Studio

Το .NET είναι ένα πλαίσιο λογισμικού (framework) που αναπτύχθηκε από τη Microsoft. Παρέχει ένα περιβάλλον εκτέλεσης και μια συλλογή βιβλιοθηκών που επιτρέπουν την ανάπτυξη μια μεγάλης γκάμας εφαρμογών για διάφορες πλατφόρμες.

Η **γλώσσα προγραμματισμού C#** (προφέρεται C-sharp) είναι μια γλώσσα προγραμματισμού που αναπτύχθηκε από τη Microsoft και είναι μια από τις κύριες γλώσσες προγραμματισμού που χρησιμοποιούνται στο .NET framework. Η C# έχει σχεδιαστεί να είναι μια απλή, σύγχρονη και πολυχρηστική γλώσσα με συντακτική σαφήνεια, ισχυρές δυνατότητες αντικειμενοστραφούς προγραμματισμού και υποστήριξη για διαχείριση μνήμης. Είναι μια από τις πιο δημοφιλής γλώσσες προγραμματισμού καθώς η ενσωμάτωση της με το .NET framework επιτρέπει στους προγραμματιστές να αξιοποιήσουν το πλούσιο σύνολο βιβλιοθηκών και υπηρεσιών του για τη δημιουργία ισχυρών και επεκτάσιμων εφαρμογών.

Το **ASP.NET** επεκτείνει την πλατφόρμα .NET με εργαλεία και βιβλιοθήκες ειδικά για τη δημιουργία διαδικτυακών εφαρμογών (web apps) και υπηρεσιών (services) με την χρήση της C#, κάνοντας το μια ιδανική επιλογή για τις ανάγκες ανάπτυξης διαδικτυακών διεπαφών (Web APIs).

Το **Entity Framework** αποτελεί μια βιβλιοθήκη αντιστοίχισης αντικειμένων-σχέσεων (Object-Relational Mapping, ORM) που αναπτύχθηκε από την Microsoft. Αυτό το εργαλείο επιτρέπει στους προγραμματιστές να αλληλεπιδρούν με τη βάση δεδομένων χρησιμοποιώντας αντικείμενα και κλάσεις, αποφεύγοντας την ανάγκη για απευθείας γραφή SQL ερωτημάτων. Ενσωματώνοντας το Entity Framework στο πλαίσιο του .NET, οι προγραμμα-

τιστές μπορούν να οργανώσουν τα δεδομένα των εφαρμογών τους με πιο αντικειμενοστραφή τρόπο, προσφέροντας ταυτόχρονα αυξημένη ασφάλεια και ευελιξία στις βάσεις δεδομένων τους. Επιπλέον, το Entity Framework συνδυάζεται άριστα με τις δυνατότητες του ASP.NET για τη δημιουργία σύγχρονων διαδικτυακών εφαρμογών και υπηρεσιών, προσφέροντας ένα ολοκληρωμένο περιβάλλον για την ανάπτυξη προηγμένων λύσεων που βασίζονται σε βάσεις δεδομένων.

Το **Visual Studio** είναι ένα Ολοκληρωμένο Περιβάλλον Ανάπτυξης (Integrated Development Environment, IDE) που επίσης αναπτύχθηκε από την Microsoft και γιαυτό έχει πολύ καλή ενσωμάτωση με τα υπόλοιπα προϊόντα της εταιρείας, κάτι που το κάνει το πλέον καταλληλότερο και ισχυρότερο εργαλείο για την ανάπτυξη εφαρμογών που κάνουν χρήση του .NET framework αλλά και του Microsoft SQL Server.

4.2.2 Microsoft SQL Server

Ο Microsoft SQL Server είναι ένα σύστημα διαχείρισης βάσεων δεδομένων που αναπτύχθηκε από την Microsoft. Πρόκειται για μια ολοκληρωμένη και πλούσια σε δυνατότητες πλατφόρμα για βάσεις δεδομένων που χρησιμοποιείται ευρέως για την αποθήκευση, διαχείριση και ανάκτηση δεδομένων σε διάφορες εφαρμογές και κλάδους.

Το SQL Server Management Studio (SSMS) είναι ένα γραφικό εργαλείο διεπαφής χρήστη (GUI) που παρέχεται από τη Microsoft για τη διαχείριση και την επίβλεψη των βάσεων δεδομένων SQL Server. Χρησιμοποιείται για την εκτέλεση μιας ευρείας γκάμας εργασιών, συμπεριλαμβανομένης της σχεδίασης και τροποποίησης της δομής της βάσης δεδομένων, καθώς και της ανάπτυξης και εκτέλεσης SQL ερωτημάτων (queries) και αποθηκευμένων διαδικασιών (stored procedures). Δίνει επίσης την δυνατότητα διαχείρισης θεμάτων ασφαλείας, όπως οι άδειες πρόσβασης χρηστών και η δημιουργία εφεδρικών αντιγράφων (backups).

Ουσιαστικά, ο SQL Server είναι το ίδιο το σύστημα που αποθηκεύει και διαχειρίζεται τα δεδομένα, ενώ το SQL Server Management Studio είναι το εργαλείο που επιτρέπει στους χρήστες να αλληλεπιδρούν με τον SQL Server μέσω μιας φιλικής προς τον χρήστη διεπαφής, καθιστώντας ευκολότερη την εκτέλεση εργασιών και τη διαχείριση της βάσης δεδομένων με αποτελεσματικό τρόπο.

4.2.3 Node.js

Η Node.js είναι μια ανοιχτού κώδικα πλατφόρμα (περιβάλλον εκτέλεσης και βιβλιοθήκη) που βασίζεται στη γλώσσα JavaScript και επιτρέπει την ανάπτυξη εκτελέσιμων προγραμμάτων σε εξυπηρετητές (serverside). Αντίθετα με την παραδοσιακή χρήση της JavaScript στη διεπαφή του χρήστη (client-side), το Node.js επιτρέπει την εκτέλεση JavaScript κώδικα στον εξυπηρετητή, επιτρέποντας τη δημιουργία δυναμικών ιστοσελίδων και εφαρμογών.

Ο ρόλος του Node.js είναι σημαντικός στην ανάπτυξη διαδικτυακών εφαρμογών και δικτυακών υπηρεσιών, καθώς η βασισμένη σε συμβάντα (event-driven) αρχιτεκτονική του παρέχει ένα περιβάλλον εκτέλεσης που είναι εξαιρετικά αποτελεσματικό στην αντιμετώπιση πολλαπλών συνδέσεων και επεξεργασία μεγάλου όγκου αιτημάτων σε πραγματικό χρόνο. Επιπλέον, το Node.js διαθέτει μια πλούσια συλλογή από πακέτα (packages) και βιβλιοθήκες,

τα οποία διευκολύνουν την ανάπτυξη εφαρμογών υψηλής απόδοσης και κλιμακωσιμότητας (scalability).

Συγκεκριμένα, **για τις ανάγκες του δικού μας συστήματος, χρησιμοποιήσαμε τη βιβλιοθήκη r-script**, η οποία προσφέρει την δυνατότητα μετάδοσης δεδομένων από και προς το περιβάλλον εκτέλεσης προγραμμάτων της γλώσσας προγραμματισμού R, επιτρέποντας έτσι την αξιοποίηση των ισχυρών εργαλείων πρόβλεψης και επεξεργασίας χρονοσειρών που ήδη υπάρχουν στο περιβάλλον της R.

Συνολικά, το Node.js αποτελεί ένα ισχυρό εργαλείο για την ανάπτυξη διαδικτυακών εφαρμογών που απαιτούν υψηλές αποκρίσεις και αποδοτική χρήση πόρων. Με τη δυνατότητα ασύγχρονης επεξεργασίας, την εκτέλεση σε πολλαπλά νήματα (multithreading) και την ευελιξία της JavaScript, οι προγραμματιστές μπορούν να αναπτύξουν γρήγορες και αποδοτικές εφαρμογές.

4.2.4 Γλώσσα προγραμματισμού R

Η R είναι μια ευρέως χρησιμοποιούμενη γλώσσα προγραμματισμού και περιβάλλον λογισμικού για στατιστική ανάλυση και οπτικοποίηση δεδομένων.

Ένα από τα κύρια πλεονεκτήματα της R είναι η πλούσια συλλογή πακέτων (packages) που με εξειδικευμένες λειτουργίες και αλγορίθμους επιτρέπουν σε ερευνητές και αναλυτές να εξερευνήσουν, να χειριστούν και να μοντελοποιήσουν τα δεδομένα τους με αποτελεσματικό τρόπο. Πακέτα όπως forecast, tseries και gratis, είναι ο βασικός λόγος που χρησιμοποιούμε την R στο σύστημα μας καθώς οι δυνατότητες που παρέχουν για ανάλυση και πρόβλεψη χρονοσειρών είναι δύσκολο να αναπαραχθούν σε άλλες πλατφόρμες.

Σε συνδυασμό με την ευκολία ενσωμάτωσης της R σε μεγαλύτερες εφαρμογές και αλυσίδες επεξεργασίας δεδομένων την καθιστούν την ιδανική επιλογή για τις ανάγκες του συστήματος. Με βιβλιοθήκες όπως η RDotNet (C#) και r-script (Node.js) μπορέσαμε εύκολα να ενσωματώσουμε προγράμματα R στην Εφαρμογή Συλλογής χρονοσειρών και στην Διεπαφή Δημιουργίας χρονοσειρών αντίστοιχα.

Τα προγράμματα (R scripts) που εκτελούνται στο σύστημα μας αναπτύχθηκαν χρησιμοποιώντας το RStudio, ένα Ολοκληρωμένο Περιβάλλον Ανάπτυξης λογισμικού (IDE) για την γλώσσα R, το οποίο με το διαδραστικό του περιβάλλον επιτρέπει την ευκολότερη χρήση της R και έτσι την ευκολότερη ανάπτυξη προγραμμάτων.

4.2.5 React

Η React είναι μια ανοιχτού κώδικα βιβλιοθήκη JavaScript που αναπτύχθηκε από την Meta (πρώην Facebook) και αποτελεί μια από τις πιο δημοφιλείς επιλογές για την κατασκευή σύγχρονων και ευέλικτων διεπαφών χρήστη.

Η React βασίζεται σε ένα σύστημα επαναχρησιμοποιήσιμων και αυτόνομων στοιχείων (component-based), κάτι που συμβάλει στην ευκολία συντήρησης και επέκτασης των εφαρμογών. Ο τρόπος λειτουργίας της React και η βάση της αποδοτικότητας της είναι ο τρόπος με τον οποίο ανιχνεύει και επανασχεδιάζει μόνο τα στοιχεία που έχουν υποστεί αλλαγές στην κατάστασή τους (state changes). Όταν η κατάσταση αλλάζει, η React εντοπίζει τις αλλαγές και ενημερώνει δυναμικά το UI, παρουσιάζοντας στον χρήστη τις ενημερωμένες πληροφορίες.

Συνολικά, η React προσφέρει έναν ευέλικτο, αποδοτικό και επεκτάσιμο τρόπο ανάπτυξης διεπαφών χρήστη, επιτρέποντας στους προγραμματιστές να δημιουργήσουν σύγχρονες και δυναμικές εφαρμογές που προσφέρουν μια εξαιρετική εμπειρία χρήστη.

4.3 Κατασκευή Βάσης Δεδομένων

Η δομή της βάσης δεδομένων βασίζεται σε ένα μοντέλο δεδομένων το οποίο ορίζει τον τρόπο που περιγράφονται τα δεδομένα, οι σχέσεις τους, η σημασία τους και οι περιορισμοί πάνω στα δεδομένα αυτά. Στην δικιά μας περίπτωση χρησιμοποιούμε το σχεσιακό μοντέλο (relational model) δεδομένων, το οποίο περιγράφει τα δεδομένα και τις αντίστοιχες σχέσεις τους ως ένα σύνολο πινάκων. Κάθε πίνακας (table) αποτελείται από στήλες (columns) με μοναδικά ονόματα. Μια εγγραφή (record) στον πίνακα αναπαριστά μια συσχέτιση (relationship) ανάμεσα σε ένα σύνολο τιμών.

Η δομή της βάσης δεδομένων καθορίστηκε από τις απαιτήσεις του συστήματος και σύμφωνα με την ομαδοποίηση των δεδομένων που επιλέξαμε να κάνουμε. Η συνολική εικόνα της βάσης δεδομένων φαίνεται στο Σχήμα 4.2 όπου παρουσιάζονται όλοι οι πίνακες και οι μεταξύ τους σχέσεις.

4.3.1 Πίνακες ανά συχνότητα χρονοσειράς

Η βασική ομαδοποίηση είναι σύμφωνα με την συχνότητα της κάθε χρονοσειράς και γιαυτό καταλήξαμε να έχουμε ένα ζευγάρι πινάκων για κάθε διαφορετική συχνότητα - ένας πίνακας για τα χαρακτηριστικά και ένας πίνακας για τις παρατηρήσεις των χρονοσειρών.

Οι δύο πίνακες (χαρακτηριστικά - παρατηρήσεις) συνδέονται με μια σχέση ένα προς πολλά (one to many) όπου μια εγγραφή από τον πίνακα χαρακτηριστικών αντιπροσωπεύει μια χρονοσειρά η οποία σχετίζεται με πολλές εγγραφές από τον πίνακα με τις παρατηρήσεις.

Πίνακες χαρακτηριστικών χρονοσειράς

Η βάση δεδομένων περιέχει έξι πίνακες χαρακτηριστικών, ένα για κάθε συχνότητα που μας ενδιαφέρει. Η αρχική προσέγγιση ήταν ένας πίνακας για όλες τις συχνότητες αλλά καθώς η ομαδοποίηση ανά συχνότητα ήταν απαραίτητη στο επόμενο επίπεδο (η Διεπαφή χρήστη φιλτράρει ανά συχνότητα) αυτό οδηγούσε σε περιττούς υπολογισμούς και αχρείαστες καθυστερήσεις στον διαμοιρασμό δεδομένων.

Οι πίνακες έχουν ακριβώς την ίδια μορφή και περιέχουν μεταδεδομένα που πήραμε κατευθείαν από την πηγή δεδομένων μας:

- Όνομα χρονοσειράς (name)
- Περιγραφή χρονοσειράς (description)
- Αριθμός παρατηρήσεων (number_of_observations)
- Πηγή δεδομένων (source)

- Βάση πηγής δεδομένων (database): αυτό αφορά πηγές δεδομένων όπως το Quandl το οποίο περιέχει πολλά διαφορετικά datasets, για παράδειγμα BUNDESBANK, ZILLOW, INSEE, κτλ.

καθώς και χαρακτηριστικά που εμείς υπολογίσαμε:

- Μοναδικός κωδικός χρονοσειράς (id)
- Ημερομηνία προσθήκης στην βάση δεδομένων (date_added)
- Κατηγορία χρονοσειράς (category_id)
- Ημερομηνία πρώτης παρατήρησης (starting_date)
- Ημερομηνία τελευταίας παρατήρησης (ending_date)
- Ελάχιστη τιμή παρατήρησης (min)
- Μέγιστη τιμή παρατήρησης (max)
- Μέση τιμή παρατήρησης (mean)
- Μέση τιμή των πρώτων διαφορών (avg_first_diff): Ο μέσος όρος των πρώτων διαφορών (first differences) των παρατηρήσεων
- Τα χαρακτηριστικά των χρονοσειρών όπως αυτά έχουν διατυπωθεί στο Κεφάλαιο 2.6.1, εξαιρουμένης της συχνότητας η οποία είναι η βάση ομαδοποίησης του πίνακα
  - Ένταση τυχαιότητας (entropy)
  - Ένταση τάσης (trend)
  - Ένταση εποχιακότητας (seasonal_strength)
  - Ένταση αυτοσυσχέτισης (e_acf1)
  - Στασιμότητα (stability)
- Οι κύριες συνιστώσες PC1 και PC2 όπως αυτές έχουν διατυπωθεί στο Κεφάλαιο 2.6.2

Πίνακες παρατηρήσεων χρονοσειράς

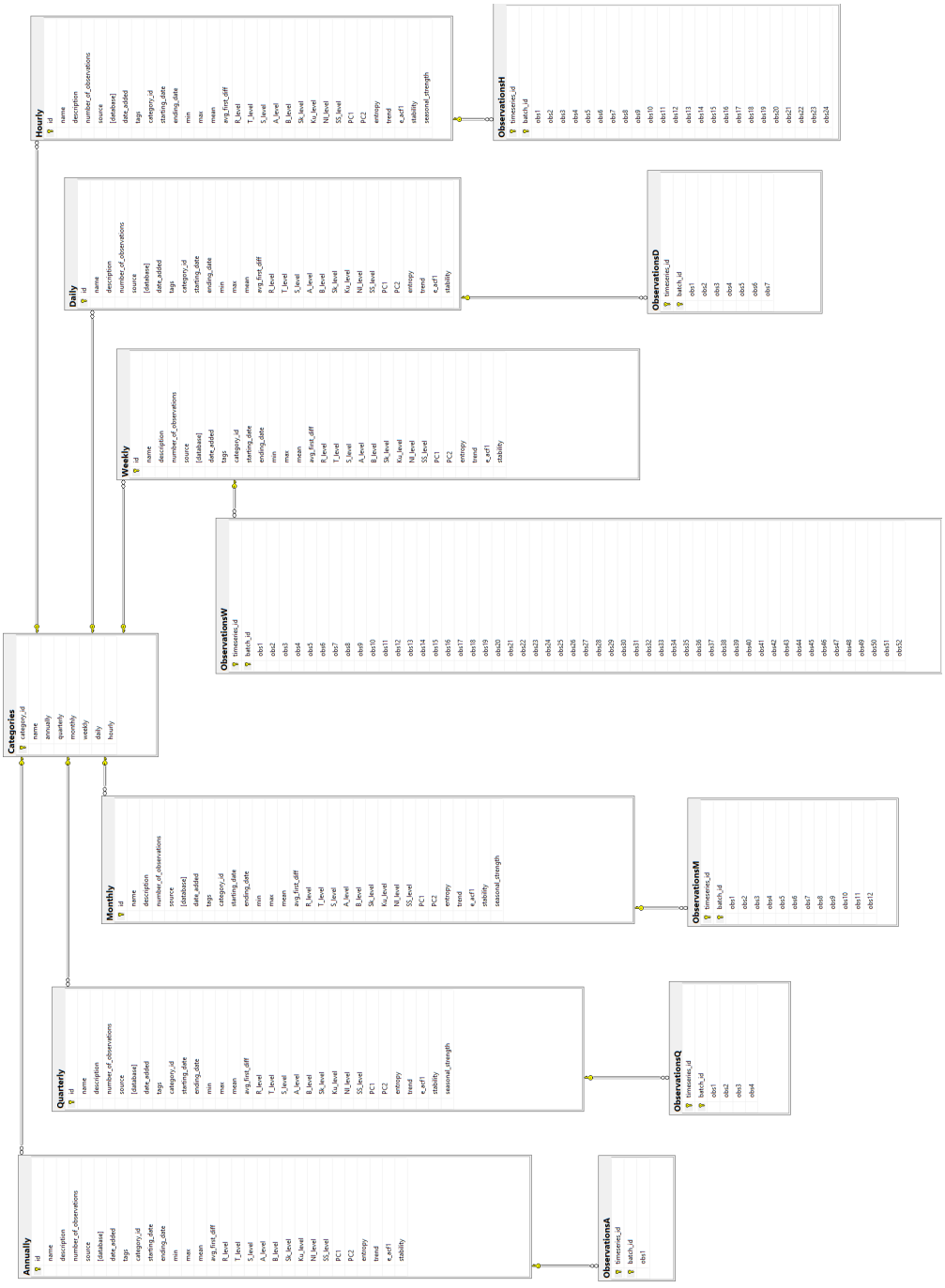
Η βάση δεδομένων περιέχει έξι πίνακες παρατηρήσεων, ένα για κάθε συχνότητας που μας ενδιαφέρει. Εδώ, σε αντίθεση με τους πίνακες χαρακτηριστικών, η μορφή των πινάκων έχει κάποιες διαφορές ανάλογα με την συχνότητα που αφορά.

Όταν οι παρατηρήσεις μια χρονοσειράς προστίθενται στον πίνακα, αυτό γίνεται σε παρτίδες (batches), με την κάθε παρτίδα να έχει μέγιστο μήκος όσο η περίοδος της εκάστοτε συχνότητας της χρονοσειράς. Για παράδειγμα ο πίνακας παρατηρήσεων για τις μηνιαίες χρονοσειρές έχει δώδεκα πεδία ανά εγγραφή, για τις εβδομαδιαίες έχει πενήντα δύο και για τις ετήσιες μόνο ένα. Άρα μια μηνιαία χρονοσειρά με 100 παρατηρήσεις θα χρειαστεί εννέα εγγραφές (παρτίδες) στον αντίστοιχο πίνακα παρατηρήσεων. Ο αύξων αριθμός παρτίδας (batch_id) καθορίζει την σειρά των παρατηρήσεων και σε συνδυασμό με τον μοναδικό κωδικό χρονοσειράς (timeseries_id) μας επιτρέπει να συσχετίσουμε τις παρατηρήσεις με την χρονοσειρά στην οποία ανήκουν.

4.3.2 Πίνακας Κατηγορίας χρονοσειρών

Η δευτερεύουσα ομαδοποίηση των χρονοσειρών βασίζεται στην κατηγορία στην οποία ανήκουν. Ο πίνακας αυτός περιέχει τις διαθέσιμες κατηγορίες και τον αριθμό χρονοσειρών ανά συχνότητα που ανήκουν στην κάθε κατηγορία.

Κάθε εγγραφή του πίνακα αντιπροσωπεύει μια κατηγορία και αποτελείται από τον μοναδικό κωδικό κατηγορίας (category_id), το όνομα της κατηγορίας (name) καθώς και ένα πεδίο ανά συχνότητα (annually, quarterly, monthly, weekly, daily, hourly) με τον συνολικό αριθμό χρονοσειρών που ανήκουν σε αυτή.



Σχήμα 4.2: *Διάγραμμα βάσης δεδομένων*

4.4 Παρουσίαση Διεπαφής Βάσης Δεδομένων

Η απευθείας αλληλεπίδραση μιας εφαρμογής με την βάση δεδομένων δεν συνηθίζεται και σε πολλές περιπτώσεις θεωρείται κακή πρακτική καθώς προκαλεί απώλεια ευελιξίας και ασφάλειας.

Για να αντιμετωπιστούν αυτά τα προβλήματα, χρησιμοποιούμε ένα επιπλέον επίπεδο επικοινωνίας που απλοποιεί τις πολυπλοκότητες του συστήματος διαχείρισης βάσεων δεδομένων και παρέχει μια απλοποιημένη διεπαφή που μπορούμε να χρησιμοποιήσουμε για πιο ασφαλή και ευέλικτη αλληλεπίδραση με τη βάση δεδομένων.

Η Διεπαφή Βάσης Δεδομένων συνδέεται απευθείας με την βάση δεδομένων και παρέχει συγκεκριμένα σημεία πρόσβασης (url endpoints) τα οποία η Διεπαφή Χρήστη χρησιμοποιεί για να αποκτήσει πρόσβαση στις πληροφορίες που χρειάζεται και τις οποίες ζητάει ο χρήστης.

Συγκεκριμένα, τα σημεία πρόσβασης είναι:

- **/api/timeseries/ids**: Φιλτράρει την βάση δεδομένων και επιστρέφει τους μοναδικούς κωδικούς (IDs) των χρονοσειρών που ανταποκρίνονται στις παραμέτρους που έχουν δοθεί. Οι διαθέσιμες παράμετροι φιλτραρίσματος είναι η κατηγορία (category), η συχνότητα (frequency) και προαιρετικά, ένα εύρος τιμής "min-max" για κάθε ένα από τα χαρακτηριστικά που έχουν διατυπωθεί στο Κεφάλαιο 2.2 (entropy, seasonality, trend, acf1, stability).

Αίτημα (Request)		Αποτέλεσμα (Response)	
Παράμετρος	Τύπος	Πεδίο	Τύπος
category*	integer	total	integer
frequency*	integer	max_observations	integer
entropy	string	ids	integer[]
seasonality	string		
trend	string		
acf1	string		
stability	string		

Πίνακας 4.1: *Πληροφορίες endpoint: api/timeseries/ids*

- **/api/timeseries**: Επιστρέφει τις χρονοσειρές (μόνο τις τιμές των παρατηρήσεων) σύμφωνα με την συχνότητα και τα IDs χρονοσειρών που δίνονται ως παράμετροι.

Αίτημα (Request)		Αποτέλεσμα (Response) []	
Παράμετρος	Τύπος	Πεδίο	Τύπος
frequency*	integer	foredeck_id	string
ids*	integer[]	name	string
		number_of_observations	integer
		starting_date	date
		ending_date	date
		Observations	double[]

Πίνακας 4.2: *Πληροφορίες endpoint: api/timeseries*

- **/api/timeseries/features**: Επιστρέφει τις χρονοσειρές (μόνο τις τιμές των χαρακτηριστικών) σύμφωνα με την συχνότητα και τα IDs χρονοσειρών που δίνονται ως παράμετροι.

Αίτημα (Request)		Αποτέλεσμα (Response) []	
Παράμετρος	Τύπος	Πεδίο	Τύπος
frequency*	integer	foredeck_id	string
ids*	integer[]	number_of_observations	integer
		entropy	double
		seasonality	double
		trend	double
		acf1	double
		stability	double

Πίνακας 4.3: *Πληροφορίες endpoint: api/timeseries/features*

- **/api/timeseries/pca**: Επιστρέφει τις κύριες συνιστώσες PC1 και PC2 για όλες τις χρονοσειρές της συχνότητας που δίνεται ως παράμετρος. Αυτή η πληροφορία μπορεί να χρησιμοποιηθεί για οπτικοποίηση και ανάλυση των χρονοσειρών σε δύο διαστάσεις.

Αίτημα (Request)		Αποτέλεσμα (Response) []	
Παράμετρος	Τύπος	Πεδίο	Τύπος
frequency*	integer	-	double[]

Πίνακας 4.4: *Πληροφορίες endpoint: api/timeseries/pca*

Όπως φαίνεται από τις περιγραφές των σημείων πρόσβασης, το φιλτράρισμα χρονοσειρών χωρίζεται σε δύο στάδια. Στο πρώτο στάδιο θέτουμε τα κριτήρια φιλτραρίσματος και λαμβάνουμε μόνο τα IDs των χρονοσειρών που μας ενδιαφέρουν και στο δεύτερο στάδιο χρησιμοποιούμε αυτά τα IDs για να ζητήσουμε την πλήρη πληροφορία. Αυτός ο τρόπος λειτουργίας επιτυγχάνει βελτίωση της απόδοσης του συστήματος και, κατ' επέκταση, της εμπειρίας του χρήστη, αφού επιτρέπει:

- Γρήγορη προεπισκόπηση των αποτελεσμάτων (πρώτο στάδιο): ο χρήστης λαμβάνει γρήγορα μια πρώτη εικόνα για την διαθεσιμότητα των δεδομένων που ζήτησε καθώς δεν απαιτείται η μεταφορά ολόκληρων των δεδομένων για κάθε αίτημα αναζήτησης.
- Γρήγορη ανάκτηση της πλήρους πληροφορίας των χρονοσειρών (δεύτερο στάδιο): Χρησιμοποιώντας τα αναγνωριστικά (IDs) που λάβαμε από το πρώτο στάδιο, εκμεταλλευόμαστε την υπάρχουσα δομή ευρετηρίου (index) στο πεδίο ID της βάσης δεδομένων για να ανακτήσουμε και να επιστρέψουμε γρήγορα τα δεδομένα των χρονοσειρών που μας ενδιαφέρουν.

4.5 Παρουσίαση Διεπαφής Δημιουργίας χρονοσειρών

Η δημιουργία νέων χρονοσειρών με ελεγχόμενα χαρακτηριστικά γίνεται βάση της μεθοδολογίας που περιγράφηκε στο Κεφάλαιο 2.7.2 και συγκεκριμένα με την χρήση του πακέτου

gratis που είναι η υλοποίηση των αλγορίθμων που περιγράφονται στην συγκεκριμένη δημοσίευση.

Για να μπορέσουμε να εκμεταλλευτούμε αυτό το πακέτο, αλλά και τις εκτεταμένες δυνατότητες για ανάλυση χρονοσειρών που υπάρχουν στο οικοσύστημα της R, υλοποιήθηκε η Διεπαφής Δημιουργίας χρονοσειρών η οποία έχει αμφίδρομη επικοινωνία με ένα περιβάλλον εκτέλεσης προγραμμάτων στην γλώσσα προγραμματισμού R. Η Διεπαφή παρέχει ένα και μόνο σημείο πρόσβασης (url endpoint) το οποίο λαμβάνει τις απαιτούμενες παραμέτρους και βάση αυτών εκτελεί το πρόγραμμα δημιουργίας νέων χρονοσειρών που έχει υλοποιηθεί.

Συγκεκριμένα, το σημείο πρόσβασης είναι το **api/timeseries/generate** το οποίο δέχεται τον συνολικό αριθμό χρονοσειρών (n) που θέλουμε να δημιουργηθούν, την συχνότητα (freq) και τον αριθμό παρατηρήσεων (length) που θέλουμε να έχουν, καθώς και ένα εύρος τιμής "min-max" για τουλάχιστο ένα από τα χαρακτηριστικά που έχουν διατυπωθεί στο Κεφάλαιο 2.6.1 (entropy, seasonality, trend, acf1, stability).

Αίτημα (Request)		Αποτέλεσμα (Response)[]	
Παράμετρος	Τύπος	Πεδίο	Τύπος
n*	integer	observations	double[]
freq*	integer	PC1	double
length*	integer	PC2	double
entropy	string	entropy	double
seasonality	string	seasonality	double
trend	string	trend	double
acf1	string	acf1	double
stability	string	stability	double

Πίνακας 4.5: *Πληροφορίες endpoint: api/timeseries/generate*

Η διεπαφή επικυρώνει (validates) τις παραμέτρους και στη συνέχεια, με χρήση της βιβλιοθήκης r-script [25], τις δίνει σαν όρισμα στο πρόγραμμα που βρίσκεται στο περιβάλλον εκτέλεσης της R. Το πρόγραμμα παράγει χρονοσειρές με τον εξής τρόπο:

1. Χρησιμοποιεί την μέθοδο generate_ts() της βιβλιοθήκης gratis [26] για να δημιουργήσει τυχαίες χρονοσειρές.
2. Υπολογίζει τα χαρακτηριστικά των νέων χρονοσειρών χρησιμοποιώντας την βιβλιοθήκη tsfeatures [27].
3. Κρατάει μόνο τις χρονοσειρές που έχουν χαρακτηριστικά εντός του εύρους που δόθηκε σαν παράμετρος και για αυτές υπολογίζει το τις κύριες συνιστώσες PC1, PC2 σύμφωνα με το PCA που έχει ήδη υπολογιστεί για τις υπάρχουσες χρονοσειρές.

Η εκτέλεση του προγράμματος επαναλαμβάνεται μέχρι να δημιουργηθεί ο συνολικός αριθμός χρονοσειρών και μετά αυτές επιστρέφονται σαν αποτέλεσμα του αιτήματος δημιουργίας χρονοσειρών.

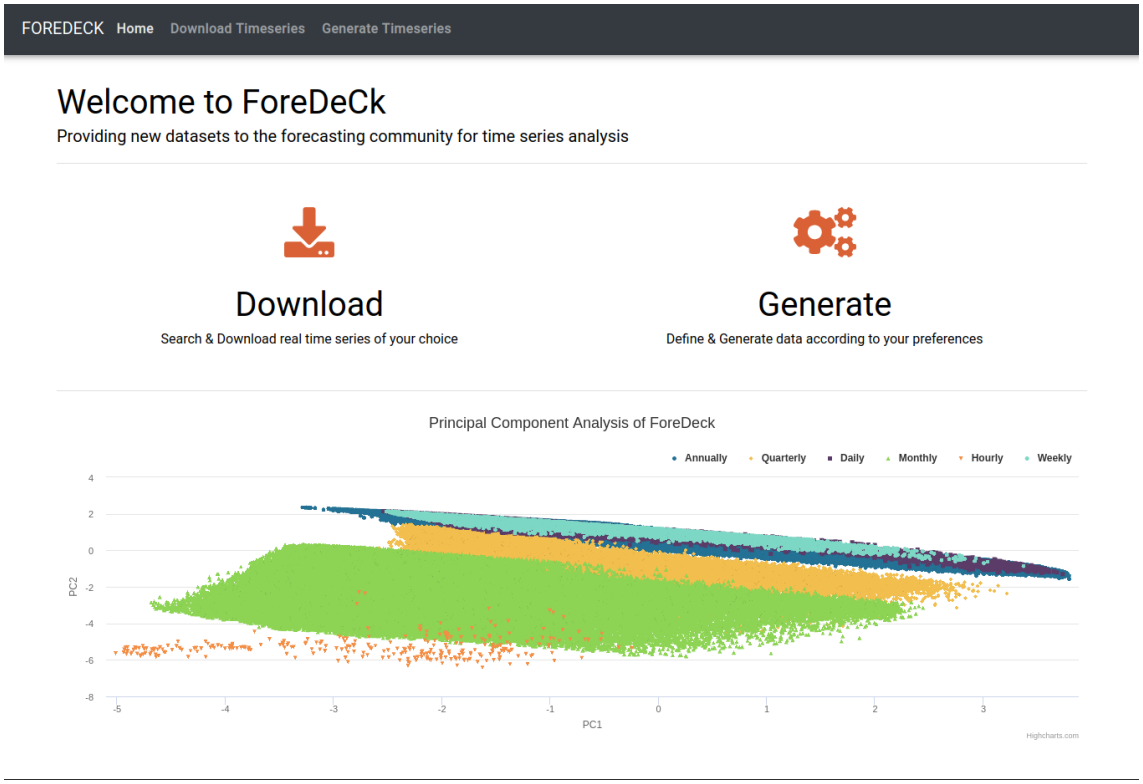
Η αρχική προσέγγιση για την δημιουργία των χρονοσειρών ήταν η χρήση της μεθόδου generate_target() της βιβλιοθήκη gratis η οποία δίνει την δυνατότητα απευθείας ορισμού των

χαρακτηριστικών που εμείς θέλουμε να έχουν οι χρονοσειρές που θα δημιουργήσει. Αλλά στην πράξη, όταν προσπαθήσαμε να δημιουργήσουμε περισσότερες από μερικές δεκάδες χρονοσειρές, αυτός ο τρόπος ήταν πολύ χρονοβόρος. Έτσι η προσέγγιση άλλαξε και καταλήξαμε στην πιο αποδοτική μέθοδο που περιγράφηκε πιο πάνω.

4.6 Παρουσίαση Διεπαφής χρήστη

Η διεπαφή χρήστη αποτελεί τον κεντρικό πυλώνα της λύσης του συστήματος, αφού συνδυάζει τα επί μέρους στοιχεία και προσφέρει την ολοκληρωμένη εμπειρία στον χρήστη. Στο παρόν κεφάλαιο, θα παρουσιάσουμε αναλυτικά την διεπαφή χρήστη και τις διαφορετικές λειτουργίες που προσφέρει.

Η αρχική σελίδα της εφαρμογής παρουσιάζει ξεκάθαρα στον χρήστη τις διαθέσιμες λειτουργίες. Ο χρήστης έχει τη δυνατότητα να αναζητήσει και να κατεβάσει πραγματικές χρονοσειρές ή να δημιουργήσει νέες χρονοσειρές ανάλογα με τις προτιμήσεις του. Παράλληλα, ο χρήστης έχει πρόσβαση σε μια οπτικοποίηση των διαθέσιμων χρονοσειρών, κάτι που το προσφέρει έναν απλό και κατανοητό τρόπο για να αντιληφθεί και να συγκρίνει τα δεδομένα που έχει στην διάθεση του.

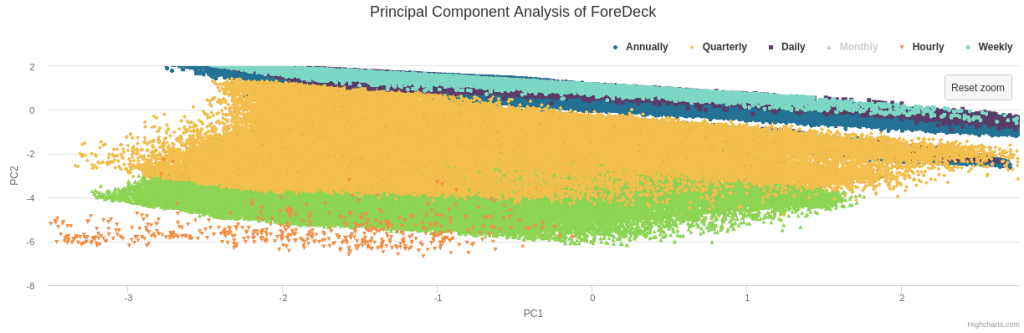


Σχήμα 4.3: *Αρχική σελίδα Διεπαφής Χρήστη*

4.6.1 Οπτικοποίηση Χρονοσειρών

Με το άνοιγμα της εφαρμογής, ξεκινά αυτόματα η λήψη των απαραίτητων δεδομένων *μέσω της διεπαφής βάσης δεδομένων* για την οπτικοποίηση των διαθέσιμων χρονοσειρών βάση των κύριων συνιστωσών του *Principal Component Analysis (PCA)*.

Ο χρήστης έχει τη δυνατότητα να αλληλεπιδράσει με το γράφημα προκειμένου να εξερευνήσει καλύτερα την κατανομή των δεδομένων. Μπορεί να πραγματοποιήσει ζουμ, δηλαδή να μεγεθύνει ή να σμικρύνει το γράφημα ή να φιλτράρει ανά συχνότητα, επιτρέποντάς του να εστιάσει σε συγκεκριμένες περιοχές.

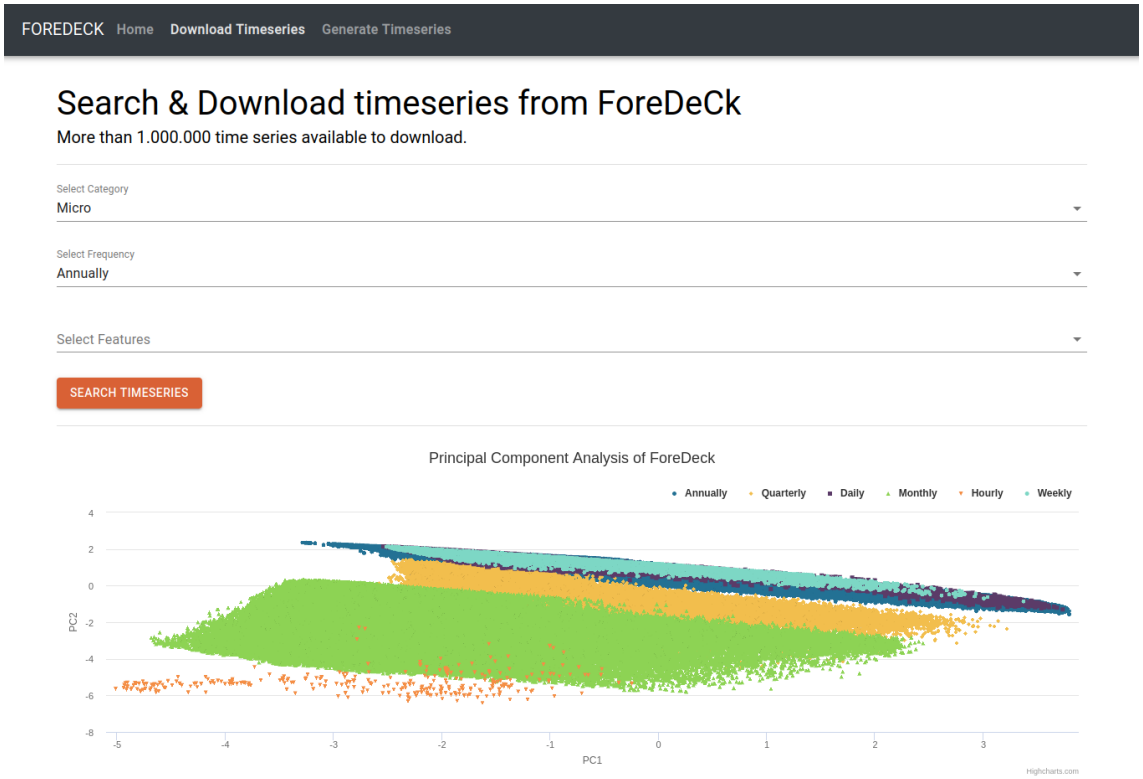


Σχήμα 4.4: *Οπτικοποίηση χρονοσειρών μέσω Principal Component Analysis (PCA)*

Η οπτικοποίηση είναι προσβάσιμη σε όλες σελίδες της εφαρμογής και, όπως θα δούμε στη συνέχεια, παρέχει επιπλέον δυνατότητες εξερεύνησης των δεδομένων, αφού κατά την λήψη και δημιουργία χρονοσειρών αυτές επίσης εμφανίζονται στο γράφημα.

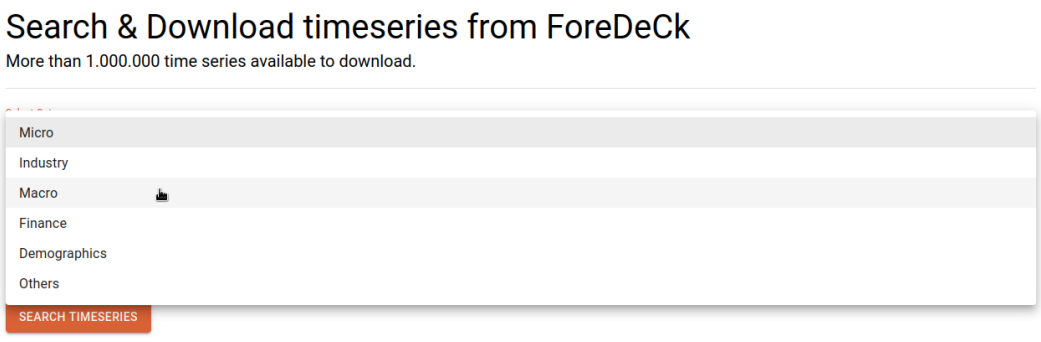
4.6.2 Αναζήτηση και Λήψη Χρονοσειρών

Στην σελίδα αναζήτησης και λήψης χρονοσειρών, ο χρήστης έχει τη δυνατότητα να αναζητήσει χρονοσειρές βάση των διαθέσιμων κριτηρίων, να πάρει μια πρώτη εικόνα για το αποτέλεσμα της αναζήτησης και στη συνέχεια να προχωρήσει στη λήψη των επιλεγμένων χρονοσειρών.

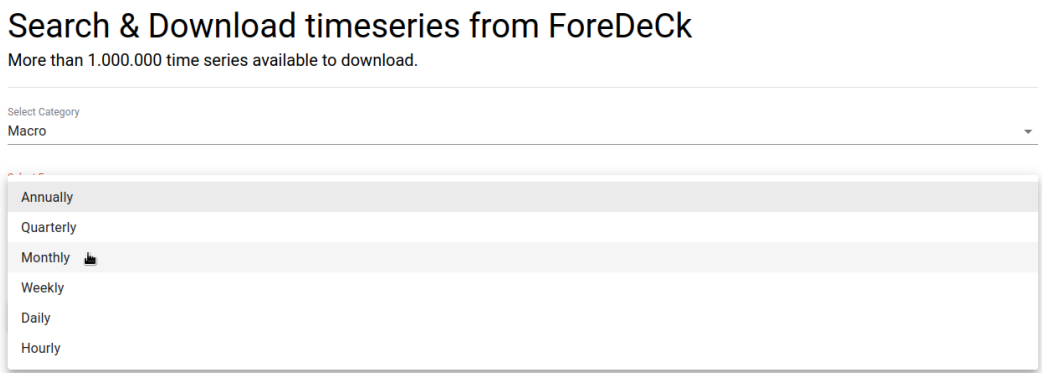


Σχήμα 4.5: *Σελίδα αναζήτησης και λήψης χρονοσειρών*

Ο χρήστης πρέπει αρχικά να καθορίσει την κατηγορία καθώς και την συχνότητα των χρονοσειρών που αναζητεί επιλέγοντας από τις προκαθορισμένες επιλογές που είναι διαθέσιμες.

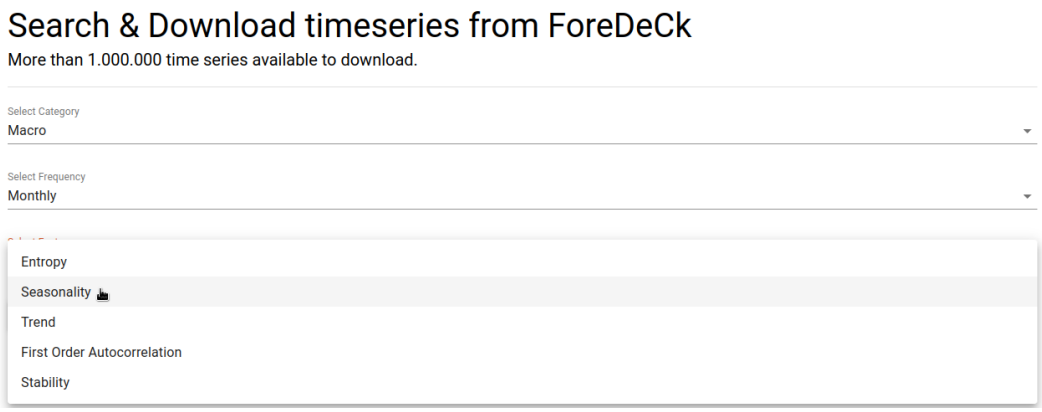


Σχήμα 4.6: *Διαθέσιμες κατηγορίες χρονοσειρών*



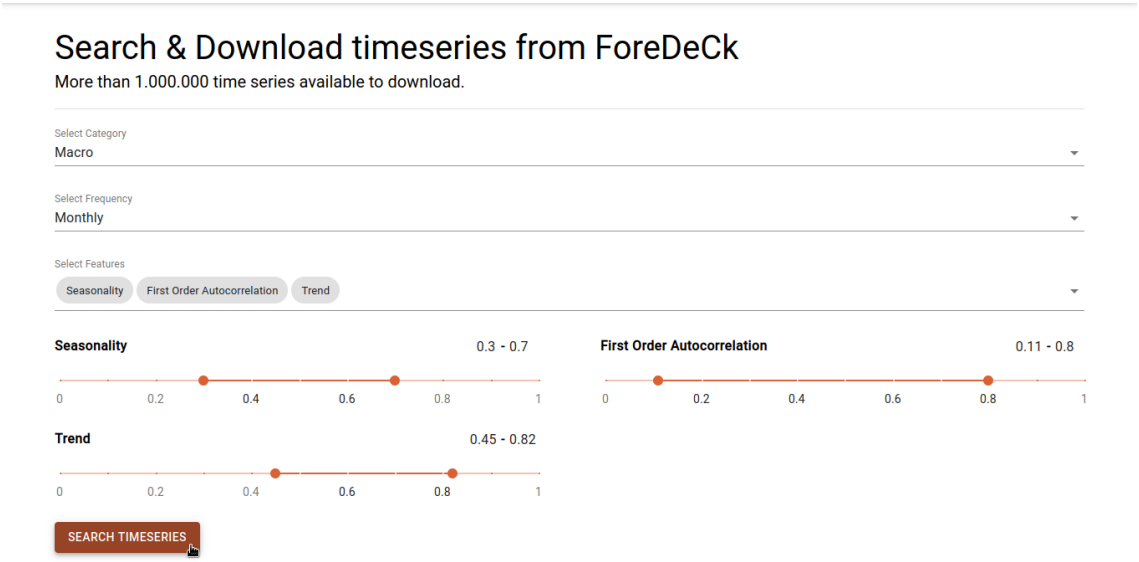
Σχήμα 4.7: *Διαθέσιμες συχνότητες χρονοσειρών*

Στη συνέχεια, μπορεί προαιρετικά να περιορίσει ακόμα περισσότερο την αναζήτηση, επιλέγοντας τα χαρακτηριστικά που επιθυμεί να πληρούν οι χρονοσειρές.



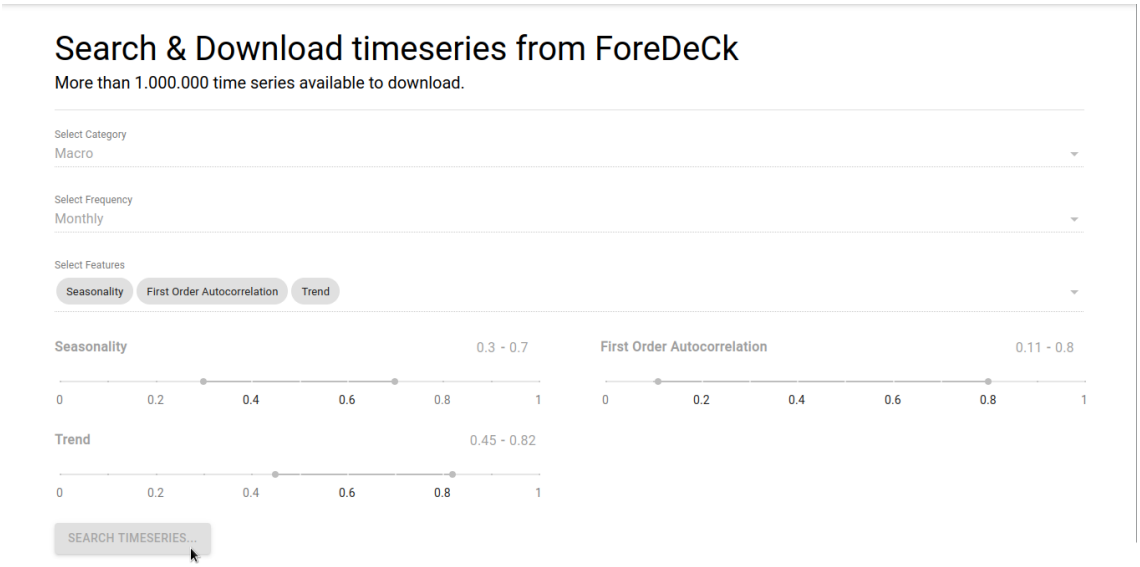
Σχήμα 4.8: *Διαθέσιμα χαρακτηριστικά χρονοσειρών*

Για τα χαρακτηριστικά που έχει επιλέξει, ο χρήστης έχει την ευελιξία να προσαρμόσει το εύρος τιμών με βάση τις ανάγκες του. Μπορεί να ορίσει τα κατώτερα και ανώτερα όρια των χαρακτηριστικών που τον ενδιαφέρουν μέσω του range slider που υπάρχει ή να πληκτρολογήσει το εύρος χειροκίνητα. Αφού θέσει όλα τα κριτήρια που τον ενδιαφέρουν, ο χρήστης μπορεί να προχωρήσει στην αναζήτηση των χρονοσειρών.



Σχήμα 4.9: *Ορισμός εύρους τιμών επιλεγμένων χαρακτηριστικών*

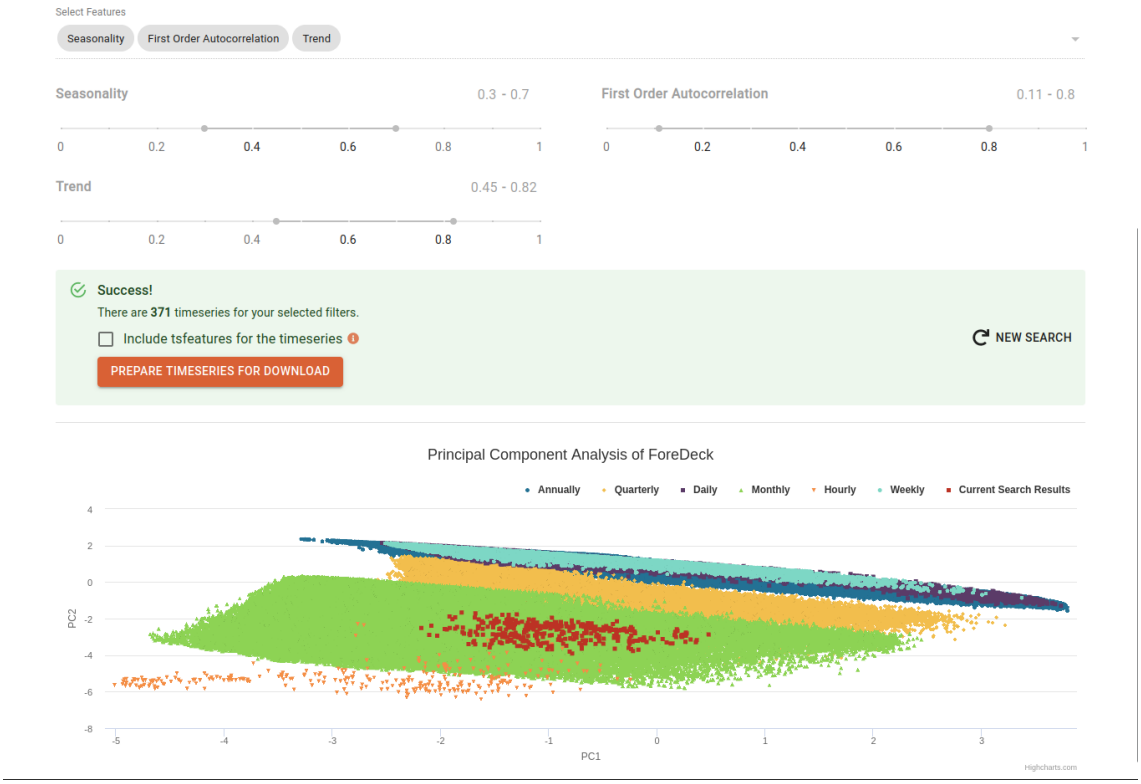
Με το πάτημα του κουμπιού, η εφαρμογή στέλνει το αντίστοιχο αίτημα στην διεπαφή βάσης δεδομένων και ταυτόχρονα παρέχει στον χρήστη ξεκάθαρες ενδείξεις ότι η αναζήτηση είναι σε εξέλιξη. Αυτό βοηθά τον χρήστη να γνωρίζει ότι πρέπει να περιμένει για την ολοκλήρωση της αναζήτησης και να μην ανησυχεί για τυχόν καθυστερήσεις ή απρόοπτα.



Σχήμα 4.10: *Αναζήτηση χρονοσειρών βάση κριτηρίων*

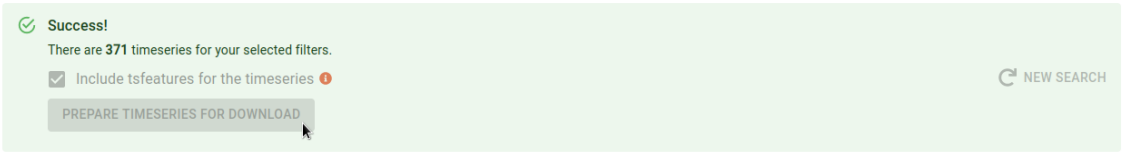
Με την ολοκλήρωση της αναζήτησης, ο χρήστης ενημερώνεται για τον αριθμό των χρονοσειρών που πληρούν τα κριτήρια που έθεσε. Επιπλέον, οι χρονοσειρές εμφανίζονται πάνω στο γράφημα, επιτρέποντας στον χρήστη να αντιληφθεί πού ακριβώς τοποθετούνται σε σχέση με το σύνολο των χρονοσειρών που υπάρχουν στο σύστημα. Αυτή η οπτική αναπαράσταση παρέχει στον χρήστη μια γρήγορη εποπτεία και κατανόηση των αποτελεσμάτων της αναζήτησης, επιτρέποντάς του να εστιάσει στις χρονοσειρές που τον ενδιαφέρουν περισσότερο, χωρίς να χρειάζεται να κατεβάσει ολόκληρες τις χρονοσειρές. Αυτό συμβάλλει στην εξοικονόμηση

χρόνου και πόρων του χρήστη αλλά και του συστήματος.



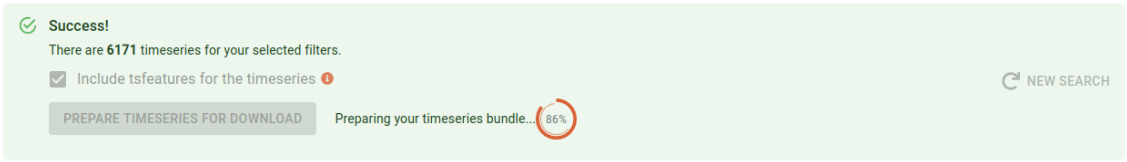
Σχήμα 4.11: *Προεπισκόπηση αναζήτησης χρονοσειρών*

Έχοντας την προεπισκόπηση της αναζήτησης στα χέρια του, ο χρήστης μπορεί επιλέξει αν θέλει να συνεχίσει την διαδικασία αναζήτησης ώστε να λάβει την πλήρη πληροφορία για τις χρονοσειρές που τον ενδιαφέρουν. Η βασική πληροφορία που θα λάβει είναι οι τιμές των παρατηρήσεων των χρονοσειρών και εάν επιλέξει, μπορεί να λάβει και τις τιμές των υπολογισμένων χαρακτηριστικών.



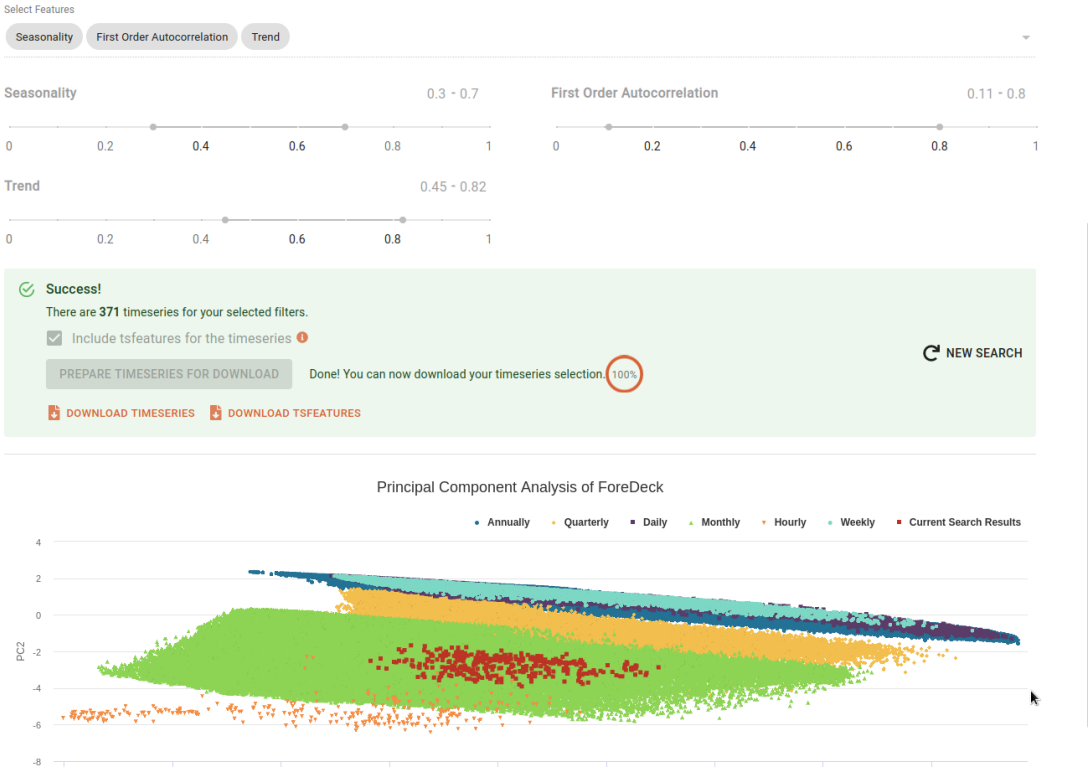
Σχήμα 4.12: *Επιλογή για προετοιμασία πλήρης πληροφορίας χρονοσειρών*

Με το πάτημα του κουμπιού, η εφαρμογή αποστέλλει αιτήματα στα σημεία πρόσβασης /api/timeseries και /api/timeseries/features της διεπαφής βάσης δεδομένων χρησιμοποιώντας τους μοναδικούς κωδικούς (IDs) των χρονοσειρών που έχουν ήδη ληφθεί στο πρώτο στάδιο της αναζήτησης. Αυτά τα αιτήματα γίνονται σε παρτίδες για να επιτευχθεί καλύτερη απόδοση του συστήματος και για να παρέχεται μια σαφής εικόνα στον χρήστη σχετικά με την πρόοδο της διαδικασίας, κάτι που είναι ιδιαίτερα χρήσιμο όταν ο αριθμός των επιλεγμένων χρονοσειρών είναι μεγάλος.



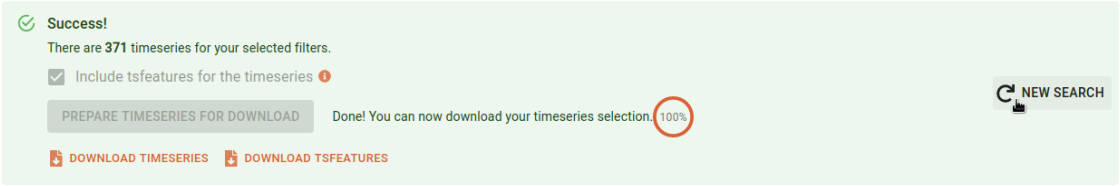
Σχήμα 4.13: *Πρόοδος προετοιμασίας πλήρης πληροφορίας χρονοσειρών*

Μετά την ολοκλήρωση της διαδικασίας αναζήτησης και προετοιμασίας των χρονοσειρών, ο χρήστης έχει τη δυνατότητα να κατεβάσει δύο διαφορετικά αρχεία σε μορφή .csv. Αυτά τα αρχεία περιέχουν την πλήρη πληροφορία για τις επιλεγμένες χρονοσειρές.



Σχήμα 4.14: *Αποτέλεσμα αναζήτησης και προετοιμασίας χρονοσειρών*

Ο χρήστης μπορεί οποιαδήποτε στιγμή να κάνει επαναφορά και να ξεκινήσει μια νέα αναζήτηση πατώντας το κουμπί "New Search".



Σχήμα 4.15: *Επιλογή νέας αναζήτησης*

Το πρώτο αρχείο περιέχει τις τιμές των παρατηρήσεων των χρονοσειρών και το δεύτερο αρχείο περιέχει τα χαρακτηριστικά των χρονοσειρών.

	A	B	C	D	E	F	G	H	I
1	Timeseries Id	Name	Number Of Observations	Starting Date	Ending Date	1.0000000	2.0000000	3.0000000	4.00
2	M1439	Average Weekly Hours of All Employees: Mining and Logging	130	2006-03-01T00:	2016-12-01T00:	42.1000000	42.9000000	42.5000000	43.20
3	M1742	Total Separations: Mining and Logging	192	2000-12-01T00:	2016-11-01T00:	28.0000000	27.0000000	13.0000000	19.00
4	M2220	Future Employment; Percentage Reporting Increases for Texas	151	2004-06-01T00:	2016-12-01T00:	46.1000000	50.0000000	44.3000000	46.50
5	M2246	Current Employment; Percentage Reporting No Change for Texas	151	2004-06-01T00:	2016-12-01T00:	59.0000000	50.0000000	59.2000000	63.00
6	M2270	Current Shipments; Percent Reporting Increases for New York	187	2001-07-01T00:	2017-01-01T00:	17.7000000	30.7000000	26.9000000	26.20
7	M2439	Future Growth Rate of Orders; Percentage Reporting Increases for Texas	151	2004-06-01T00:	2016-12-01T00:	51.3000000	50.7000000	55.7000000	56.30
8	M2441	Future New Orders; Percentage Reporting Increases for Texas	151	2004-06-01T00:	2016-12-01T00:	56.6000000	61.2000000	58.6000000	66.20
9	M2546	Future Shipments; Percentage Reporting Increases for Texas	151	2004-06-01T00:	2016-12-01T00:	61.8000000	70.1000000	67.6000000	71.40
10	M2549	Future Shipments; Percent Reporting Increases for FRB - Philadelphia District	585	1968-05-01T00:	2017-01-01T00:	56.0000000	65.0000000	46.0000000	47.00
11	M3943	Producer Price Index by Commodity for Travel Arrangement Services: Arrangement of Cruises and Tours	97	2008-12-01T00:	2016-12-01T00:	100.0000000	100.0000000	101.1000000	102.40
12	M3963	Producer Price Index by Industry: Hotels and Motels Except Casino Hotels: Other Receipts	73	2010-12-01T00:	2016-12-01T00:	100.0000000	98.8000000	98.1000000	98.20
13	M4455	Producer Price Index by Industry: Lessors of Nonresidential Buildings: Leasing of Other Nonresidential Buildings and Facilities	169	2002-12-01T00:	2016-12-01T00:	100.0000000	100.0000000	99.2000000	104.10
14	M4634	Producer Price Index by Industry: Nonscheduled Air Freight Chartering: Domestic Nonscheduled Freight Services	103	2008-06-01T00:	2016-12-01T00:	100.0000000	102.9000000	100.7000000	100.70
15	M4886	Import (Harmonized System): Refrigerators freezers heat pumps; and parts	121	2006-12-01T00:	2016-12-01T00:	100.0000000	102.7000000	102.7000000	102.40
16	M5150	Conditioned Wool Output for France	211	1922-01-01T00:	1939-07-01T00:	8732.0000000	7925.0000000	9366.0000000	9334.00
17	M5158	Slab Zinc Shipments for United States	99	1940-01-01T00:	1948-03-01T00:	59.8000000	53.9000000	52.8000000	50.10
18	M5179	Prepared Roofing Shipments for United States	117	1919-01-01T00:	1928-09-01T00:	1054.0000000	1253.0000000	1517.0000000	1641.00
19	M5180	Prepared Roofing Shipments for United States	138	1932-04-01T00:	1943-09-01T00:	2312.0000000	1661.0000000	1563.0000000	1749.00
20	M5202	Number of Residential Buildings Constructed and Inspected for Berlin Germany	132	1923-01-01T00:	1933-12-01T00:	97.0000000	144.0000000	238.0000000	135.00
21	M5210	Eastbound Freight Shipments for Chicago IL	154	1887-09-01T00:	1900-06-01T00:	46.0000000	41.2000000	48.5000000	53.90
22	M5237	Railway Freight Traffic Total Including Free Hauled for Great Britain	144	1920-01-01T00:	1931-12-01T00:	1673.0000000	1577.0000000	1811.0000000	1575.00
23	M5244	Loaded Wagon-Miles for Great Britain	144	1920-01-01T00:	1931-12-01T00:	310.0000000	291.0000000	330.0000000	289.00
24	M5246	Net Ton-Miles Per Freight Train-Mile for Great Britain	144	1920-01-01T00:	1931-12-01T00:	131.5000000	133.3000000	136.1000000	133.60
25	M5249	Freight Train-Hours for Great Britain	144	1920-01-01T00:	1931-12-01T00:	1725.0000000	1626.0000000	1782.0000000	1513.00
26	M5429	Raw Silk Stocks at Warehouses for New York NY and Hoboken NJ	261	1919-12-01T00:	1941-08-01T00:	77.6000000	68.0000000	65.0000000	52.80
27	M5543	Foodstuffs Imports Value for France	120	1878-01-01T00:	1887-12-01T00:	69.4000000	77.5000000	83.7000000	92.00
28	M5566	Raw Silk Imports for United States	254	1919-01-01T00:	1940-02-01T00:	2781.0000000	3103.0000000	2114.0000000	3394.00
29	M5567	Raw Cotton Exports for United States	216	1938-01-01T00:	1955-12-01T00:	673.0000000	420.2000000	450.1000000	402.70
30	M5574	Positions Open Illinois Free Employment Offices for Illinois	173	1920-08-01T00:	1934-12-01T00:	24635.0000000	22891.0000000	20375.0000000	14540.00
31	M5661	Average Actual Hours of Work Per Week Rubber Products Manufacturing for United States	120	1947-01-01T00:	1956-12-01T00:	40.8000000	40.6000000	39.9000000	39.50

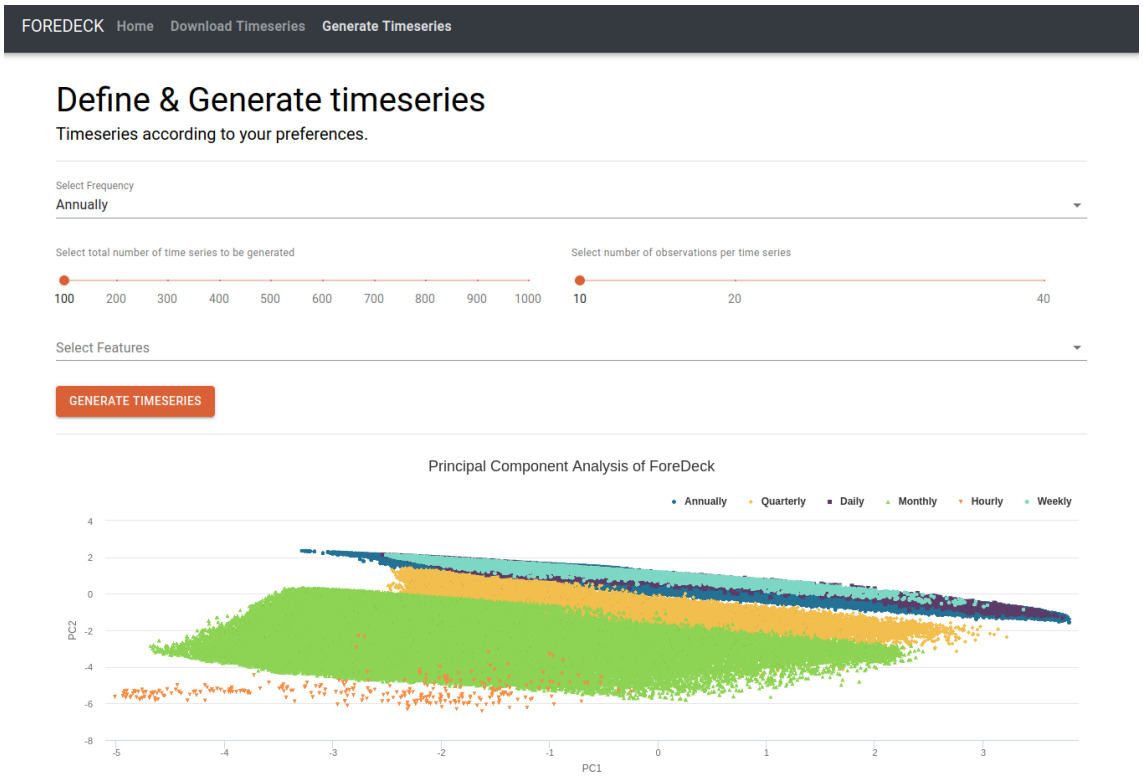
Σχήμα 4.16: *Ανακτηθέντες βασικές πληροφορίες χρουοσειρών*

	A	B	C	D	E	F	G
1	TimeseriesId	NumberOfObservations	entropy	trend	e_acf1	stability	seasonal_strength
2	M1439	130	0.749412	0.709334	0.187945	1	0
3	M1742	192	0.736976	0.788493	0.142938	1	1
4	M2220	151	0.805158	0.726406	0.150583	1	0
5	M2246	151	0.933968	0.522932	0.128093	0	0
6	M2270	187	0.891678	0.578922	0.13084	0	1
7	M2439	151	0.852593	0.671319	0.198367	1	0
8	M2441	151	0.807672	0.693661	0.19006	1	0
9	M2546	151	0.724503	0.790091	0.222682	1	0
10	M2549	585	0.760149	0.812467	0.274665	1	1
11	M3943	97	0.695614	0.771665	0.5338	1	0
12	M3963	73	0.771581	0.702807	0.452612	1	0
13	M4455	169	0.55099	0.769517	0.297139	1	1
14	M4634	103	0.771105	0.736635	0.361745	0	0
15	M4886	121	0.717184	0.819145	0.376013	1	0
16	M5150	211	0.86424	0.585285	0.317463	0	0
17	M5158	99	0.858346	0.575324	0.470228	0	0
18	M5179	117	0.83035	0.480427	0.24392	0	1
19	M5180	138	0.613104	0.770171	0.173309	1	1
20	M5202	132	0.772909	0.621239	0.215626	1	0
21	M5210	154	0.774078	0.677011	0.260275	0	1
22	M5237	144	0.850184	0.585341	0.46662	0	0
23	M5244	144	0.872533	0.593215	0.316393	0	0
24	M5246	144	0.846783	0.503433	0.524206	0	0
25	M5249	144	0.840502	0.599439	0.468525	0	0
26	M5429	261	0.716469	0.799305	0.60551	1	1
27	M5543	120	0.826086	0.649523	0.147486	0	1
28	M5566	254	0.779411	0.659915	0.180039	1	0
29	M5567	216	0.7898	0.700787	0.534101	0	0
30	M5574	173	0.790411	0.745421	0.242032	0	0
31	M5661	120	0.722315	0.767631	0.575705	1	0

Σχήμα 4.17: *Ανακτηθέντες πληροφορίες για χαρακτηριστικά χρουοσειρών*

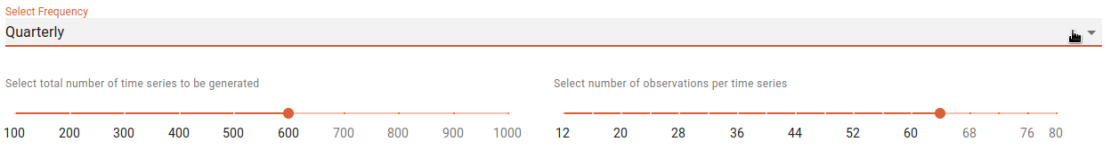
4.6.3 Δημιουργία Χρονοσειρών

Στην σελίδα δημιουργίας χρονοσειρών, ο χρήστης έχει τη δυνατότητα να δημιουργήσει προσαρμοσμένες χρονοσειρές βάση των χαρακτηριστικών που αυτός επιθυμεί να έχουν.



Σχήμα 4.18: *Σελίδα δημιουργίας χρονοσειρών*

Ο χρήστης πρέπει αρχικά να καθορίσει την συχνότητα, τον συνολικό αριθμό χρονοσειρών που θέλει να δημιουργήσει καθώς και τον αριθμό των παρατηρήσεων που θα έχει κάθε χρονοσειρά.



Σχήμα 4.19: *Επιλογές για δημιουργία τριμηνιαίων χρονοσειρών*

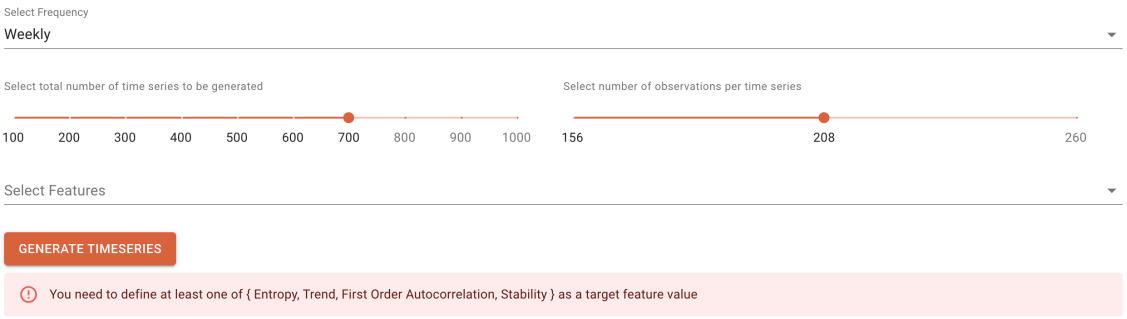
Οι επιλογές που προσφέρονται είναι προκαθορισμένες, με τις διαθέσιμες επιλογές για τον αριθμό παρατηρήσεων της χρονοσειράς να αλλάζουν ανάλογα με την επιλεγμένη συχνότητα.

Αυτός ο περιορισμός οφείλεται στον τρόπο λειτουργίας της διεπαφής δημιουργίας χρονοσειρών. Κατά την ανάπτυξη του συστήματος, πραγματοποιήθηκαν αρκετές δοκιμές που έδειξαν πώς με την αύξηση του συνολικού αριθμού των χρονοσειρών και παρατηρήσεων προς δημιουργία, το πρόγραμμα φτάνει σε ένα όριο. Σε αυτό το όριο, ο χρόνος που απαιτείται για τη δημιουργία γίνεται πολύ μεγάλος, υποβαθμίζοντας την εμπειρία αλληλεπίδρασης του χρήστη με το σύστημα. Για να αποφύγουμε αυτό το σενάριο, πραγματοποιήθηκαν μετρήσεις και αξιολογήσεις (benchmarks), και καταλήξαμε σε αυτές τις προκαθορισμένες επιλογές για

κάθε συχνότητα. Αυτό γίνεται προκειμένου να διασφαλιστεί η βέλτιστη απόδοση και η ευχρηστία του συστήματος, παρέχοντας συνάμα στον χρήστη την ευελιξία να προσαρμόσει την ποσότητα και την ακρίβεια των δημιουργούμενων χρονοσειρών, ανάλογα με τις ανάγκες και τις προτιμήσεις του.

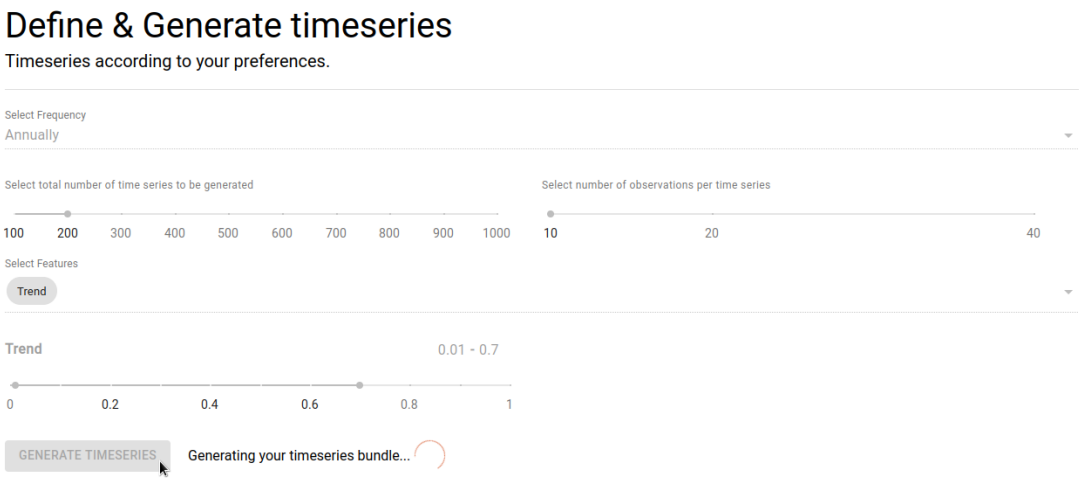
Σχήμα 4.20: *Επιλογές για δημιουργία μηνιαίων χρονοσειρών*

Στη συνέχεια, ο χρήστης πρέπει να επιλέξει τα χαρακτηριστικά που επιθυμεί να πληρούν οι χρονοσειρές που θα δημιουργηθούν. Τα διαθέσιμα χαρακτηριστικά είναι τα ίδια με αυτά που προσφέρονται στη σελίδας αναζήτησης και λήψης χρονοσειρών. Ωστόσο, για αυτήν τη λειτουργία, ο χρήστης πρέπει να ορίσει ένα εύρος τιμής για τουλάχιστον ένα από τα χαρακτηριστικά.



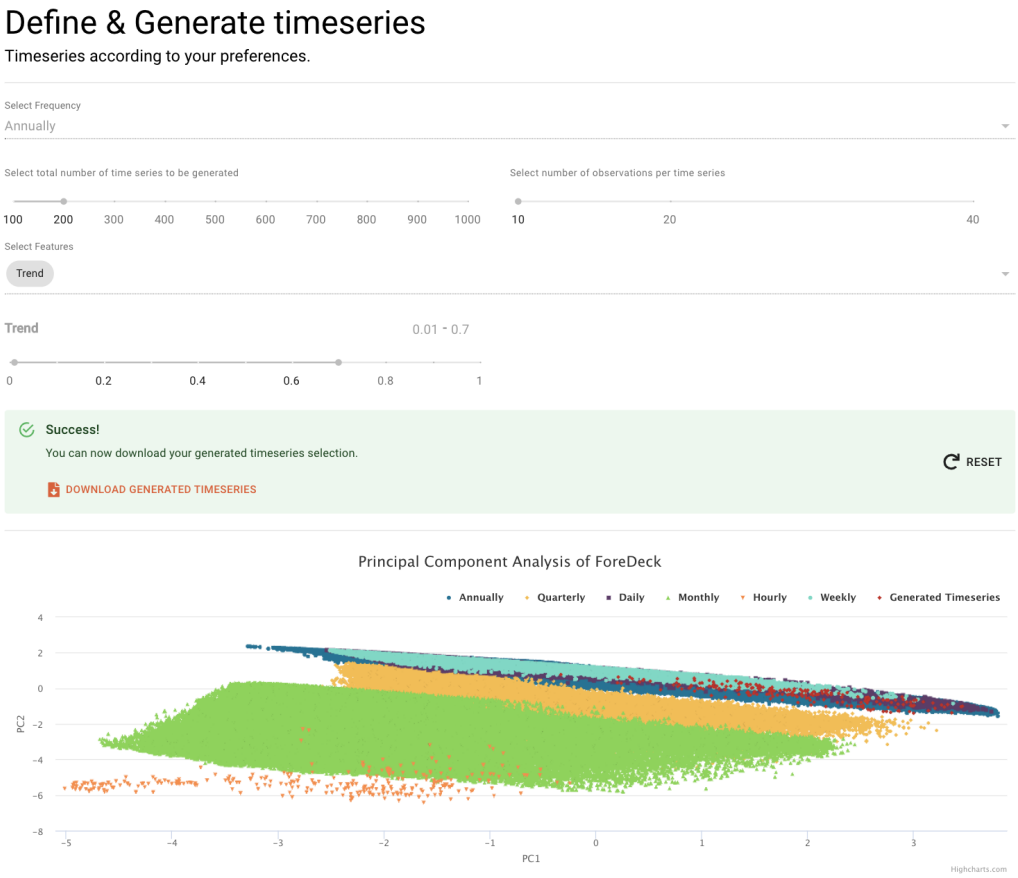
Σχήμα 4.21: *Παράδειγμα σφάλματος για μη επιλογή χαρακτηριστικών*

Αφού ο χρήστης ορίσει τα κριτήρια που τον ενδιαφέρουν, μπορεί να προχωρήσει στη δημιουργία των χρονοσειρών. Με το πάτημα του κουμπιού, η εφαρμογή στέλνει το αντίστοιχο αίτημα στην διεπαφή δημιουργίας χρονοσειρών, παρέχοντας ταυτόχρονα στον χρήστη ξεκάθαρες ενδείξεις ότι η διαδικασία βρίσκεται σε εξέλιξη.



Σχήμα 4.22: *Αίτημα δημιουργίας χρονοσειρών με συγκεκριμένα χαρακτηριστικά*

Με την ολοκλήρωση της διαδικασίας, οι νέες χρονοσειρές εμφανίζονται πάνω στο γράφημα και ο χρήστης μπορεί να τις κατεβάσει. Ένας από τους λόγους που κάποιος θα θελήσει να δημιουργήσει νέες χρονοσειρές είναι επειδή οι υπάρχουσες δεν καλύπτουν τις ανάγκες του, οπότε αυτή η οπτική αναπαράσταση είναι χρήσιμη καθώς βοηθά τον χρήστη να τις αξιολογήσει σε σχέση με το σύνολο των υπαρχουσών και να προσαρμόσει την προσέγγισή του ανάλογα με τις ανάγκες του.



Σχήμα 4.23: *Αποτέλεσμα δημιουργίας χρονοσειρών με συγκεκριμένα χαρακτηριστικά*

Ο χρήστης έχει πρόσβαση σε ένα και μόνο αρχείο που περιέχει τις χρονοσειρές που έχουν δημιουργηθεί μαζί με τα αντίστοιχα χαρακτηριστικά τους.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	entropy	trend	e_acf1	stability	seasonal_strength	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.00
2	0.06	0.16	-0.25	0	0	4.70	2.20	10.71	9.96	9.37	1.00	6.50	11.74	1.38	4.18
3	0.07	0.38	0.27	0	0	19.45	15.01	9.33	2.23	10.55	11.50	21.22	34.47	24.23	13.46
4	0.33	0.09	-0.41	0	0	3.74	6.25	4.42	1.00	5.10	3.20	4.19	3.90	4.40	4.64
5	0.58	0.14	-0.39	0	0	47.80	57.30	77.65	43.92	50.31	1.00	54.15	47.15	150.88	42.44
6	0.31	0.47	-0.34	0	0	683.64	714.14	634.92	650.62	516.90	469.14	236.40	272.28	1.00	651.28
7	0.55	0.52	0.38	0	0	13.76	14.61	15.53	13.98	11.06	7.82	1.00	5.03	9.29	16.76
8	0.69	0.58	-0.04	0	0	19.36	17.58	17.07	25.77	38.50	28.19	25.49	36.52	34.49	30.98
9	0	0.57	-0.03	0	0	22.06	35.53	9.48	1.00	4.61	8.42	3.84	3.67	14.15	32.05
10	0.3	0.05	-0.83	0	0	4.06	4.49	11.82	1.00	10.96	3.91	11.74	3.04	14.62	1.92
11	0.9	0.34	-0.17	0	0	1.00	4.45	36.54	7.07	2.90	13.21	39.99	35.32	13.17	41.78
12	0.1	0.69	-0.07	0	0	13.66	13.35	16.31	0.65	0.94	1.34	2.11	1.85	2.84	5.62
13	0.33	0.29	0.23	0	0	5.77	15.31	17.66	16.83	25.77	26.18	1.00	6.23	26.38	37.82
14	0.3	0.03	-0.87	0	0	11.89	1.32	11.69	2.26	11.36	2.63	12.43	2.94	13.14	1.00
15	0.32	0.35	-0.63	0	0	117.28	238.90	105.56	92.79	220.05	14.06	110.16	77.22	87.96	1.00
16	0.81	0.51	-0.09	0	0	18.51	22.28	33.50	53.86	38.51	66.81	60.18	33.52	41.31	71.15
17	0.65	0.55	-0.25	0	0	0.97	4.04	2.81	5.32	3.60	6.88	7.46	3.58	3.07	2.60
18	0.46	0.08	-0.86	0	0	20.43	8.61	21.09	7.92	20.45	7.41	20.64	5.00	24.48	1.00
19	0.76	0.55	0.31	0	0	22.61	16.15	19.44	25.70	37.47	50.95	44.04	35.18	25.91	36.09
20	0.65	0.4	0.15	0	0	280.04	205.41	160.33	272.30	213.78	226.18	106.47	1.00	357.30	524.74
21	0.49	0.55	-0.3	0	0	31.03	35.59	38.46	39.44	41.18	43.26	5.90	21.31	5.63	19.85
22	0.95	0.51	-0.3	0	0	16.90	22.10	9.95	10.88	20.02	9.18	7.52	5.61	7.33	8.56
23	1	0.43	-0.19	0	0	416.64	8.95	10.24	27.12	28.44	53.56	52.57	92.09	86.90	138.88
24	0.83	0.48	0.14	0	0	2.06	2.49	5.39	5.22	4.55	4.23	2.25	1.00	3.36	1.87
25	0.77	0.28	-0.14	0	0	12.08	6.35	3.12	2.68	8.74	9.04	22.42	26.37	35.29	1.00
26	0.3	0.57	-0.14	0	0	2.49	6.33	5.87	8.67	4.13	5.07	4.17	8.97	8.65	11.50
27	0.79	0.2	-0.39	0	0	39.25	4.11	18.46	26.72	22.80	29.13	28.82	28.10	31.07	34.68
28	0.3	0.7	-0.58	0	0	9.50	7.92	7.75	3.96	3.81	1.00	6.32	1.62	3.94	2.54
29	0.56	0.67	-0.32	0	0	67.83	108.80	164.22	232.03	152.11	203.72	112.34	152.09	38.49	42.55
30	0.64	0.45	0.06	0	0	20.72	20.08	11.56	13.08	15.76	15.53	17.89	18.37	16.83	19.83
31	0.71	0.09	-0.33	0	0	14.19	2.42	4.87	11.47	4.66	4.03	7.81	12.17	1.00	5.02
32	0.8	0.47	0.22	0	0	10.34	12.38	13.28	16.13	22.47	16.11	11.18	6.51	6.14	11.94
33	0.31	0.61	0.02	0	0	16.55	13.55	12.74	12.40	17.23	17.51	16.18	16.82	19.83	19.81
34	0.71	0.58	0.43	0	0	14.67	10.17	7.68	4.55	5.50	8.97	9.20	9.73	3.89	1.00

Σχήμα 4.24: *Πλήρης πληροφορία για χρονοσειρές που δημιουργήθηκαν*

Κεφάλαιο 5

Συμπεράσματα και μελλοντικές προεκτάσεις

Στο κεφάλαιο αυτό γίνεται μία σύντομη ανακεφαλαίωση και αναφέρονται ενδεχόμενες μελλοντικές κατευθύνσεις προς επέκταση και βελτίωση του συστήματος FOREDeCk.

Η παρούσα διπλωματική εργασία αποσκοπεί στην παροχή νέων συνόλων δεδομένων για την δημιουργία ενός ποιοτικού και αξιόπιστου εργαλείου προς υποστήριξη της ερευνητικής κοινότητας. Το σύστημα που έχει υλοποιηθεί παρέχει πάνω από ένα εκατομμύριο πραγματικές χρονοσειρές με ποικίλα χαρακτηριστικά που καλύπτουν κάθε τομέα, από την βιομηχανία και τον τουρισμό μέχρι την υγεία και το περιβάλλον. Ακόμα, για πιο εξειδικευμένες ανάγκες που ίσως να μην καλύπτονται από το υπάρχον περιεχόμενο της δεξαμενής δεδομένων, ο χρήστης μπορεί ανά πάσα στιγμή να δημιουργήσει τεχνητές χρονοσειρές με ελεγχόμενα χαρακτηριστικά της αρεσκείας του.

Η περιορισμένη ποικιλομορφία των δεδομένων, ιδίως σε μεγάλα σύνολα δεδομένων, αναδεικνύεται συχνά ως ανησυχία στη βιβλιογραφία. Γι' αυτόν τον λόγο, από την αρχή δόθηκε ιδιαίτερη προσοχή στη διαδικασία συλλογής δεδομένων καθώς και στην μετέπειτα διαδικασία φιλτραρίσματος, με σκοπό να αποφευχθούν τέτοιες αδυναμίες στο τελικό σύνολο δεδομένων. Επιπλέον, σημαντικό κομμάτι της πλατφόρμας αποτελεί η δυνατότητα οπτικοποίησης των δεδομένων στον δισδιάστατο χώρο που έχει οριστεί στο πλαίσιο της μελέτης των Kang et al., 2017 [9], κάτι που προσφέρει ένα επιπλέον εργαλείο για την ανάδειξη σχέσεων και τη σύγκριση των χρονοσειρών που επιλέγει ο χρήστης.

Όπως έχει αναφερθεί προηγουμένως, ο ευρύτερος στόχος του FOREDeCk είναι η υποστήριξη της ερευνητικής κοινότητας των τεχνικών προβλέψεων με νέα σύνολα δεδομένων και τα κατάλληλα εργαλεία ώστε να μπορεί να εκμεταλλευτεί αυτά τα δεδομένα με τον καλύτερο δυνατό τρόπο. Βάσει αυτού, υπάρχουν πολλές ενδεχόμενες προεκτάσεις και βελτιώσεις που θα μπορούσαν να γίνουν. Αυτές μπορούν να ομαδοποιηθούν σε τρεις βασικές κατηγορίες:

Εμπλουτισμός του συνόλου δεδομένων
Προσθήκη νέων χρονοσειρών, ειδικότερα από συχνότητες και τομείς όπου η κάλυψη είναι περιορισμένη. Αυτό μπορεί να επιτευχθεί με δύο τρόπους: αφενός, μπορούμε να ακούσουμε την ανταπόκριση και τις ανάγκες της κοινότητας που ήδη χρησιμοποιεί το σύνολο δεδομένων, προκειμένου να εντοπίσουμε τις προτεραιότητες για προσθήκη νέων χρονοσειρών. Αφετέρου, μπορούμε να δημιουργήσουμε μηχανισμούς όπου η ίδια η κοινότητα μπορεί να προτείνει

νέες χρονοσειρές που θα ενσωματωθούν στο μέλλον ως μέρος του βασικού συνόλου δεδομένων. Και οι δύο προσεγγίσεις ενισχύουν τη διαφάνεια και την ανοιχτότητα του συστήματος, δημιουργώντας ένα πλούσιο και διαρκώς εξελισσόμενο πόρο για την ερευνητική κοινότητα.

Μια πτυχή που απαιτεί προσοχή σε μια τέτοια προέκταση, είναι ότι η προσθήκη νέων χρονοσειρών ή ο εμπλουτισμός των υπαρχόντων με περισσότερες παρατηρήσεις, επηρεάζει άμεσα το βασικό σύνολο δεδομένων. Αυτό μπορεί να δυσκολέψει τη δημιουργία ενός κοινού σημείου αναφοράς για όσους χρησιμοποιούν τα δεδομένα, περιορίζοντας τη δυνατότητά τους να συγκρίνουν αποτελέσματα και να αντλούν συμπεράσματα από αυτά.

Μια πιθανή προσέγγιση για την αντιμετώπιση αυτού του προβλήματος θα μπορούσε να είναι η αποδοχή των διαφορετικών επεκτάσεων του συνόλου δεδομένων ως εναλλακτικές εκδόσεις, για παράδειγμα μια νέα εκδοχή ανά χρονική περίοδο (FOREDeCk2023, FORE-DeCk2024, κτλ.). Αυτό θα διευκόλυνε την αναγνώριση των διαφορετικών εκδοχών του συνόλου δεδομένων από την κοινότητα και θα δημιουργούσε διαφάνεια ως προς την εξέλιξη των δεδομένων στον χρόνο.

Τεχνικές βελτιώσεις

Η βέλτιστη εμπειρία χρήσης ήταν μια από τις κυριότερες απαιτήσεις του συστήματος, γιάυτο και κατά τη διάρκεια της υλοποίησης αφιερώθηκε σημαντικός χρόνος σε βελτιστοποιήσεις που αποσκοπούσαν στην αύξηση της απόδοσης, της ταχύτητας απόκρισης και των χρόνων φόρτωσης των δεδομένων. Μεταξύ αυτών συμπεριλαμβάνεται η επανασχεδίαση της βάσης δεδομένων, το caching στον server, καθώς και το άμεσο κατέβασμα των δεδομένων οπτικοποίησης στον client, προκειμένου να αποφευχθούν αχρείαστες καθυστερήσεις.

Αυτές οι αλλαγές επικεντρώθηκαν κυρίως στο κομμάτι του διαμοιρασμού και της οπτικοποίησης των δεδομένων, με αποτέλεσμα να παραμένουν ακόμα περιθώρια για βελτιώσεις στο κομμάτι των τεχνητών χρονοσειρών. Όπως αναφέρθηκε και στην αντίστοιχη ενότητα, έχουν τεθεί κάποια όρια στις δυνατότητες δημιουργίας τεχνητών χρονοσειρών, λόγω ακριβώς της καθυστέρησης που παρατηρείται σε συγκεκριμένα σενάρια δημιουργίας, προκειμένου να διασφαλιστεί η ευχρηστία του συστήματος. Αυτά τα όρια ίσως μπορούν να αφαιρεθούν με κάποιες αλλαγές στον τρόπο λειτουρίας της γεννήτριας. Συγκεκριμένα, οι δημιουργοί της γεννήτριας GRATIS, η οποία χρησιμοποιείται στο σύστημα για τη δημιουργία χρονοσειρών, έχουν κάνει κάποιες βελτιώσεις στο αντίστοιχο πακέτο gratis της R, οι οποίες θεωρητικά βελτιώνουν τους χρόνους δημιουργίας. Όμως, λόγω περιορισμένου χρόνου, οι νέες συναρτήσεις του πακέτου δεν πρόλαβαν να δοκιμαστούν και να ενσωματωθούν στο σύστημα. Παρ' όλα αυτά, αντιπροσωπεύουν έναν δυνητικό τρόπο βελτίωσης της δημιουργίας των χρονοσειρών που θα μπορούσε να εξερευνηθεί στο μέλλον.

Προηγμένη Διαδραστικότητα

Το σύστημα θα μπορούσε να παρέχει στους χρήστες περισσότερες δυνατότητες για αλληλεπίδραση με τα δεδομένα που τους ενδιαφέρουν μέσω της διαδικτυακής πλατφόρμας. Για παράδειγμα:

- **Διαδραστικές γραφικές παραστάσεις**: Μια ενδιαφέρουσα προσθήκη θα ήταν η δυνατότητα για τον χρήστη να επιλέγει χρονοσειρές και να τις αναπαριστά σε γραφικές

παραστάσεις, επιτρέποντάς του να πραγματοποιεί διαδραστικές συγκρίσεις. Αυτό θα επέτρεπε στον χρήστη να αντιληφθεί καλύτερα τη μορφή, τις διαφορές και τις ομοιότητες μεταξύ των χρονοσειρών χωρίς την ανάγκη για λήψη και τοπική οπτικοποίηση τους. Επιπλέον, λαμβάνοντας υπόψη ότι ήδη διαθέτουμε επιπλέον πληροφορίες από τις στατιστικές αναλύσεις, μπορούμε να εκμεταλλευτούμε αυτά τα στοιχεία για να παρέχουμε περισσότερη πληροφορία στον χρήστη όταν χρησιμοποιεί αυτήν τη διαδραστική αναπαράσταση.

- **Ενσωμάτωση περισσότερων δυνατοτήτων από το περιβάλλον εκτέλεσης της γλώσσας R**: Αφού υπάρχουν ήδη τα θεμέλια για αμφίδρομη επικοινωνία με το περιβάλλον εκτέλεσης R, είναι εφικτό να εκμεταλλευτούμε περαιτέρω αυτή τη σύνδεση. Μια ενδιαφέρουσα προοπτική είναι η ενσωμάτωση περισσότερων λειτουργιών στην πλατφόρμα κατευθείαν από το περιβάλλον εκτέλεσης της R. Αυτό θα επέκτεινε σημαντικά τις δυνατότητες ανάλυσης των χρονοσειρών που προσφέρουμε. Με αυτήν την ενίσχυση, οι χρήστες θα μπορούν να εκτελούν πιο πολύπλοκες αναλύσεις απευθείας από το περιβάλλον της εφαρμογής, επιτρέποντας τους να λαμβάνουν άμεσα αποτελέσματα και συμπεράσματα χωρίς την ανάγκη για εξωτερικά εργαλεία. Αυτό θα καθιστούσε την πλατφόρμα ακόμα πιο πολύτιμη για ερευνητές και επαγγελματίες που ασχολούνται με την ανάλυση χρονοσειρών, προσφέροντας ένα πιο ολοκληρωμένο και ευέλικτο εργαλείο για την εξερεύνηση και την κατανόηση των δεδομένων.

- **Ενσωμάτωση εξωτερικών δεδομένων**: Αυτή η επέκταση θα επιτρέπει στους χρήστες να ενσωματώνουν τα δικά τους δεδομένα στις υπάρχουσες οπτικοποιήσεις και αναλύσεις που προσφέρει η πλατφόρμα. Συγκεκριμένα, θα μπορούσε να δίνεται η δυνατότητα να ενσωματώσουν τα δικά τους δεδομένα στον υπάρχοντα χώρο ανάλυσης PCA, επιτρέποντάς τους να αντιληφθούν καλύτερα τη θέση και τη σημασία των δικών τους δεδομένων σε σχέση με αυτά που προσφέρει ήδη η πλατφόρμα. Με αυτόν τον τρόπο, οι χρήστες θα μπορούν να εξερευνήσουν τη συσχέτιση και τις διαφορές μεταξύ των δικών τους δεδομένων και των υπαρχόντων, προσφέροντας μια πιο προσαρμοσμένη και εξατομικευμένη εμπειρία ανάλυσης.

- **Λογαριασμοί χρηστών και αναλυτικά στοιχεία χρήσης**: Μια επέκταση του συστήματος θα μπορούσε να περιλαμβάνει τη δυνατότητα δημιουργίας λογαριασμών χρηστών στην πλατφόρμα. Αυτό θα απαιτεί από τους χρήστες να δημιουργήσουν έναν προσωπικό λογαριασμό προτού αξιοποιήσουν πλήρως τις λειτουργίες της πλατφόρμας. Αυτή η προσέγγιση συμβαδίζει με τον τρόπο λειτουργίας των μεγάλων πλατφορμών δεδομένων, όπως το Quandl, όπου η δημιουργία λογαριασμού είναι υποχρεωτική για πρόσβαση σε επιπλέον υπηρεσίες.

  Οι λογαριασμοί χρηστών θα μας επιτρέπουν να διαχειριζόμαστε πιο αποτελεσματικά τα δεδομένα και τις ενέργειες των χρηστών. Θα μπορούμε να παρέχουμε αναλυτικά στοιχεία χρήσης για κάθε λογαριασμό, όπως πόσες χρονοσειρές έχει κατεβάσει, τι αναλύσεις έχει εκτελέσει, και πόσο χρόνο έχει δαπανήσει στην πλατφόρμα. Αυτά τα αναλυτικά στοιχεία χρήσης θα μας βοηθήσουν να κατανοήσουμε πως οι χρήστες αλληλεπιδρούν με την πλατφόρμα και ανοίγει τον δρόμο για περαιτέρω βελτιώσεις,

αφού μπορούν να αναγνωριστούν τα σημεία όπου οι χρήστες χρειάζονται περισσότερη υποστήριξη ή επιπλέον λειτουργίες.

Επιπλέον, μέσω των λογαριασμών χρηστών, ανοίγεται το ενδεχόμενο για μια επιπλέον επέκταση του τρόπου αλληλεπίδρασης των χρηστών με την πλατφόρμα. Μπορούμε να προσφέρουμε πρόσβαση στη δεξαμενή δεδομένων μέσω ενός ξεχωριστού API, με τη χρήση προσωπικών API keys που θα αντιστοιχούν σε κάθε λογαριασμό. Αυτό παρέχει μια επιπλέον στρώση ευελιξίας για τους χρήστες που επιθυμούν να αλληλεπιδρούν με τα δεδομένα μέσω προγραμματιστικών διεπαφών.

Βιβλιογραφία

[1] Hannah Ritchie, Lucas Rodés-Guirao, Edouard Mathieu, Marcel Gerber, Esteban Ortiz-Ospina, Joe Hasell και Max Roser. *Population Growth of India. Our World in Data*, 2023. https://ourworldindata.org/population-growth.

[2] Athanasopoulos G. Hyndman, R.J. *Forecasting: principles and practice, 3rd edition*, 2021. https://OTexts.com/fpp3.

[3] Vipul Mehra. *Forecasting USD to INR foreign exchange rate using Time Series Analysis techniques like HoltWinters Simple Exponential Smoothing, ARIMA and Neural Networks*. 2017.

[4] Ευάγγελος Σπηλιώτης. *Ανάπτυξη πλαισίου αυτοματοποιημένης προέκτασης χρονοσειρών μέσω της γενίκευσης της μεθόδου πρόβλεψης θ*. Διδακτορική Διατριβή, Εθνικό Μετσόβιο Πολυτεχνείο, 2017.

[5] Yanfei Kang, Rob J. Hyndman και Feng Li. *GRATIS: GeneRAting TIme Series with diverse and controllable characteristics. Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(4):354–376, 2020.

[6] Rob J. Hyndman. *A brief history of time series forecasting competitions*. https://robjhyndman.com/hyndsight/forecasting-competitions/.

[7] Keith Ord. *Commentaries on the m3-competition. International Journal of Forecasting*, 17:537–541, 2001.

[8] Michael Clements και David Hendry. *Explaining the results of the M3 forecasting competition. International Journal of Forecasting*, 17, 2001.

[9] Yanfei Kang, Rob J. Hyndman και Kate Smith-Miles. *Visualising forecasting algorithm performance using time series instance spaces. International Journal of Forecasting*, 33(2):345–358, 2017.

[10] Evangelos Spiliotis, Andreas Kouloumos, Vassilis Assimakopoulos και Spyros Makridakis. *Are forecasting competitions data representative of the reality? International Journal of Forecasting*, 2018.

[11] Spyros Makridakis, Vassilis Assimakopoulos και Evangelos Spiliotis. *Objectivity, reproducibility and replicability in forecasting research. International Journal of Forecasting*, 34, 2018.

[12] Spyros Makridakis, Evangelos Spiliotis και Vassilis Assimakopoulos. *The M4 Competition: Results, findings, conclusion and way forward*. *International Journal of Forecasting*, 34, 2018.

[13] Φώτιος Πετρόπουλος και Βασίλειος Ασημακόπουλος. *Επιχειρησιακές Προβλέψεις*. Εκδόσεις Συμμετρία, 2013.

[14] Ben Fulcher. *Feature-based time-series analysis*. 2017.

[15] Rob J. Hyndman, Earo Wang και Nikolay Laptev. *Large-Scale Unusual Time Series Detection*. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, σελίδες 1616–1619, 2015.

[16] Spyros Makridakis και Michèle Hibon. *The M3-Competition: results, conclusions and implications*. *International Journal of Forecasting*, 16(4):451–476, 2000. The M3-Competition.

[17] H. Hotelling. *Analysis of a complex of statistical variables into principal components*. *Journal of Educational Psychology, 24(6), 417-441*, 24(6):417–441, 1933.

[18] Rob J. Hyndman, Yanfei Kang και Kate Smith-Miles. *Exploring time series collections used for forecast evaluation*. 36th International Symposium on Forecasting (ISF 2016), June 19 – 22, 2016, Santander, Spain.

[19] *Nasdaq Data Link (previously Quandl)*. `https://data.nasdaq.com/search`.

[20] City of Melbourne. *Melbourne's Open Data Platform*. `https://data.melbourne.vic.gov.au/`.

[21] New York State. *New York's Open Data Portal*. `https://data.ny.gov/`.

[22] Stanford NLP Group. *Stemming and lemmatization*. `https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html`.

[23] Princeton University. *WordNet*. `https://wordnet.princeton.edu/`.

[24] R Foundation for Statistical Computing. *R: A Language and Environment for Statistical Computing*. `https://www.r-project.org/`.

[25] Josh Katz. *rscript: A simple little module for passing data from NodeJS to R (and back again)*. `https://github.com/joshkatz/r-script`.

[26] *GRATIS: GeneRAting TIme Series with diverse and controllable characteristics*. `https://github.com/ykang/gratis`.

[27] Rob J. Hyndman, Earo Wang και Yanfei Kang. *tsfeatures: Time Series Feature Extraction*, 2018. `https://github.com/robjhyndman/tsfeatures//`.