

Μέτρα εγγύτητας βασισμένα στους γράφους ορατότητας

Εαρινό Εξάμηνο 2023- 2024

ΔΙΑΤΜΗΜΑΤΙΚΟ ΔΙΔΡΥΜΑΤΙΚΟ ΠΜΣ στην Οικονομική Φυσική

Όνομα: Πράπας Ευθύμιος Αλκιβιάδης

Όνομα Επιβλέποντα Καθηγητή: Καθηγητής Καρακασίδης Θεόδωρος,
Χαρακόπουλος Αβραάμ

Αριθμός μητρώου: 00064

Έτος: 2024

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ
ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΦΥΣΙΚΗΣ



ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ
ΤΜΗΜΑ ΦΥΣΙΚΗΣ

Βόλος

Υπεύθυνη Δήλωση

Βεβαιώνω ότι είμαι συγγραφέας αυτής της μεταπτυχιακής διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στη μεταπτυχιακή διπλωματική εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η πτυχιακή εργασία προετοιμάστηκε από μένα προσωπικά ειδικά για τις απαιτήσεις του Διατμηματικού προγράμματος μεταπτυχιακών σπουδών στην 'Οικονομική Φυσική-Χρηματοοικονομικές Προβλέψεις' Τμήματα Οικονομικών Επιστημών-Φυσικής, Πανεπιστημίου Θεσσαλίας και Φυσικής Διεθνούς Πανεπιστημίου της Ελλάδας.

Περίληψη

Περίληψη

Η παρούσα εργασία μελετά κάποια νέα μέτρα εγγύτητας για 2 μετοχές που βασίζονται στη θεωρία των γράφων ορατότητας. Ορίζονται συνολικά 3 οικογένειες μέτρων που μπορούν να χρησιμοποιηθούν για την αξιολόγηση της εγγύτητας 2 μετοχών. Μετά τη θεωρητική θεμελίωση τους και την αξιολόγηση τους στο πλαίσιο των υπολοίπων μέτρων εγγύτητας, προχωρούμε σε έναν εμπειρικό έλεγχο όπου 18 μετοχές του δείκτη NASDAQ κατά την περίοδο 2020-2023 κατηγοριοποιούνται σε 3 κατηγορίες. Ελέγχουμε κατά πόσο μπορούμε να επιτύχουμε τη σωστή συσταδοποίηση με χρήση των νέων προτεινόμενων μέτρων και συγκρίνουμε με βάση την ανάλυση με συσχετίσεις. Ελέγχουμε επίσης κατά πόσο τα μέτρα αυτά σχετίζονται με την αιτιότητα κατά Granger. Τα αποτελέσματα δείχνουν πως παρά την πιο περιορισμένη πληροφορία που περιέχουν οι δείκτες, είναι κοντά στα αποτελέσματα της συσχέτισης και συγκεκριμένα στην περίπτωση του πρώτου έτους της ανάλυσης μας (2/2020-2/2021) έχουμε καλύτερα αποτελέσματα από ότι δίνει η συσχέτιση. Επίσης με σημαντική ακρίβεια μπορούμε να προβλέψουμε αν υπάρχει συσχέτιση κατά Granger μεταξύ 2 μετοχών στο δείγμα μας. Καταλήγουμε στο ότι τα φαινόμενα χρήζουν περαιτέρω διερεύνησης και σε μεγαλύτερη κλίμακα, ενώ η μέθοδος αποτελεί μία μίξη της παραδοσιακής μεθοδολογίας για την ταξινόμηση μετοχών με μέτρα παραγόμενα από γράφους ορατότητας.

Λέξεις-Κλειδιά

Λέξεις-Κλειδιά: Γράφοι Ορατότητας, Συσταδοποίηση Μετοχών, Ομοιότητα χρονοσειρών

Abstract

This thesis is studying measures of similarity for pairs of stocks that are based on visibility graphs. In total, 3 families of measures are defined that can be used to ascertain similarity between 2 stocks. After going through their theoretical underpinnings and their evaluation within the framework of other similarity measures, we proceed to test them empirically on 18 stock indices during the period 2020-2023, that belong to 3 sectors of NASDAQ. We check whether we can achieve the correct clustering using the new measures and we compare to the benchmark of correlation analysis. We also check the association of these new measures with Granger causality. Results show that despite the smaller informational content of these measures, their results are close to correlation analysis and even overperform in the case of the first year of our data (2/2020-2/2021). We can also predict with reasonable precision whether there is a Granger causality connection between 2 stocks in our sample. We conclude that these phenomena need to be studied on a larger scale, while the method suggested is a mix between the classical methodology for stock taxonomy and new measures based on visibility graphs.

Key-Words

Key-Words: Visibility graphs, stock clustering, time-series similarity

Περιεχόμενα

1	Εισαγωγή	6
1.1	Σχέση Οικονομικής, Σύνθετων Δικτύων και Μηχανικής Μάθησης	10
2	Ανασκόπηση Βιβλιογραφίας	13
2.1	Συσχετίσεις μεταξύ τιμών μετοχών στο χρηματιστήριο	14
2.2	Η μέθοδος των γραφημάτων ορατότητας	18
2.2.1	Περιγραφή	18
2.2.2	Σχετικές εφαρμογές	21
2.3	Συσταδοποίηση	23
2.3.1	Μετρικές ομοιότητας	25
2.3.2	Μετρικές αξιολόγησης συσταδοποίησης	28
2.4	Δέντρα αποφάσεων και τυχαία Δάση	30
2.5	Μοντέλα VAR και Granger Αιτιότητα	32
3	Μεθοδολογία	34
3.1	Μεθοδολογικό πλαίσιο	34
3.1.1	Ερευνητικά Ερωτήματα	36
3.1.2	Ερευνητική Στρατηγική	37
3.1.3	Συλλογή δεδομένων και υπολογιστικά εργαλεία	39
3.1.4	Περιορισμοί της έρευνας και ηθικά ζητήματα	41
3.2	Εφαρμογή Μεθοδολογίας	42
3.2.1	Προεπεξεργασία δεδομένων	42
3.2.2	Τοπικά διανύσματα ορατότητας	44
3.2.3	Συσταδοποίηση με Συσχέτιση	47
3.2.4	Δημιουργία πίνακα αποστάσεων γράφων ορατότητας	48
3.2.5	Ομαδοποίηση με χρήση των μέτρων	49
3.2.6	Σύγκριση ομαδοποιήσεων	49

3.2.7	Μετρική ομοιότητας ομαλής ορατότητας	50
3.2.8	Εκτίμηση αιτιότητας	51
3.2.9	Δέντρα αποφάσεων και τυχαίο Δάσος	52
4	Αποτελέσματα	53
4.1	Αποτελέσματα Ταξινόμησης με Ιεραρχική Συσταδοποίηση	53
4.2	Αποτελέσματα Συσταδοποίησης με μεθόδους Τυχαίου Δάσους	55
4.3	Ανάλυση ομαλών χρονοσειρών ορατότητας με βάση αυτοπαλινδρόμηση	57
4.4	Αποτελέσματα αιτιότητας κατά Granger	58
4.5	Πρόβλεψη αιτιότητας κατά Granger	59
5	Συζήτηση	59
6	Μελλοντική Έρευνα	62
7	Αναφορές	65
8	Παραρτήματα	72

1 Εισαγωγή

Ήδη από το 16ο αιώνα εμφανίζονται χρηματιστηριακές αγορές στην Ολλανδία με σημαντικό στοιχείο τους τη μη προβλεψιμότητα. Με την ανάπτυξη των μεθόδων των οικονομικών καθώς και συναφών τεχνολογικών εργαλείων, αναζητούμε τρόπους να προβλέψουμε την πορεία των τιμών των μετοχών. Η δυναμικότητα της αγοράς είναι εντυπωσιακή και ενώ γενικοί κανόνες για τη συμπεριφορά τους έχουν βρεθεί, έχει παρατηρηθεί ότι είναι πρακτικά αδύνατο να επινοηθεί μία μέθοδος που θα κερδίζει με συνέπεια χρήματα. Θεωρητικά, αυτό αποδίδεται στην αρχή του market efficiency (Laopodis, 2013).

Ο κύριος τρόπος που διαθέτουμε για την περιγραφή των τιμών μίας μετοχής είναι η χρονοσειρά. Ένα από τα εργαλεία για τη διερεύνηση των σχέσεων των τιμών μετοχών είναι η συσχέτιση μεταξύ των χρονοσειρών τους. Έχει χρησιμοποιηθεί για την πρόβλεψη και παίζει βασικό ρόλο στο μοντέλο CAPM (Fama, French, 2004). Η συσχέτιση χρησιμοποιείται συχνά ως κριτήριο για την εγγύτητα των μετοχών (Marti, 2021). Πολλές αναλύσεις έχουν γίνει εστιάζοντας στη συσχέτιση μετασχηματισμένων χρονοσειρών μετοχών λόγω της πρακτικής χρησιμότητας της, με απώτερο στόχο τη διαχείριση χαρτοφυλακίου.

Ένα σημαντικό πρόβλημα της διαχείρισης χαρτοφυλακίου αποτελεί η βελτιστοποίηση του (Tola et al., 2007). Σημαντική συνεισφορά στη θεωρία της διαχείρισης αποτελεί η εργασία του Markowitz (1952) στην οποία αναλύεται η διαδικασία εύρεσης ενός χαρτοφυλακίου υπό εξιδανικευμένες συνθήκες (Tola et al., 2007). Το πρόβλημα ανάγεται στην επιλογή κατάλληλων βαρών για τις μετοχές, όπου τα βάρη περιγράφουν τη σύνθεση του χαρτοφυλακίου (Wu et al., 2022). Σημαντικό ρόλο στην ανάλυση αυτή έχει η έννοια της διαφοροποίησης των μετοχών για τη μείωση του ατομικού ρίσκου (Wu et al., 2022). Για τον έλεγχο της διαφοροποίησης των μετοχών χρησιμοποιείται η συσχέτιση, με ένα βέλτιστο χαρτοφυλάκιο να ελαχιστοποιεί το ρίσκο για δοθείσα απόδοση (Fama, French, 2004). Αυτό μπορεί να επιτευχθεί με τη χρήση

αρνητικά συσχετισμένων μετοχών. Ταυτόχρονα, η ομαδοποίηση μετοχών με βάση τη συσχέτιση, μπορεί να βοηθήσει στη διαφοροποίηση μετοχών ώστε να μειωθεί το ρίσκο (Nanda et al., 2010).

Ωστόσο, η συσχέτιση χαρακτηρίζεται ως ελλιπές μέτρο λόγω βασικών αδυναμιών της. Πρώτα από όλα, είναι ευαίσθητη στις χρονικές μετατοπίσεις. Δύο χρονοσειρές που έχουν την ίδια συμπεριφορά αλλά μετατοπισμένη στο χρόνο δεν θα αναγνωριστούν απαραίτητα ως όμοιες. Η συμπεριφορά αυτή οφείλεται στο ότι η συσχέτιση αποτελεί μέτρο lockstep (Wang et al., 2012) και λαμβάνει υπόψη μόνο τη συμπεριφορά των συγκρινόμενων χρονοσειρών την ίδια χρονική στιγμή. Ταυτόχρονα, η συσχέτιση δίνει μεγάλο βάρος σε ακραίες τιμές και είναι ασταθής μέθοδος (Marti et al., 2021). Παράλληλα, προκύπτουν προβλήματα σύγκλισης από τη στιγμή που παρατηρούμε μόνο ένα μέρος της χρονοσειράς. Η στατιστική αβεβαιότητα που ενέχουν τέτοιες μετρικές δημιουργούν προβλήματα στην αξιοποίηση των αποτελεσμάτων μετέπειτα (Tola et al., 2007).

Συνεπώς δημιουργείται η ανάγκη για νέους τρόπους αξιολόγησης της εγγύτητας μετοχών. Με το ζήτημα της σύγκρισης γενικότερα δύο χρονοσειρών ασχολείται ο τομέας της ομοιότητας χρονοσειρών. Με τη χρήση του τομέα αυτού ορίζονται μέτρα αποστάσεων μεταξύ χρονοσειρών, βασισμένα είτε σε συσχετίσεις είτε σε άλλες ιδιότητες των δεδομένων και έτσι αποκτάται μία καθολική εικόνα για την 'εγγύτητα' των μετοχών.

Ένας καινούριος τρόπος οπτικοποίησης χρονοσειρών προήλθε από τη δουλειά του Lacasa (2008) και των συνεργατών του. Θα αναφερόμαστε σε αυτή την οπτικοποίηση ως γραφήματα ορατότητας. Η βασική ιδέα είναι πως αντί να παρατηρούμε τις τιμές μίας χρονοσειράς με βάση τη χρονική τους στιγμή, τις παρατηρούμε σε σχέση με τις τιμές σε άλλες χρονικές στιγμές.

Πέρα από τη χρησιμότητα των γραφημάτων ορατότητας ως προς την οπτικοποίηση, ενδιαφερόμαστε για την εύρεση πληροφορίας μέσα από αυτά. Κάθε γράφος ορατότη-

τας διαθέτει ήδη κάποια πληροφορία χαρακτηριστική για τη χρονοσειρά (μπορούμε να διαχωρίσουμε τυχαίες από μη τυχαίες χρονοσειρές) αλλά επιδιώκουμε την ανάλυση της πληροφορίας που μπορεί να εξαχθεί από το συνδυασμό 2 γραφημάτων ορατότητας. Μέχρι τώρα, η ανάλυση γίνεται συνήθως ατομικά σε κάθε γράφημα και τα συμπεράσματα/μετρικές συγκρίνονται διαχρονικά (Yu, 2013) ή με άλλα προϊόντα. Σκοπός αυτής της έρευνας είναι η εισαγωγή εργαλείων που λαμβάνουν υπόψη ταυτόχρονα και συνδυαστικά την πληροφορία των γραφημάτων ορατότητας 2 χρονοσειρών.

Η ανάλυση αυτή γίνεται σε βήματα. Το πρώτο βήμα είναι η εξαγωγή κατάλληλων μέτρων. Για το σκοπό αυτό χρησιμοποιούμε μία τοπικά περιορισμένη μορφή του πίνακα ορατότητας. Χρησιμοποιούμε ένα μέτρο της τοπικής κυρτότητας του συνόλου, υπολογίζοντας το ποσοστό των συνδέσεων που υλοποιούνται σε ένα περιορισμένο παράθυρο της χρονοσειράς μας. Λαμβάνουμε έτσι ένα διάνυμα που αντιπροσωπεύει την τοπική συμπεριφορά του γραφήματος. Θα υλοποιήσουμε επίσης μέτρα απόστασης μεταξύ των πινάκων ορατότητας, όπως η απόσταση Jaccard. Με τη χρήση αυτών των μέτρων θα έχουμε τρόπους να καταγράψουμε την ομοιότητα 2 γραφημάτων ομοιότητας δίχως να καταφύγουμε στη χρήση της θεωρίας γράφων.

Με τη χρήση των εξαχθέντων μέτρων, μπορούμε πλέον να υλοποιήσουμε τους 2 κύριους στόχους της εργασίας οι οποίοι είναι η διερεύνηση της σχέσης των μέτρων με την προβλεψιμότητα των χρονοσειρών καθώς και η ομαδοποίηση μετοχών με βάση αυτά τα μέτρα. Η ανάλυση μας θα συνεχίσει με βάση την αιτιότητα κατά Granger. Η έννοια αυτή βασίζεται στη διερεύνηση της παλινδρόμησης τιμών μία χρονοσειράς με τις τιμές μίας άλλης χρονοσειρά, με την κατάλληλη υστέρηση. Η χρήση του εργαλείου της παλινδρόμησης οδηγεί σε ένα στατιστικό έλεγχο σημαντικότητας που δηλώνει αν υπάρχει αιτιότητα. Παρόλο που η ονομασία αιτιότητα είναι παραπλανητική, μιας και δεν αποδεικνύεται κάποια αιτιότητα με την καθημερινή έννοια με τη χρήση αυτής της μεθόδου (Shojaie, Fox, 2022), μπορεί να διαπιστωθεί με αυτή τη μέθοδο αν μία χρονοσειρά εμπεριέχει σημαντική πληροφορία για την πρόβλεψη των τιμών άλλης

χρονοσειράς. Μία ανάλυση των μετρικών με βάση τη στατιστική σημαντικότητα της αιτιότητας κατά Granger οδηγεί σε συμπεράσματα για τη σχέση των δύο μεγεθών.

Παράλληλα, μπορούν να χρησιμοποιηθούν αυτά τα μέτρα για την ομαδοποίηση μετοχών, ορίζοντας μία ομοιότητα κατά ορατότητα. Στην πραγματικότητα των εμπειρικών δεδομένων, η έννοια του καλού μέτρου καταμερισμού είναι αυθαίρετη (Murphy, 2022) και συνεπώς για το δείγμα μας θα συγκρίνουμε τον καταμερισμό με καταμερισμούς που χρησιμοποιούν διαφορετικά μέτρα απόστασης. Παράλληλα, θα χρησιμοποιήσουμε και εσωτερικές μετρικές αξιολόγησης, δηλαδή μετρικές που αξιολογούν την ομαδοποίηση των δεδομένων χωρίς αναφοράς σε εξωτερικά δεδομένα (Murphy, 2022).

Τέλος, μερικά λόγια για τα δεδομένα που χρησιμοποιήθηκαν. Στη σύγχρονη εποχή, η πρόσβαση σε χρηματιστηριακά δεδομένα είναι σχετικά απλή. Χρησιμοποιήθηκαν δεδομένα από την ιστοσελίδα Yahoo Finance για 18 εταιρείες του δείκτη NASDAQ. Οι εταιρείες που επιλέχθηκαν ήταν από 3 κατηγορίες (Nasdaq, 2024), την υγεία, την τεχνολογία καθώς και βασικά είδη κατανάλωσης. Επιλέχθηκαν 6 μετοχές για κάθε κατηγορία. Στην κατηγορία της τεχνολογίας περιλήφθηκαν οι μετοχές FAANG δηλαδή οι META, AMAZON, APPLE, NETFLIX, ALPHABET (GOOGLE) καθώς και η NVIDIA, ενώ στις άλλες κατηγορίες επιλέχθηκαν οι 6 μετοχές που ήταν διαθέσιμες στο δείκτη NASDAQ-100. Ο περιορισμός σε λίγες μετοχές ήταν απαραίτητος λόγω των περιορισμών μίας μεταπτυχιακής εργασίας καθώς και των πολλαπλών τρόπων ανάλυσης των δεδομένων. Τα στοιχεία για όλες τις εταιρείες αφορούσαν κοινό χρονικό διάστημα 5 ετών, αλλά χρησιμοποιήθηκαν οι τιμές από 31-01-2020 έως 01-09-2023. Τα 2 κύρια διαστήματα ανάλυσης μας είναι από 31-01-2020 έως 01-02-2021 και από 01-09-2022 έως 01-09-2023.

1.1 Σχέση Οικονοφυσικής, Σύνθετων Δικτύων και Μηχανικής Μάθησης

Οι απαρχές της σύγχρονης Οικονοφυσικής ως διεπιστημονικό εγχείρημα μπορούν να τοποθετηθούν στο 1987 με το συνέδριο ‘Τα οικονομικά ως ένα πολύπλοκο εξελισσόμενο σύστημα’ (Bouchaud, 2019). Είναι σημαντικό να σημειωθεί πως μέθοδοι της φυσικής είχαν συνδυαστεί πρότερα σε προβλήματα της φυσικής, όπως στην περίπτωση του μοντέλου βαρύτητας για το εμπόριο (Simeone, 2023), η εφαρμογή μοντέλων της φυσικής όπως ο τυχαίος περίπατος στην περίπτωση της κίνησης των μετοχών από τον Bachelier (Simeone, 2023) καθώς και οι χρήση νόμων δύναμης σε εμπειρικές κατανομές οικονομικών μεταβλητών (Yee, 2021). Οι μεμονωμένες αυτές εφαρμογές όμως δεν έδωσαν μία προσέγγιση που θα είχε σημαντική επίδραση τουλάχιστον σαν υποπεδίο είτε της οικονομικής είτε της φυσικής (Yee, 2021) μέχρι τα μέσα της δεκαετίας του 1990, (Kutner et al., 2019).

Η σύγχρονη οικονομοφυσική, που συμφωνείται κατά κανόνα πως άρχισε στη δεκαετία του 1990 (Yee, 2021), (Bouchaud, 2019) έχει επεκτείνει τις μεθόδους και τη θεματολογία της. Η οικονομοφυσική μπορεί να θεωρηθεί ότι εισήγαγε τη χρήση της θεωρίας Χάους (και γενικότερα μεθόδους Πολύπλοκων Συστημάτων), την Κβαντική Μηχανική καθώς και τη Στατιστική Μηχανική στην ανάλυση της οικονομικής (Simeone, 2023). Θεματικά, ασχολείται με νόμους διατήρησης, τους νόμους κλίμακας, νόμους τυχαίου περιπάτου εφαρμοσμένων στις αγορές, τη μελέτη φαινομένων μεταπτώσεων φάσης, τη μελέτη κοινωνικών φαινομένων με μεθόδους οικονομοφυσικής καθώς και τη μελέτη σύνθετων δικτύων και την ανάλυση της δυναμικής τους εξέλιξης (Kutner et al., 2019). Μεθοδολογικά, τα ερευνητικά αποτελέσματα μπορούν να καταταχθούν σε 2 διακριτές οικογένειες με βάση την επισκόπηση του τομέα της Οικονοφυσικής από τους Chakraborti κ.ά. (2011). Ο πρώτος τομέας αφορά τη χρήση μαθηματικών νόμων για την ανάλυση χαρακτηριστικών οικονομικών μεταβλητών του χρηματιστηρίου όπως η τιμή μετοχών, η διακύμανση τους, η συσχέτιση μεταξύ τους

καθώς και η ανάλυση των αποτελεσμάτων αυτών με τη χρήση της θεωρίας τυχαίων πινάκων και τη θεωρία γράφων. Σε δεύτερο επίπεδο, η οικονομοφυσική μελετά μοντέλα με βάση πράκτορες (Agent-based Systems). Μεθοδολογικά, τα ερευνητικά αποτελέσματα της Οικονομοφυσικής μπορούν να διαχωρισθούν σε 3 είδη συνεισφορών: εμπειρικές μελέτες για την περιγραφή κατανομών, τεχνικές συνεισφορές καθώς και η ανάπτυξη μοντέλων (Bouchaud, 2019).

Η δική μας εργασία εμπίπτει στις κατηγορίες της ανάλυσης τιμών χρηματιστηριακών μετοχών με τη χρήση μεθόδων σύνθετων δικτύων και η συνεισφορά είναι τεχνική. Ο ρόλος των Σύνθετων συστημάτων και συγκεκριμένα δικτύων στην Οικονομοφυσική είναι κεντρικός και αρχίζει από τις πρώτες δημοσιεύσεις του τομέα, με την ανάλυση του (Mantegna, 1999). Τα Σύνθετα δίκτυα αποτελούν σημαντικό τομέα της Οικονομοφυσικής και έχει κανονικοποιηθεί η χρήση τους για την ανάλυση χρηματιστηριακών μεταβλητών (Raddant, Matteo, 2023), ειδικά για την ανάλυση των συσχετίσεων μεταξύ των αποδόσεων προϊόντων και την αντιμετώπιση τους ως δυναμικά δίκτυα (Raddant, Matteo, 2023). Ένας περιγραφικός ορισμός των σύνθετων συστημάτων είναι δύσκολο να συμπεριλάβει όλα τα είδη των σύνθετων συστημάτων, αλλά μπορούμε να έχουμε ένα λειτουργικό ορισμό, όπου τα σύνθετα συστήματα συμπεριλαμβάνουν την αλληλεπίδραση πολλών πρακτόρων ή παραγόντων, τη μη-στασιμότητα (τα χαρακτηριστικά του συστήματος αλλάζουν με το χρόνο) και το ότι είναι ανοιχτά συστήματα, δηλαδή μπορούν να δεχθούν επιδράσεις από το περιβάλλον τους (Simeone, 2023). Οι χρηματιστηριακές αγορές ικανοποιούν σχεδόν τετριμμένα αυτά τα κριτήρια. Ένα άλλο στοιχείο που προσδιορίζει τα σύνθετα συστήματα είναι η δυνατότητα μικρών μεταβολών να επιφέρουν σημαντικές διαφορές με κάποια μη μηδενική πιθανότητα (Bouchaud, 2019). Η ύπαρξη οικονομικών κρίσεων δείχνει ότι οι αγορές ικανοποιούν και αυτό το κριτήριο. Κάθε σύνθετο σύστημα μπορεί να παραστεθεί ως ένα σύνθετο δίκτυο με αλληλεπιδράσεις μεταξύ των αλληλεπιδρώντων τμημάτων (Simeone, 2023).

Η τεχνητή νοημοσύνη αποτελεί ένα ραγδαία αναπτυσσόμενο τομέα με σημαντικές και πολυσχιδείς εφαρμογές. Μία βιβλιογραφική ανασκόπηση δείχνει ότι η αναφορά σε τεχνητή νοημοσύνη σε δημοσιεύσεις οικονομικών δείχνει αυξητική τάση (Ruiz-Real et al., 2020), (Hoang, Wiegatz, 2023). Η μηχανική μάθηση αποτελεί υποτομέα της τεχνητής νοημοσύνης που ασχολείται με την ανάπτυξη αλγορίθμων μάθησης (Διαμαντάρας, Μπότσης, 2019) και διαθέτει εφαρμογές σε μικροοικονομικό αλλά και μακροοικονομικό επίπεδο. Οι παραπάνω λόγοι από μόνοι τους δε δικαιολογούν την καταλληλότητα των εργαλείων της μηχανικής μάθησης στο παρόν πρόβλημα. Θα δείξουμε ότι η χρήση αυτών των μεθόδων συνάδει με την παράδοση και τους σκοπούς της οικονομοφυσικής στο πλαίσιο αυτής της εργασίας.

Πρώτα από όλα, η μηχανική μάθηση χρησιμοποιείται για τη δημιουργία προβλέψεων. Η οικονομοφυσική σαν κλάδος αποτελεί τομέα που δίνει ιδιαίτερη έμφαση στις προβλέψεις (Yee, 2021), σε αντίθεση με την οικονομετρία που δίνει έμφαση στην αιτιακή σχέση μεταξύ παραγόντων (Hoang, Wiergatz, 2020). Μία άλλη οπτική που δικαιολογεί αυτή τη θεώρηση δίνεται από τους Kelly και Xiu (2023) που χαρακτηρίζουν τη διαδικασία καθορισμού τιμών ως ένα είδος πρόβλεψης και συνεπώς κατάλληλο τομέα για τη χρήση μεθόδων μηχανικής μάθησης. Παράλληλα, όσο η μηχανική μάθηση περιορίζεται στην εκτίμηση στατιστικών μοντέλων, η διαφοροποίηση μεταξύ των παραδοσιακής οικονομετρίας και μηχανικής μάθησης είναι δυσδιάκριτη (Kelly, Xiu, 2023). Η οικονομοφυσική επίσης χαρακτηρίζεται από αυτάρκεια μεθόδων (Yee, 2021), με τη χρήση πολλών μεθόδων εκτός της κλασικής οικονομικής επιστήμης και την εισαγωγή μεθόδων της φυσικής με δραστικό τρόπο χωρίς να υπάρχει απαραίτητα διάλογος με την κλασική οικονομική σκέψη (Yee, 2021). Συνεπώς, ο κλάδος της οικονομοφυσικής αποτελεί σημαντικό υποψήφιο για την εισαγωγή μεθόδων από άλλους κλάδους αντίστοιχα με τη μεταφορά μοντέλων της φυσικής (Yee, 2021) και θα μπορούσε να αποτελέσει δίοδο για τη χρήση μηχανικής μάθησης στα παραδοσιακά οικονομικά. Τέλος, η χρήση της μηχανικής μάθησης μπορεί να δώσει έμφαση στην υ-

λοποίηση νέων μεγεθών για την επεξήγηση φαινομένων (Hoang, Wiergatz, 2020) και η συγκεκριμένη εργασία στοχεύει στην εισαγωγή και αξιολόγηση νέων προτεινόμενων μέτρων. Στη συγκεκριμένη περίπτωση, έχουμε πως η υλοποίηση νέων αλγορίθμων μάθησης αποτελεί σημαντικό χαρακτηριστικό του τομέα. Η χρήση αλγορίθμων όπως ο PMFG είναι απαιτητική σε υπολογιστικούς πόρους και οδηγεί σε μία φυσική συσταδοποίηση των υποκείμενων δεδομένων (Kutner et al., 2019). Η συσταδοποίηση μετοχών αποτέλεσε σημαντικό μέρος της ανάλυσης δικτύων (Mantegna, 1998), (Kutner et al., 2019) πριν χρησιμοποιηθούν τεχνικές μηχανικής μάθησης και έχει το αντίστοιχο στο φυσικό καταμερισμό σε τομείς των χρηματιστηριακών μετοχών.

Θα πρέπει να σημειωθεί πως παρά τη χρήση μεθόδων μηχανικής μάθησης, η σύνδεση της εργασίας με την Οικονομική παραμένει ουσιαστική για την επεξήγηση των φαινομένων και αποτελεσμάτων της εργασίας. Το πλαίσιο παραμένει ο κλασικός τομέας της ανάλυσης μετοχών ως σύνθετα δίκτυα, με έμφαση στη μεταβαλλόμενη φύση των χαρακτηριστικών του δικτύου. Η προσπάθεια επεξήγησης και πρόβλεψης ιδιοτήτων της χρηματιστηριακής αγοράς αποτελεί μέρος του ορισμού της Οικονομικής που δίνεται από το Simeone (2023). Παράλληλα, δίνεται έμφαση στη μελέτη παραδοσιακών μεγεθών της οικονομικής ανάλυσης όπως η συσχέτιση και η αιτιότητα κατά Granger.

2 Ανασκόπηση Βιβλιογραφίας

Ο ρόλος της βιβλιογραφικής ανασκόπησης είναι καίριος σε μία επιστημονική εργασία. Επιτρέπει την εύρεση των ερωτημάτων που ταλανίζουν τους ερευνητές καθώς και την αποτίμηση της ήδη υπάρχουσας γνώσης στο πεδίο ανάλυσης (Latham, 2022). Συνεισφέρει έτσι στο διαχωρισμό της συμβολής της εργασίας σε σύγκριση με την υπάρχουσα γνώση. Στη βιβλιογραφική ανασκόπηση είναι σημαντικό να περιληφθούν αναφορές σε κλασικά έργα αλλά και να δοθεί έμφαση στις πρόσφατες εξελίξεις στο

μελετούμενο πεδίο, ενώ γίνεται κριτική αποτίμηση της συνεισφοράς του κάθε έργου (Latham, 2022).

Η συγκεκριμένη εργασία περιλαμβάνει και το κομμάτι της υλοποίησης των αλγορίθμων και της αποτίμησης της χρήσης τους στην πράξη. Μπορεί να ενταχθεί εν μέρει υπό την αιγίδα της μεθοδολογίας έρευνας σχεδιασμού. Στη μεθοδολογία της έρευνας σχεδιασμού, η διασύνδεση με τη βάση γνώσης, που στην περίπτωση μας είναι η γνώση για εργαλεία και μεθοδολογίες ώστε να εκτιμηθεί η ομοιότητα μεταξύ χρονοσειρών, αποτελεί στοιχείο που εξασφαλίζει την επιστημονικότητα και την πρωτοτυπία των αποτελεσμάτων (Johannesson, Perjons, 2014). Στη συγκεκριμένη περίπτωση, η πυρηνική θεωρία που μάς καθοδηγεί στο σχεδιασμό των εργαλείων είναι η οικονομική με την προσέγγιση των γράφων ορατότητας, η οικονομετρία χάρη στη χρήση των χρονοσειρών και του υποδείγματος VAR, με περιφερειακές συνεισφορές και από άλλες επιστήμες, όπως το πεδίο των scientometrics για χρήσεις μετρικών ομοιότητας, η μηχανική μάθηση με σημαντικά εργαλεία ανάλυσης και η στατιστική.

2.1 Συσχετίσεις μεταξύ τιμών μετοχών στο χρηματιστήριο

Το βασικό μοντέλο για τη σύγκριση των αποτελεσμάτων μας είναι η ανάλυση των Mantegna, Stanley (1999). Αναλύουμε εδώ κάποιες από τις συνεισφορές στη βιβλιογραφία όσον αφορά τη βασική μεθοδολογία και μεθοδολογικές διαφοροποιήσεις.

Η βασική μεθοδολογία έγκειται στην επιλογή ενός αριθμού από μετοχές/προϊόντα, στον υπολογισμό των λογαριθμικών αποδόσεων και στον υπολογισμό της συσχέτισης κατά Pearson (Marti et al., 2021). Δημιουργείται ένα δίκτυο με κόμβους τις μετοχές και ως αποστάσεις μεταξύ των κόμβων (τιμές των ακμών) ένας μετασχηματισμός των συσχετίσεων (Onnela et al., 2003), ο οποίος απαιτείται να είναι ultrametric έτσι ώστε να κατοπτρίζεται η ιεραρχική φύση της ταξινόμησης των μετοχών (Mantegna, Stanley, 1999). Τέλος, εφαρμόζεται ένας αλγόριθμος, όπως ο Minimum Spanning Tree για τον υπολογισμό ενός δικτύου σε μορφή δέντρου, με αριθμό ακμών 1 λιγότερο

από τις μετοχές (Marti et al., 2021) το οποίο περιέχει πληροφορία για την ταξινόμηση των μετοχών. Το τελευταίο αυτό βήμα περιγράφεται γενικευμένα ως φιλτράρισμα του γράφου και υπάρχουν πολλές μεθοδολογίες για την υλοποίηση του (Saha et al., 2022).

Έχουμε ότι η χρήση της συσχέτισης κατά Pearson προϋποθέτει γραμμικότητα και συχνά οι μετασχηματισμοί που υιοθετούνται έχουν ως χαρακτηριστικό τη στατιστική σημαντικότητα των αποτελεσμάτων (Mantegna, 1998), (Onnela et al., 2003), (So et al., 2021). Η ισχύς αυτών των μέτρων εξαρτάται από τη συμμόρφωση των κατανομών με τις υποθέσεις της κανονικής κατανομής (Marti et al., 2021) και της ανυπαρξίας τάσης, γεγονός που μπορεί να οδηγήσει σε λανθασμένα αποτελέσματα (Hartman, Hlinka, 2018). Ανάλυση σε 3 χρηματιστηριακούς δείκτες καταδεικνύει ότι το κύριο πρόβλημα αποτελεί η μη κανονικότητα των κατανομών ενώ η ενυπάρχουσα τάση στις χρονοσειρές κάποιων μετοχών μπορεί να παρουσιάσει προβλήματα (Hartman, Hlinka, 2018). Άλλα προβλήματα που μπορούν να εμφανιστούν είναι η μεγάλη στατιστική αβεβαιότητα των εκτιμήσεων συσχέτισης καθώς και η ευαισθησία της συσχέτισης σε ακραίες τιμές (Marti et al., 2021). Επίσης, το γεγονός ότι οι χρονοσειρές που επεξεργαζόμαστε είναι πεπερασμένες αλλά συχνά αναλύονται σαν απεριόριστου μήκους οδηγούν σε περαιτέρω προβλήματα στους υπολογισμούς (Macmahon, Garlaschelli, 2015).

Οι Onnela, Chakraborti, Kaski το 2023 χτίζοντας πάνω στα αποτελέσματα του Mantegna εφαρμόζουν τη μεθοδολογία στο χρηματιστήριο της Νέας Υόρκης για μία 20ετία που περιλαμβάνει τη Μαύρη Παρασκευή σε 87 μετοχές. Επιλέγουν ένα παράθυρο 1000 ημερών για αυτό το διάστημα και οι συσχετίσεις υπολογίζονται σε αυτό το διάστημα με διαφορά 21 ημέρες (195 υπολογισμοί). Σκοπός της εργασίας είναι η μελέτη των χρονικά μεταβαλλόμενων στοιχείων του δένδρου που προκύπτει από την εφαρμογή της μεθοδολογίας. Τα αποτελέσματα δείχνουν μεγαλύτερες συσχετίσεις κατά την κρίση και ανώμαλη συμπεριφορά. Παράλληλα, μελετάται κατά πόσο η συ-

σταδοποίηση αντιστοιχεί σε κατηγοριοποίηση από έναν τρίτο οργανισμό. Αντίστοιχα με την ανάλυση του Mantegna (1999) έχουμε πως οι συστάδες αντιστοιχούν στις κατηγοριοποιήσεις του Forbes. Αποκλίσεις μπορούν να αποδοθούν μεταξύ άλλων στις επενδύσεις της κάθε εταιρείας (ειδικά χρηματοπιστωτικά ιδρύματα), στις διαφορετικές κατηγοριοποιήσεις και σε ανεπικαιροποίητα στοιχεία (Onnela et al., 2003). Πρέπει να σημειωθεί όμως πως ο διαχωρισμός του Forbes λαμβάνει υπόψη δυναμικά στοιχεία των μετοχών και έτσι εν μέρει πρόκειται περί αυτοεκπληρούμενης προφητείας.

Οι Tumminello et al., 2005 συνεισφέρουν στη βιβλιογραφία εισάγωντας ένα νέο είδος φιλτραρίσματος του δικτύου που διατηρεί περισσότερη πληροφορία σε σχέση με τα MST, το Planar Maximally Filtered Graph, PMFG. Τα PMFG περιέχουν το MST του κάθε δικτύου. Η προσέγγιση των ερευνητών όσον αφορά το μέτρο ομοιότητας είναι ίδια, με τη χρήση της συσχέτισης και την υλοποίηση του σε 100 μετοχές. Υπάρχουν και άλλες προσεγγίσεις για την υλοποίηση αυτού του βήματος φιλτραρίσματος, όπως είναι τα TMFG που υπερτερούν έναντι των PMFG στην υπολογιστική πολυπλοκότητα (Saha et al., 2022), γράφους προϊόντων (asset graphs) με την αγνόηση συσχετίσεων κάτω από ένα όριο καθώς και μεθόδους τυχαίων πινάκων (Macmahon, Garlaschelli, 2015).

Η υλοποίηση της 'κλασικής' μεθόδου φαίνεται στην εργασία των Birch, Pantelous, Soramaki, 2015 που μελετούν 30 μετοχές του DAX 30 χρησιμοποιώντας MST για φιλτράρισμα. Οι ερευνητές επέλεξαν 23 ημέρες ως το διάστημα παράθυρο για τη συσχέτιση. Μία ακόμη επιλογή τους είναι η χρήση της διόρθωσης Bonferroni για τον έλεγχο πολλαπλών στατιστικών υποθέσεων. Οι ερευνητές καταλήγουν σε μία συσταδοποίηση των μετοχών που είναι φυσική, αλλά δίνουν κυρίως έμφαση στη μείωση των συσχετίσεων που παρατηρήθηκε κατά την περίοδο 2008-2009 της μεγάλης οικονομικής κρίσης.

Η κύρια συνεισφορά της παρούσας εργασίας έγκειται στη χρήση νέου μέτρου ομοιότητας των χρονοσειρών των μετοχών. Συγκεκριμένα για το πρόβλημα της ο-

μαδοποίησης, πέρα από τα λοιπά μειονεκτήματα της μεθόδου έχουμε πως η χρήση των συσχετίσεων δεν μπορεί να διαχωρίσει πλήρως τις ομάδες μετοχών ακόμη και αν θεωρητικά είχαμε ομάδες μετοχών με μηδενική συσχέτιση και μοναδιαία συσχέτιση (Macmahon, Garlaschelli, 2015). Έχουν προταθεί διάφορες μετρικές που έχουν υλοποιηθεί σε άρθρα, ενδεικτικά η χρήση μερικής συσχέτισης (Wang et al., 2018) σε 57 χρηματιστηριακές αγορές, με αποτελέσματα σημαντικά διαφορετικά από τη χρήση απλής συσχέτισης, η χρήση τεχνικών αμοιβαίας πληροφορίας (Guo et al., 2018), (Hartman, Hlinka, 2018) που μπορούν να συλλάβουν μη γραμμικότητες καθώς και η χρήση δικτύων βασισμένα στην αιτιότητα κατά Granger στα οποία θα γίνει ιδιαίτερη μνεία.

Οι Basu et al. (2015) υλοποιούν ένα δίκτυο αιτιότητας το οποίο λαμβάνει υπόψη την κατανομή σε ομάδες των κόμβων μεταξύ όμοιων κόμβων. Εφαρμόζουν τον έλεγχο σε πολλαπλές επαναλήψεις και καταλήγουν στη δημιουργία ενός πολυδιάστατου δικτύου που εξαρτάται κρίσιμα όμως από την ομοιότητα των αρχικών παρατηρήσεων. Οι (Papana et al., 2017) υλοποιούν ένα δίκτυο κατά Granger και χρησιμοποιούν μερική αμοιβαία πληροφορία, αναγνωρίζοντας ότι στην περίπτωση μόνο 2 μεταβλητών η μετρούμενη αιτιότητα μεταξύ τους μπορεί να οφείλεται σε τρίτες μεταβλητές. Οι ερευνητές καταλήγουν ότι ναι μεν τα κριτήρια αιτιότητας κατά Granger είναι ευαίσθητα στην επιλογή των όρων υστέρησης, αλλά δεν είναι ευαίσθητα στο θόρυβο, τουλάχιστον στην περίπτωση προσομοίωσης (Papana et al., 2017). Οι Siggiridou et al., (2019) έχουν επεκτείνει αυτού του είδους τα δίκτυα σε πολυμετάβλητες χρονοσειρές και έχουν θέσει ένα γενικότερο πλαίσιο μεθόδων ανίχνευσης αιτιότητας.

Τέλος, οι μελέτες που είναι πιο κοντά στις εφαρμογές μας είναι των Tang et al. (2019) και των Shirokikh et al. (2022). Στην πρώτη έρευνα έχουμε συσταδοποίηση χρηματιστηριακών αγορών, δημιουργώντας ένα δίκτυο του οποίου οι έδρες έχουν βάρη ίσα με την τιμή του στατιστικού F που προκύπτει από τον έλεγχο ύπαρξης αιτιότητας. χρησιμοποιεί παράθυρα μήκους 50 ημερών, περίπου 2 μήνες συναλλαγών. Επιλέχθηκε

ένα ποσοστό των ακμών προς διατήρηση, με βάση την τιμή του στατιστικού ελέγχου της αιτιότητας κατά Granger. Στη 2η εργασία, μελετώνται στο σύνολο 7240 μετοχές. Γίνεται συσταδοποίηση με κάποιους περιορισμούς μιας και τυπικά κάθε έδρα μεταξύ μετοχών έχει μη μηδενικό βάρος, αν και συχνά αμελητέο. Το πλήθος των μετοχών στην πράξη υποχρέωσε τους ερευνητές να υποθέσουν ότι όλα τα ζεύγη μετοχών αν έχουν υστέρηση, θα παρουσιάζουν υστέρηση 1 λαγκ. Λόγω του μικρού αριθμού μετοχών, μπορέσαμε να αναλύσουμε διεξοδικά τη σχέση υστέρησης για κάθε ζεύγος μετοχών.

2.2 Η μέθοδος των γραφημάτων ορατότητας

Τα γραφήματα ορατότητας προτάθηκαν από τους Lacasa, Bartolo Luque και Nuno το 2008. Αποτέλεσαν μία νέα μέθοδο ανάλυσης χρονοσειρών με τη χρήση μεθόδων από τον τομέα των πολύπλοκων συστημάτων (complex systems). Η βάση της εφαρμογής της μεθόδου είναι η μετατροπή της χρονοσειράς σε ένα γράφο και η ανάλυση μέτρων που ορίζονται σε γράφους για το χαρακτηρισμό των χρονοσειρών από τις οποίες προέκυψαν (Lacasa et al., 2008). Οι γράφοι ορατότητας έχουν χρησιμοποιηθεί σε πλήθος επιστημονικών πεδίων όπου χρησιμοποιούνται χρονοσειρές, με σημαντικές μελέτες στη φυσική, στη γεωφυσική καθώς και στην ιατρική (Zou et al., 2019).

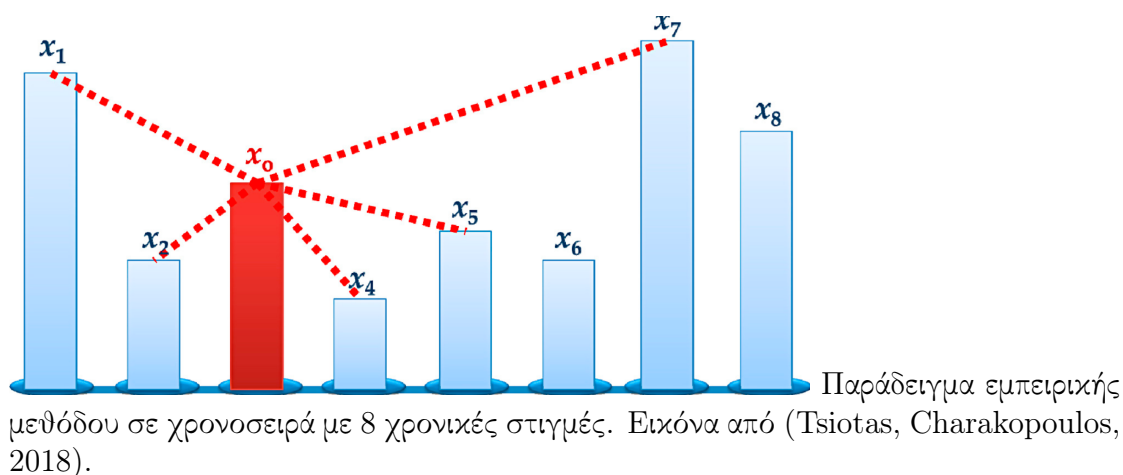
2.2.1 Περιγραφή

Ορίζουμε πρώτα κάποια στοιχεία για τους γράφους. Ένας γράφος αποτελείται από 2 σύνολα, ένα σύνολο κορυφών V καθώς και ένα σύνολο ακμών, E που αποτελεί μία δυαδική σχέση στο $V \times V$ και δείχνει ποιες κορυφές συνδέονται μεταξύ τους. Ο βαθμός μίας κορυφής αποτελεί τον αριθμό των κορυφών με τις οποίες συνδέεται η κορυφή.

Για δοθείσα χρονοσειρά, μπορούμε να βρούμε το γράφο ορατότητας της με τον εξής εμπειρικό τρόπο. Σχεδιάζουμε τη χρονοσειρά σε ένα γράφημα με ράβδους και

1	2	3	4	5	6	7	8
-	1	1	1	1	1	0	0
1	-	1	0	0	0	0	0
1	1	-	1	1	0	1	0
0	0	1	-	1	0	0	0
0	0	1	1	-	1	1	0
0	0	0	0	1	-	1	0
1	0	1	0	1	1	-	1
0	0	0	0	0	0	1	-

ορίζουμε πως 2 σημεία της χρονοσειράς θα συνδέονται αν η γραμμή που συνδέει τις 2 τιμές στο γράφημα, δεν τέμνει κάποια μπάρα. Στην ακόλουθη εικόνα, η 3η στη σειρά χρονική στιγμή συνδέεται με τις χρονικές στιγμές Με βάση το σχήμα, μπορούμε να



συμπληρώσουμε τον πίνακα γειτνίασης του τελικού γράφου, ο οποίος σε αυτή την περίπτωση θα είναι:

Είναι σημαντικό να παρατηρήσουμε ότι το αποτέλεσμα του αλγορίθμου είναι ένας γράφος με κορυφές τις χρονικές στιγμές και όχι τις τιμές της χρονοσειράς. Τυπικά, ας ορίσουμε μία χρονοσειρά τις χρονικές στιγμές $t_i, i \in [1, N]$ με τιμή τη χρονική στιγμή t_i την τιμή $y_i, i \in [1, N]$. Η συνθήκη που δίνει αν 2 χρονικές στιγμές t_a, t_b

είναι συνδεδεμένες είναι η εξής:

$$y_c < y_a + (y_b - y_a) * \frac{(t_c - t_a)}{(t_b - t_a)}, \forall c : a < c < b$$

Ακολουθούμε εδώ τον ορισμό του (Lacasa et al., 2008) που έχει χρησιμοποιηθεί και σε χρηματοοικονομικές εφαρμογές (Yu, 2013), αν και σε κάποιες εφαρμογές δεν απαιτείται η αυστηρή ανισότητα (Yang et al., 2009).

Αυτά είναι τα βήματα για την εύρεση ενός "φυσικού γράφου ορατότητας", όπως ονομάζεται στη βιβλιογραφία (Zou et al., 2019). Υπάρχουν και άλλες παραλλαγές, με κυριότερη τον οριζόντιο γράφο ορατότητας. Μελετήθηκε αρχικά από τους (Luque et al., 2009) και αποτελεί μία απλοποίηση του φυσικού γράφου ορατότητας. Ένας από τους κύριους λόγους για τη χρήση του είναι η μείωση της υπολογιστικής πολυπλοκότητας (Silva et al., 2021). Η διαδικασία δημιουργίας οριζόντιου γράφου ορατότητας είναι αντίστοιχη, αλλά η συνθήκη για τη γειτνίαση 2 χρονικών στιγμών είναι η εξής απλούστερη:

$$y_c < \min(y_a, y_b), \forall c : a < c < b$$

. Τα πλεονεκτήματα της μεθόδου περιλαμβάνουν την απλούστερη γεωμετρική συνθήκη, την απλοποίηση του προβλήματος της εύρεσης του γράφου καθώς και τη δυνατότητα αναλυτικής επίλυσης μαθηματικών προβλημάτων που δημιουργούνται από αυτούς τους γράφους (Luque et al., 2009).

Παράλληλα, υπάρχουν και άλλες μεθοδολογίες για την υλοποίηση γράφων ορατότητας. Μία από αυτές είναι οι γράφοι ορατότητας πεπερασμένης διαπερασιμότητας (Limited Penetrability Visibility Graphs) (Zhou et al., 2012). Η μέθοδος εδώ τροποποιείται ως εξής: για μία παράμετρο N , αν η απόσταση των κόμβων/χρονικών στιγμών είναι μεγαλύτερη από N , οι δύο κόμβοι δε συνδέονται, ανεξαρτήτως της συνθήκης. Επίσης, μία άλλη παράμετρος M καθορίζει πόσες φορές μπορεί να μην ισχύει η ανισότητα (η αντίστοιχη για είτε οριζόντιους είτε φυσικούς γράφους) για τις

ενδιάμεσες χρονικές στιγμές. Η προσέγγιση αυτή μειώνει την επίδραση του θορύβου στα αποτελέσματα (Zhou et al., 2012). Ταυτόχρονα όμως, εξασφαλίζει ότι οι μεταπτώσεις για μικρά χρονικά διαστήματα δε λαμβάνονται υπόψη, αφού οποιεσδήποτε 2 χρονικές στιγμές απέχουν λιγότερο από M , συνδέονται με ακμή.

2.2.2 Σχετικές εφαρμογές

Οι γράφοι ορατότητας διαθέτουν πολλές εφαρμογές στη φυσική, στην ιατρική αλλά και στα οικονομικά (Azizi, 2024). Οι Nunez, Lacasa, Gomez (2009) στη σύνοψη των εξελίξεων τους δίνουν έμφαση στη χρήση των γράφων ορατότητας για την εκτίμηση του εκθέτη Hurst σε φυσικές διαδικασίες. Έμφαση δίνεται σε απλές διαδικασίες για τις οποίες υπάρχουν και θεωρητικά αποτελέσματα, όπως η κλασματική κίνηση Brown. Οι Lacasa et al., (2008) έδειξαν ότι στην κλασματική κίνηση Brown ο γράφος ορατότητας δίνει ένα δίκτυο χωρίς κλίμακα (scale-free). Τα κύρια χαρακτηριστικά του δικτύου είναι η κατανομή των βαθμών των κόμβων, η τοπική υπερσυγκέντρωση ακμών καθώς και άλλα περιγραφικά μεγέθη που προκύπτουν από τη θεωρία γράφων. Σύμφωνα με τους ερευνητές, ο γράφος ορατότητας είναι ένα βήμα προς «μια θεωρία γραφημάτων για χρονοσειρές».

Μία άλλη χρήση των γράφων ορατότητας είναι για το διαχωρισμό συστημάτων. Με τη χρήση της απόκλισης Kullback-Leibler μπορούμε να διακρίνουμε αν ένα σύστημα είναι αντιστρέψιμο ή όχι. Ο τεχνικός ορισμός για την έννοια της αντιστρεψιμότητας δηλώνει πως ένα σύστημα είναι αντιστρέψιμο αν μία υλοποιημένη ακολουθία τιμών του αντεστραμμένη έχει την ίδια κατά συνθήκη πιθανότητα εμφάνισης με την αρχική ακολουθία (Nunez, Lacasa, Gomez 2009). Με ακριβή αποτελέσματα μπορούμε να διακρίνουμε τυχαία από ντετερμινιστικά συστήματα και διατηρητικά συστήματα από συστήματα που έχουν αποσβέσεις με βάση την κατανομή των βαθμών των εδρών.

Σημαντικά αποτελέσματα έχουμε και στην περίπτωση των οριζόντιων γράφων ορατότητας, οπότε κατανομές βαθμού κόμβων είναι διαθέσιμες για την περίπτωση

τχαίων χρονοσειρών (Lacasa et al., 2009). Οι Gutin et al. (2010) καταλήγουν σε ένα χαρακτηρισμό όλων των πιθανών γράφων οριζόντιας ορατότητας. Συγκεκριμένα πρέπει να ικανοποιούν 2 τεχνικές συνθήκες της συνδυαστικής. Επίσης, αποτελέσματα για παραλλαγές των γράφων ορατότητας δόθηκαν από τους Wang et al. (2018) στην περίπτωση των Limited Penetrability Horizontal Visibility Graphs.

Υπάρχουν επίσης αμιγώς οικονομικές εφαρμογές. Οι Serafino et al. (2017) προτείνουν μία διαδικασία για την ανίχνευση χρηματοοικονομικών φουσκών. Χρησιμοποιούν ένα κυλιόμενο παράθυρο, αντίστοιχα με το κυλιόμενο παράθυρο στο (Onnale et al. 2003) για τη συσχέτιση. Στη συγκεκριμένη εργασία, προτείνεται ο έλεγχος της κατανομής των βαθμών του δικτύου καθώς και η υλοποίηση ενός διαχρονικού δείκτη, μεγάλες τιμές του οποίου υποδεικνύουν πιθανή φούσκα. Ο Stephen (2015) προτείνει ένα παρόμοιο πλαίσιο, αλλά αντί για διαχρονική εποπτεία κάποιων δεικτών συνδέει τα δίκτυα μεταξύ τους διαχρονικά, δημιουργώντας ένα δίκτυο δικτύων. Τα αποτελέσματα αυτής της διαδικασίας είναι πειστικά σε επίπεδο τεχνητών παραδειγμάτων αλλά η υλοποίηση του αλγορίθμου για χρηματιστηριακές αγορές της Αμερικής δίνει δύσκολα ερμηνεύσιμα αποτελέσματα.

Παράλληλα, υπάρχουν έρευνες που επικεντρώνονται σε ένα συγκεκριμένο αγαθό και αναλύουν τη δυναμική του δικτύου που προκύπτει από το γράφο ορατότητας. Ενδεικτικά αναφέρουμε την εργασία των Liu et al. (2020) πάνω στο Bitcoin καθώς και την εργασία του Yu (2013) πάνω στο χρυσό. Είναι όμως δύσκολη η σύγκριση των ιδιοτήτων αυτών με άλλα αγαθά και δεν μπορεί να γίνει απευθείας, παρά μόνο με τη χρήση δεικτών χωρίς την απλότητα της ερμηνείας της συσχέτισης (κατανομή βαθμών κόμβων).

Συνοψίζοντας, διατρέξαμε κάποιες από τις κύριες εφαρμογές των γράφων ορατότητας. Αποτελούν ισχυρό εργαλείο για τη μελέτη δυναμικών φαινομένων και μετασχηματίζουν τα δεδομένα από στοιχεία που αφορούν μετοχές σε στοιχεία που αφορούν χρονικές στιγμές. Αυτό αποτελεί και τη μεγάλη αδυναμία του εργαλείου

στην ταξινόμηση μετοχών, μιας και είναι ιδανικό πλέον για να κάνει διακρίσεις ανάλογα με τη χρονική στιγμή και γενικότερα για την παρακολούθηση της δυναμικής εξέλιξης ενός συστήματος. Στην εργασία αυτή διατηρούμε την πρωτότυπη θεώρηση των Lacasa et al. (2008) αλλά δε μεταφέρουμε τα αποτελέσματα στο πεδίο του χρόνου, αντίθετα δημιουργούμε μέτρα ομοιότητας για χρονοσειράς που μπορούν να χρησιμοποιηθούν στο παραδοσιακό υπόδειγμα ταξινόμησης μετοχών που αναλύσαμε.

2.3 Συσταδοποίηση

Η συσταδοποίηση αποτελεί σημαντικό εργαλείο για τη δημιουργία προτύπων πληροφορόρησης (Βλαχάβας κ.ά., 2020). Ο στόχος είναι ο διαχωρισμός των δειγμάτων (στην περίπτωση μας χρονοσειρές τιμών μετοχών) σε συστάδες έτσι ώστε χρονοσειρές στην ίδια ομάδα να 'μοιάζουν' όσο το δυνατόν περισσότερο ενώ χρονοσειρές σε διαφορετικές ομάδες πρέπει να διαφέρουν (Βλαχάβας κ.ά., 2020). Μερικά από τα στοιχεία που χρειάζονται για να εκτελεσθεί η ομαδοποίηση/συσταδοποίηση είναι η επιλογή χαρακτηριστικών, η επιλογή ενός μέτρου εγγύτητας που θα προσδιορίζει ποιες παρατηρήσεις μοιάζουν μεταξύ τους, ένα κριτήριο ομαδοποίησης καθώς και ένα αλγόριθμο ομαδοποίησης (Κουτρούμπας, Θεοδωρίδης 2012).

Ορίζουμε εδώ τεχνικά τη σημασία της m-ομαδοποίησης. Έστω ένα σύνολο $X = s_1, s_2, \dots, s_N$. Τότε μία m-ομαδοποίηση του X είναι μία διαμέριση του X σε m μη κενά υποσύνολα $C_i \forall i \in [1, m]$. Κάθε στοιχείο του X ανήκει σε ένα μόνο C_i και τα σύνολα ικανοποιούν $X = \cup_{i=1}^m C_i$ (Κουτρούμπας, Θεοδωρίδης 2012). Το προβληματικό σημείο είναι η διαμέριση του συνόλου με έναν τρόπο που θα επιτυγχάνεται ο στόχος, οπότε η επιλογή κατάλληλης συνάρτησης ομοιότητας μεταξύ 2 στοιχείων είναι καίρια.

Υπάρχουν 3 κατηγορίες αλγορίθμων ομαδοποίησης, οι αλγόριθμοι που βασίζονται σε διαμερίσεις, οι ιεραρχικοί αλγόριθμοι καθώς και οι πιθανοκρατικοί αλγόριθμοι (Βλαχάβας κ.ά., 2020). Στη μελέτη μας, θα βασιστούμε στους ιεραρχικούς αλγόριθμους, για τους οποίους υπάρχουν 2 προσεγγίσεις, οι προσαυξητικοί agglomerative

και οι διαιρετικοί divisive (Alpaydin, 2020). Οι αλγόριθμοι αυτοί ανάλογα με τον ορισμό των αποστάσεων σημείων από συστάδες δε χρειάζονται το σύνολο της αρχικής πληροφορίας αλλά μόνο τον πίνακα αποστάσεων μεταξύ σημείων (Murphy, 2022). Στη συσσωρευτική μέθοδο που περιγράφεται και από το Mantegna (1999), έχουμε πως ξεκινάμε από τόσες συστάδες όσα έχουμε σημεία. Έπειτα ιεραρχικά συνενώνουμε συστάδες μέχρι να παραμείνει μόνο 1 συστάδα. Κατά τη διάρκεια αυτής της διαδικασίας, θα υπάρξουν από N έως 1 συστάδες, οπότε μπορούμε να επιλέξουμε το στιγμιότυπο του αλγόριθμου με τον επιθυμητό αριθμό συστάδων. Η ίδια η επιλογή του αριθμού συστάδων αποτελεί μία δύσκολη διαδικασία, με αρκετές εμπειρικές διαδικασίες να έχουν αναδειχθεί, όπως είναι η προανάλυση με PCA, το θέσιμο μέγιστου σφάλματος/απόκλισης (Alpaydin, 2020), ή η μέθοδος elbow για αριθμητικά δεδομένα (Βλαχάβας κ.ά., 2020). Είναι επίσης αποδεκτό το a priori θέσιμο μίας τιμής, όπως κάνουμε στην εφαρμογή μας (Alpaydin, 2020), με βάση βέβαια το διαχωρισμό που προτείνει ο δείκτης NASDAQ.

Τέλος, απαιτείται η ανάλυση σε ξεχωριστό επίπεδο του καθορισμού αποστάσεων συστάδων και σημείων. Τα μεγέθη αυτά δεν ορίζονται μονοσήμαντα από το μέτρο εγγύτητας. Μερικές συνήθεις παραδοχές είναι πως η απόσταση συστάδας και σημείου ισούται με τη μικρότερη απόσταση μεταξύ του σημείου και κάποιου σημείου της συστάδας single-link clustering και η απόσταση συστάδας και σημείου ισούται με τη μεγαλύτερη απόσταση μεταξύ του σημείου και κάποιου σημείου της συστάδας complete link clustering (Βλαχάβας κ.ά., 2020). Υπάρχουν και άλλοι αλγόριθμοι που δε θα μάς απασχολήσουν εδώ, όπως είναι η μέθοδος σταθμισμένης μέση απόστασης ομάδων ζεύγους, η μη σταθμισμένη μέση απόσταση ομάδων καθώς και ο αλγόριθμος ελάχιστης διακύμανσης. Οι διαδικασίες αυτές βρίσκονται σε ένα φάσμα μεταξύ των 2 διαδικασιών που αναφέραμε (Βλαχάβας κ.ά., 2020). Εξετάζοντας τα χαρακτηριστικά των αποτελεσμάτων, ο αλγόριθμος απλού δεσμού δημιουργεί συχνά ομάδες που είναι επιμήκεις, το λεγόμενο chaining effect (Βλαχάβας κ.ά., 2020). Αντίθετα, ο αλ-

γόριθμος πλήρους δεσμού δημιουργεί συμπαγείς ομάδες. Και οι δύο αλγόριθμοι είναι σχετικά ευαίσθητοι σε ακραία σημεία (Timm, 2002). Ένα από τα πλεονεκτήματα και των δύο αλγορίθμων σύνδεσης έναντι εναλλακτικών είναι ότι για την υλοποίησή τους απαιτείται μόνο η μήτρα αποστάσεων και όχι οι ίδιες οι τιμές των χρονοσειρών.

Το τελικό αποτέλεσμα της συσταδοποίησης συνήθως απεικονίζεται ως μία ιεραρχική δομή που ονομάζεται δενδρόγραμμα (Alpaydin, 2020). Η επιλογή ενός αριθμού ομάδων σε αυτή την περίπτωση ανάγεται στην επιλογή ενός επιπέδου τομής του δενδρογράμματος έτσι ώστε να τέμνεται ο αντίστοιχος αριθμός γραμμών κάθετα (Βλαχάβας κ.ά., 2020). Τα τελικά 'φύλλα' του δέντρου αντιστοιχούν στα στοιχεία του συνόλου προς διαμέριση (Alpaydin, 2020).

Σε τεχνικό επίπεδο, λόγω της επαναληπτικής διαδικασίας οι αλγόριθμοι ιεραρχικής συσταδοποίησης δεν είναι αποδοτικοί για μεγάλα σύνολα προς διαμέριση, αλλά αυτό δεν αποτελεί πρόβλημα στην παρούσα εφαρμογή.

Συχνά στην περίπτωση χρονοσειρών μετοχών έχουμε τις τιμές των μετοχών ως μοναδικό χαρακτηριστικό, αν και άλλες περιπτώσεις αναδεικνύεται και ο όγκος των συναλλαγών έχουν μελετηθεί (Podobnik, Stanley, 2008) και οι ερευνητές προχωρούν σε πολυδιάστατη ανάλυση. Στην παρούσα μελέτη επιλέχθηκε η τιμή της μετοχής ως μόνο χαρακτηριστικό.

2.3.1 Μετρικές ομοιότητας

Οι μετρικές ομοιότητας που ορίζονται σε κάθε περίπτωση εξαρτώνται από τη φύση των δεδομένων. Τα δεδομένα που είναι αριθμητικά παρέχουν περισσότερες επιλογές από τα (διατεταγμένα) κατηγορικά δεδομένα όπου η δυνατότητα αριθμητικής επεξεργασίας είναι περιορισμένη. Παράλληλα, υπάρχουν ειδικά μέτρα για την περίπτωση των δυαδικών δεδομένων (Βλαχάβας κ.ά., 2020) καθώς και διακριτά δεδομένα (Θεοδωρίδης, Κουτρούμπας, 2014).

Μία κοινή κατηγοριοποίηση των μέτρων ομοιότητας διαχωρίζει ανάμεσα στις τε-

χνικές όπου υπάρχει συγχρονισμός και τα τμήματα των χρονοσειρών εξετάζονται για ομοιότητα κατά το ίδιο χρονικό διάστημα και στις δυναμικές τεχνικές όπου υπάρχει ελαστικότητα ως προς την προσαρμογή των σημείων που μοιάζουν οι χρονοσειρές (Wang et al., 2012). Στις 'στατικές' μεθόδους περιλαμβάνεται και η Ευκλείδεια απόσταση (Serra, Arcos, 2024). Η Ευκλείδεια απόσταση αποτελεί εξειδίκευση της μετρικής Minkowski (Βλαχάβας κ.ά, 2020), που δίνεται από τον ακόλουθο τύπο για 2 χρονοσειρές x, y , μήκους N :

$$D(x, y) = \sqrt[q]{\sum_{i=1}^{i=N} (x_i - y_i)^q}$$

Η μετρική Minkowski λαμβάνει υπόψη τις αποκλίσεις μεταξύ ζευγών τιμών των χρονοσειρών για κάθε χρονική στιγμή και τις συναθροίζει. Για $q = 2$ έχουμε την Ευκλείδεια απόσταση, για $q = 1$ τη μετρική Manhattan (άθροισμα απόλυτων τιμών).

Η ομοιότητα συνημιτόνου ή αποτελεί ένα άλλο στατικό μέτρο. Βρίσκει το συνημίτονο της γωνίας μεταξύ 2 διανυσμάτων, με μεγαλύτερα συνημίτονα να δείχνουν μεγαλύτερη ομοιότητα (Βλαχάβας κ.ά, 2020). Μία άλλη αντίστοιχη τεχνική είναι η συσχέτιση κατά Pearson. Η τιμή του μέτρου για 2 χρονοσειρές x, y , μήκους N δίνεται από:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 * \sum (y_i - \bar{y})^2}}$$

Κατά βάση, αποτελεί παραλλαγή της ομοιότητας συνημιτόνου αφού χρησιμοποιεί το εσωτερικό γινόμενο των αποκλίσεων των διανυσμάτων από το μέσο τους, παρά το εσωτερικό γινόμενο των ίδιων των διανυσμάτων. Στα οικονομικά και ειδικά στο θέμα των μετοχών, η χρήση της συσχέτισης συνηθίζεται σε χρονοσειρές που δείχνουν τις επιστροφές των μετοχών ανά χρονική περίοδο (Mantegna, 1999), (Dolphin et. al, 2021), (Anderson et al., 2007). Ένα από τα προβλήματα της απευθείας εφαρμογής της συσχέτισης είναι η ύπαρξη τάσεων στη χρονοσειρά, που δε λαμβάνονται υπόψη

από τη συσχέτιση Pearson (Dolphin et. al, 2021).

Προχωρούμε τώρα σε στατικά μέτρα για δυαδικά δεδομένα. Η πρώτη μετρική που συναντάται είναι η κωδικοποίηση Hamming, που στην περίπτωση δυαδικών δεδομένων αποτελεί απλά τον αριθμό των συνιστωσών στα οποία διαφέρουν τα δυαδικά διανύσματα. Η κωδικοποίηση Hamming για δυαδικά δεδομένα ταυτίζεται με τη μετρική Minkowski για $q = 1, 2$. Συμπυκνώνοντας την επεξήγηση του Βλαχάβας και των συνεργατών του (2020), για 2 ζεύγη δυαδικών χρονοσειρών, υπάρχουν 3 ενδεχόμενα για κάθε χρονική στιγμή t . Είτε, οι τιμές διαφέρουν, με τη μία να είναι ίση με 0 και την άλλη ίση με 1, είτε οι δύο τιμές είναι ίσες με 0, είτε οι δύο τιμές είναι ίσες με 1. Αντίστοιχα ορίζουμε τον αριθμό των χρονικών στιγμών για τις οποίες ισχύει η κάθε συνθήκη ως M_{01}, M_{00}, M_{11} . Τότε, ένας απλός δείκτης είναι ο SMC (simple matching coefficient) ο οποίος ορίζεται ως εξής (Βλαχάβας κ.ά, 2020):

$$SMC = \frac{M_{00} + M_{11}}{M_{00} + M_{11} + M_{01}}$$

Δηλαδή ο SMC δείχνει την αναλογία των φορών που οι δύο χρονοσειρές έχουν την ίδια τιμή. Ο δείκτης SMC είναι κατάλληλος για περιπτώσεις όπου οι καταστάσεις 0 και 1 ενδιαφέρουν εξίσου τον ερευνητή. Σε πολλές περιπτώσεις, έχουμε πως οι καταστάσεις με τιμή 1 υποδηλώνουν παρουσία και διαθέτουν σημαντικό ενδιαφέρον για την ανάλυση μας. Στην περίπτωση αυτή έχουμε το δείκτη Jaccard που αποτυπώνει το λόγο του αριθμού των χρονικών στιγμών στις οποίες και οι 2 χρονοσειρές είναι ίσες με 1, προς τον αριθμό των χρονικών στιγμών που έστω 1 χρονοσειρά είναι ίση με 1 (van Eck, 2009). Αριθμητικά έχουμε:

$$X = \frac{M_{11}}{M_{11} + M_{01}}$$

Τα μέτρα απόστασης αντίστοιχα για τις 2 μεθόδους ορίζονται ως $1 - SMC$, $1 - J$ αφού και οι δύο δείκτες είναι μικρότεροι της μονάδας (Βλαχάβας κ.ά, 2020).

Τέλος, κάνουμε αναφορά σε δυναμικούς δείκτες ομοιότητας χρονοσειρών. DTW-Dynamic Time Warping είναι ένας από αυτούς που έχει χρησιμοποιηθεί με επιτυχία για ταξινόμηση (Wang et al., 2012), αν και δεν έχει χρησιμοποιηθεί αρκούντως στα οικονομικά (Frances, Wiemann, 2020). Ο αλγόριθμος DTW βρίσκει ένα μέτρο ομοιότητας το οποίο βασικά επαναπροσδιορίζει τη χρονική ακολουθία έτσι ώστε να ελαχιστοποιηθεί ένα κόστος ευθυγράμμισης (Θεοδωρίδης κ.ά., 2011). Χρησιμοποιεί μεθόδους δυναμικού προγραμματισμού (Θεοδωρίδης, Κουτρούμπας, 2014). Αποτελεί μία τεχνική που έχει δείξει καλύτερα αποτελέσματα από τη συσχέτιση κατά ΠEARSON (Frances, Wiemann, 2020) και αποτελεί μία σημαντική μέθοδο για την αξιολόγηση νέων κριτηρίων ομοιότητας (Serra, Arcos, 2014).

Τέλος, μία άλλη μέθοδος σύγκρισης χρονοσειρών είναι η μοντελοποίηση τους με κοινά μοντέλα και έπειτα η σύγκριση των συντελεστών/ παραμέτρων των μοντέλων (Serra, Arcos, 2014). Στη συγκεκριμένη περίπτωση, θα υλοποιηθεί η ίδια ανάλυση με τις χρονοσειρές που προκύπτουν από την επεξεργασία του γράφου ορατότητας.

2.3.2 Μετρικές αξιολόγησης συσταδοποίησης

Η μηχανική μάθηση στην οποία η συσταδοποίηση παίζει σημαντικό ρόλο, έχει δώσει σημαντικά αποτελέσματα σε σχέση με τις μετρικές αξιολόγησης τους. Δύο σημαντικές παράμετροι που πρέπει να ληφθούν υπόψη σε κάθε μετρική είναι το κατά πόσο οι ομάδες είναι συμπαγείς και το κατά πόσο οι ομάδες είναι διαχωρισμένες, που μπορούν να προσεγγισθούν με τη διασπορά τιμών σε κάθε ομάδα καθώς και τα αντίστοιχα κριτήρια δεσμού για τη συσταδοποίηση (Χαλκίδη, Βαζιργιάννης, 2005).

Για τους σκοπούς της εργασίας αυτής, ιδιαίτερη έμφαση θα δοθεί στη σύγκριση 2 συσταδοποιήσεων. Στη γενική περίπτωση, ενώ η συσταδοποίηση είναι αλγόριθμος της ανεπίβλεπτης μηχανικής μάθησης και συνεπώς τα στοιχεία δεν έχουν ετικέτες (Murphy, 2022), όταν εφαρμόζουμε συσταδοποίηση σε ετικετοποιημένα δεδομένα, μπορούμε να χρησιμοποιήσουμε αυτά τα εξωτερικά δεδομένα για την αξιολόγηση της

ποιότητας της συσταδοποίησης (Murphy, 2022).

Ακολουθώντας τους Χαλκίδη και Βαζιργιάννη (2005), έστω δύο τμηματοποιήσεις των στοιχείων του X $P = [P_1, P_2, \dots, P_n]$ και $R = [R_1, R_2, \dots, R_m]$. Τότε για κάθε ζεύγος διαφορετικών στοιχείων του $Q(x_1, x_2)$ (χρονοσειρές μετοχών στην περίπτωση μας) μπορούμε να διακρίνουμε 4 περιπτώσεις. Είτε τα δύο σημεία ανήκουν στις ίδιες συστάδες και στις 2 περιπτώσεις ($x_1 \in P_i, x_2 \in P_i, x_1 \in C_j, x_2 \in C_j$), είτε τα δύο σημεία ανήκουν σε διαφορετικές συστάδες και για τις 2 συσταδοποιήσεις, είτε τα δύο σημεία ανήκουν στην ίδια συστάδα για την P και σε διαφορετικές για τη C , είτε τα δύο σημεία ανήκουν στην ίδια συστάδα για την C και σε διαφορετικές για τη P . Έστω ότι έχουμε τον αριθμό των ζευγών για τα οποία ισχύει η κάθε συνθήκη και είναι ίσες με a, d, b, c . Παράλληλα, έχουμε πως τα συνολικά ζεύγη διαφορετικών σημείων είναι ίσα με $N(N - 1)/2$ όπου N το σύνολο των σημείων του X .

Ορίζουμε τότε 2 βασικά μέτρα ομοιότητας συσταδοποίησης, το δείκτη Rand $Ra(R, P) = \frac{a+d}{N(N-1)/2}$ καθώς και το δείκτη Jaccard $J(P, R) = \frac{a}{a+b+c}$ (Χαλκίδη, Βαζιργιάννη, 2005) με αντίστοιχη λογική με τους δείκτες ομοιότητας στοιχείων. Ειδικά ο δείκτης Rand μπορεί να οριστεί ως η αναλογία των σωστών αποφάσεων του P με βάση τη συσταδοποίηση του R (Murphy, 2022). Και οι 2 δείκτες κυμαίνονται από 0 έως 1 και υψηλές τιμές υποδεικνύουν ομοιότητα στη συσταδοποίηση.

Ένας άλλος απλός δείκτης είναι η 'καθαρότητα' της συσταδοποίησης. Για κάθε συστάδα της συσταδοποίησης, μπορούμε να ορίσουμε την καθαρότητα της ως την αναλογία των δειγμάτων με τη συχνότερη ετικέτα σε κάθε συστάδα. Η 'καθαρότητα' της συσταδοποίησης έπειτα δίνεται από το σταθμισμένο μέσο της καθαρότητας όλων των συστάδων, με βάρη τον αριθμό των σημείων σε κάθε συστάδα (Murphy, 2022). Οι δείκτες καθαρότητας, Rand και Jaccard θα αποτελέσουν τα κύρια εργαλεία εξωτερικής αξιολόγησης των συσταδοποιήσεων μας.

Παράλληλα υπάρχουν και εσωτερικά μέτρα αξιολόγησης των συστάδων, από τα οποία το κυριότερο που θα χρησιμοποιήσουμε είναι ο δείκτης σχιαγράφησης, που

χρησιμοποιείται και για την επιλογή του αριθμού των συστάδων (Liu et al., 2010). Ο δείκτης σχιαγράφησης για ένα σημείο i είναι ίσος με $|1 - a/b|$ όπου ορίζουμε ως a τη μέση απόσταση του σημείου i από τα υπόλοιπα σημεία της συστάδας όπου ανήκει και ως b "τη μέση απόσταση του σημείου από την κοντινότερη συστάδα" (Βλαχάβας κ.ά, 2020). Η απόλυτη τιμή χρησιμοποιείται σπάνια, μιας και στις περισσότερες εφαρμογές $a < b$ (Βλαχάβας κ.ά, 2020). Ο δείκτης σχιαγράφησης μίας συσταδοποίησης είναι ο μέσος των δεικτών για κάθε σημείο.

2.4 Δέντρα αποφάσεων και τυχαία Δάση

Τα δέντρα αποφάσεων αποτελούν σημαντικό εργαλείο για την κατηγοριοποίηση δεδομένων, την παλινδρόμηση (Διαμαντάρας, Μπότσης, 2019) αλλά και την πρόβλεψη δεδομένων (Χαλκίδη, Βαρζιγιάννης, 2005). Αποτελεί επαναληπτικό μηχανισμό επιβλεπόμενης μάθησης και απαιτούνται τα δείγματα με τα χαρακτηριστικά τους καθώς και η κατηγοριοποίηση τους σε κλάσεις. Η βασική ιδέα του μηχανισμού των δέντρων αποφάσεων είναι πως αφού οριστεί ο πολυδιάστατος χώρος όπου βρίσκονται τα δείγματα, ο χώρος αυτός υπόκειται σε κατάτμηση σε "κανονικά" χωρία, όπου σε κάθε χωρίο η ίδια κλάση αποδίδεται σε όλα τα δείγματα (Θεοδωρίδης, Κουτρούμπας, 2014).

Η κατάτμηση του χώρου γίνεται με βάση ένα συγκεκριμένο κριτήριο που κρίνει την ανομοιογένεια του κάθε χωρίου και σε κάθε βήμα διαιρεί ένα χωρίο κατά μία διάσταση για να αυξηθεί η επιμέρους ομοιογένεια των χωρίων (Διαμαντάρας, Μπότσης, 2019). Η συνήθης επιλογή των χωρίων είναι τα ορθογώνια χωρία, που προκύπτουν από δυαδικούς διαχωρισμούς ανά διάσταση, δηλαδή υπερεπίπεδα τομής που δίνονται από εξισώσεις όπως $x_j = s$, όπου s είναι η συννοριακή γραμμή. Τα συνήθη κριτήρια μέτρησης της ανομοιογένειας είναι 2, το κόστος Gini και η εντροπία. Σύμφωνα με τον ορισμό στο Διαμαντάρας, Μπότσης (2019), για δοθέν χωρίο με συνολικά N πρότυπα και K κλάσεις έχουμε τους εξής ορισμούς για την εντροπία E και το κόστος

Gini G :

$$E = - \sum_{k=1}^K p(k) * \log_2 p(k)$$
$$G = - \sum_{k \neq l} p(k) * p(l)$$

όπου $p(k)$ το ποσοστό των προτύπων που διαθέτουν την κλάση k . Η διαμέριση που επιλέγεται ελαχιστοποιεί το σταθμισμένο μέσο του δείκτη ανομοιογένειας ανά χωρίο.

Ο μηχανισμός ονομάζεται δέντρο απόφασης γιατί το τελικό αποτέλεσμα μπορεί να παρασταθεί ως μια δενδροειδής δομή, στην οποία οι κόμβοι αποτελούν ελέγχους συνθηκών ενώ τα φύλλα φέρουν ονόματα κλάσεων και δείχνουν την τελική απόφαση που λαμβάνεται για την κατηγοριοποίηση ενός προτύπου (Διαμαντάρας, Μπότσης, 2019).

Μερικά από τα πλεονεκτήματα των δένδρων αποφάσεων είναι η ευκολία κατανόησης τους και το λογαριθμικό βάθος σε σχέση με το πλήθος δεδομένων που οδηγούν σε χαμηλό κόστος υπολογισμών, ωστόσο είναι αδύναμοι αλγόριθμοι μάθησης weak learners και μεταξύ άλλων εμφανίζουν αστάθεια σε σχέση με τις αρχικές εισόδους (Διαμαντάρας, Μπότσης, 2019). Η λύση για αυτό το ζήτημα βρίσκεται στη μέθοδο του boosting, που αποτελεί μία μέθοδο συνάνθροισης των αποτελεσμάτων πολλών ασθενών μοντέλων για τη δημιουργία ισχυρού μοντέλου με το σκεπτικό της σοφίας του πλήθους (Βλαχάβας κ.ά., 2020). Η αρχή αυτή προϋποθέτει τη μείωση της συσχέτισης των δένδρων.

Η διαδικασία του τυχαίου δένδρου αναλύεται στην εργασία του Ho (1995) και μερικά από τα βασικά σημεία είναι η δημιουργία n υποσυνόλων του αρχικού συνόλου δεδομένων μας με δειγματοληψία όπου επιτρέπεται η επανάληψη, με κάθε υποσύνολο να έχει πληθικότητα ίση με την πληθικότητα του αρχικού συνόλου δειγμάτων. Από κάθε υποσύνολο θα κατασκευαστεί ένα δέντρο απόφασης με βάση ένα υποσύνολο χαρακτηριστικών του συνόλου δειγμάτων. Η ταξινόμηση ενός δείγματος γίνεται για κάθε δένδρο ξεχωριστά και η τελική κλάση ταξινόμησης είναι η κλάση που είναι η

πλειοψηφία των αποτελεσμάτων των δέντρων.

Τα τυχαία δάση αποτελούν ισχυρούς ταξινομητές που δεν υπερπροσαρμόζουν στα δεδομένα (Βλαχάβας κ.ά., 2020) σε αντίθεση με τα δέντρα αποφάσεων (Διαμαντάρας, Μπότσης, 2019) Αυτό οφείλεται στην ελάχιστη συσχέτιση μεταξύ των υποκείμενων δέντρων αποφάσεων, που υλοποιούνται σε διαφορετικά υποσύνολα και με διαφορετικά χαρακτηριστικά (Βλαχάβας κ.ά., 2020).

2.5 Μοντέλα VAR και Granger Αιτιότητα

Συχνά η μοντελοποίηση χρονοσειρών παρουσιάζει ιδιομορφίες όπως τη σειριακή συσχέτιση που σημαίνει πως οι τιμές ενός μεγέθους είναι συσχετισμένες στο χρόνο (Diebold, 2019). Συνήθως η ανάλυση χρονοσειρών υποθέτει κάποιο αιτιακό μηχανισμό ή τουλάχιστον μία διαίρεση σε επεξηγηματικές και εξαρτημένες μεταβλητές (Sims, 1980). Ένας τρόπος αντιμετώπισης των 2 θεμάτων είναι η χρήση των Vector Autoregressive Models-VAR. Στα μοντέλα αυτά υπάρχει συμμετρία μεταξύ των μεταβλητών, με κάθε χρονοσειρά να επηρεάζει και ως εξαρτημένη και ως ανεξάρτητη μεταβλητή κάθε άλλη μέσω χρονικών υστερήσεων (Asteriou, Hall, 2019). Συγκεκριμένα σε διανυσματική μορφή όπου $X_t = [x_1, x_2 \dots x_n]$ και x_i οι τιμές των n χρονοσειρών που μοντελοποιούνται τη στιγμή T , έχουμε πως η χρονοσειρά μοντελοποιείται ως εξής, αν ληφθούν υπόψη p υστερήσεις (Asteriou, Hall, 2019):

$$X_t = A_0 + \sum_{i=1}^p A_i X_{t-i}$$

Η εκτίμηση των παραμέτρων μπορεί να γίνει πολύ απλά με μεθόδους γραμμικής παλινδρόμησης στις επιμέρους εξισώσεις που προκύπτουν από τη διανυσματική μορφή του μοντέλου (Diebold, 2019), οδηγώντας τη διαδικασία εκτίμησης στις σταθερές ιδιότητες του αλγορίθμου OLS. Η επιλογή της μεθόδου αυτή διαθέτει και αρνητικά στοιχεία, με ιδιαίτερη έμφαση να δίνεται στις αυστηρές προϋποθέσεις για την εφαρμο-

γή της. Μερικές από αυτές είναι η γραμμικότητα του μοντέλου, η γνώση της υστέρησης καθώς και η συμπεριφορά των υπολοίπων που πρέπει να είναι στάσιμη (Shojaie, Fox, 2022), (Neusser, 2016). Συγκεκριμένα η απαίτηση της ασθενούς στασιμότητας ικανοποιείται σπάνια, μιας και απαιτείται σταθερός μέσος όρος πεπερασμένη διασπορά και ανεξάρτητη αυτοσυσχέτιση στο χρόνο, ανεξαρτήτως υστερήσεων (Neusser, 2016) που υποδηλώνει σταθερές διασπορές στο χρόνο (Diebold, 2019).

Το μοντέλο VAR μπορεί να συλλάβει δυναμικές αλληλεπιδράσεις στο χρόνο μεταξύ μεταβλητών και προσφέρεται για τον ορισμό ενός μεγέθους προβλεψιμότητας μεταξύ χρονοσειρών, της αιτιότητας κατά Granger που βασίζεται σε εκτίμηση μοντέλων VAR. Όταν εισήχθη στην επιστημονική βιβλιογραφία, η αιτιότητα κατά Granger θεωρήθηκε ως ένα μέτρο της αιτιότητας μεταξύ των χρονοσειρών και ως ικανό να αποδώσει αιτιακά φαινόμενα (Σηροθαie, Φοξ, 2022). Ωστόσο, αποτελεί ένα μέτρο του κατά πόσο μία χρονοσειρά περιέχει πληροφορία που μπορεί να χρησιμοποιηθεί γραμμικά για την πρόβλεψη των άλλων χρονοσειρών που εμπλέκονται στο μοντέλο VAR (Diebold, 2019) με περιορισμένη χρησιμότητα σε αιτιακά μοντέλα (Shojaie, Fox, 2022). Ακολουθούμε την παρουσίαση των Asteriou, Hall (2019) για την ανάλυση του ελέγχου στην περίπτωση 2 μεταβλητών χρονοσειρών x, y . Για να ελέγξουμε αν η χρονοσειρά x προκαλεί την y θεωρούμε τις 2 εξισώσεις με βαθμό υστέρησης p :

$$y_t = A_0 + \sum_{i=1}^p A_i y_{t-i} + \sum_{i=1}^p B_i x_{t-i}$$

$$y_t = A_0 + \sum_{i=1}^p A_i y_{t-i}$$

Το πρώτο μοντέλο είναι η εξίσωση που προκύπτει από το VAR(p) για τη χρονοσειρά y , ενώ η δεύτερη εξίσωση μπορεί να θεωρηθεί ως ο περιορισμός της πρώτης (συνεπώς θα έχει μεγαλύτερο άθροισμα τετραγωνικών υπολοίπων RSS). Η μηδενική μας υπόθεση H_0 είναι ότι η x δεν επηρεάζει την y και συνεπώς ότι $\sum_i B_i = 0$. Η H_1 είναι

ότι $\sum_i B_i \neq 0$. Η στατιστική ελέγχου δίνεται σε σχέση με τα RSS των 2 εξισώσεων. Έχουμε πως

$$F = \frac{(RSS_2 - RSS_1)/p}{RSS_1/(p+1)}$$

όπου η στατιστική F ακολουθεί την κατανομή $F_{p,p+1}$.

Δύο σχόλια είναι σημαντικό να αναφερθούν εδώ. Η επιλογή του σωστού όρου υστέρησης αποτελεί σημαντική παράμετρο για την υλοποίηση των γραμμικών παλινδρομήσεων και οι συνήθεις περιορισμοί του μοντέλου της απλής γραμμικής παλινδρόμησης εφαρμόζονται και εδώ για την περίπτωση της εκτίμησης παραμέτρων (ασυσχέτιστα υπόλοιπα, σταθερή διασπορά) (Diebold, 2019). Επίσης, η σύνδεση μέσω αιτιότητας δεν είναι αντιμεταθετική, δηλαδή αν η x προκαλεί την y τότε δεν είναι απαραίτητο ότι και η y θα προκαλεί τη x κατά Granger. Συνεπώς μπορούμε να έχουμε 4 διαφορετικά ενδεχόμενα, με το φαινόμενο της αμοιβαίας αιτιότητας να παρουσιάζει ιδιαίτερο ενδιαφέρον (Diebold, 2019).

3 Μεθοδολογία

3.1 Μεθοδολογικό πλαίσιο

Η επιλογή μεθοδολογίας για την υλοποίηση ενός ερευνητικού πρότζεκτ αποτελεί σημαντική παράμετρο για την επιτυχία του (Denscombe, 2010). Κύρια ερωτήματα για την επιλογή της ερευνητικής στρατηγικής αποτελούν το κατά πόσο είναι κατάλληλη για την απάντηση του προβλήματος, κατά πόσο είναι υλοποιήσιμη και κατά πόσο έχει επικίνδυνες επιπτώσεις ηθικά (Denscombe, 2010). Τα ερευνητικά ερωτήματα μάς καθοδηγούν στην επιλογή ενός πλαισίου ιδανικού για ποσοτική έρευνα. Ειδικά στον τομέα των οικονομικών, ένα κατάλληλο πλαίσιο αποτελεί ο ποζιτιβισμός, με την έμφαση που δίνει στην αναγωγή των φαινομένων σε απλούστερα, τον έλεγχο υποθέσεων καθώς και το μετασχηματισμό εννοιών σε μετρήσιμα μεγέθη (Patterson

et al., 2016).

Ένα σύννηδες υπόδειγμα που ακολουθείται συχνά στην επιστημονική μέθοδο είναι η υποθετικοπαραγωγική προσέγγιση που μπορεί να συμπυκνωθεί σε 4 βήματα. Πρώτον, η δημιουργία ερωτήσεων για ένα φαινόμενο, κατά δεύτερον η διαμόρφωση υποθέσεων, κατά τρίτον η παραγωγή προβλέψεων με βάση τις υποθέσεις και τέλος ο έλεγχος των προβλέψεων (Johannesson, Perjons, 2014). Σε αυτό το σχήμα, έχουμε πως οι υποθέσεις, συνήθως κάποια αιτιακή σχέση (Johannesson, Perjons, 2014) αποτελούν ένα μέσο για να παραχθούν μετρήσιμες προβλέψεις, που μπορούν έπειτα να ελεγχθούν ώστε να αξιολογηθεί η καταλληλότητα των υποθέσεων.

Η υποθετικοπαραγωγική προσέγγιση είναι κατάλληλη για τον έλεγχο ερωτημάτων αλλά δεν είναι προσαρμοσμένη για την αντιμετώπιση προβλημάτων, παρα μόνο αν διαμορφωθούν ερωτήματα ειδικά σε σχέση με κάποια πτυχή του προβλήματος. Στην παρούσα εργασία, είναι σημαντική η υλοποίηση νέων μετρικών με βάση υπάρχουσες έννοιες για την αντιμετώπιση προβλημάτων. Χρησιμοποιείται εν μέρει το υπόδειγμα της επιστήμης του σχεδιασμού. Η επιλογή αυτή δικαιολογείται από την ανάγκη δημιουργίας ενός αντικειμένου για πρακτική εφαρμογή σε πρόβλημα, δηλαδή τις μεθόδους των γράφων ορατότητας σε χρονοσειρές για συσταδοποίηση. Σύμφωνα με τον ορισμό των Johannesson, Perjons (2014), 'η επιστήμη του σχεδιασμού αποτελεί την επιστημονική μελέτη και δημιουργία αντικειμένων για χρήση των ανθρώπων με σκοπό την επίλυση προβλημάτων γενικού ενδιαφέροντος'. Η συσταδοποίηση μετοχών έχει σημαντικό γενικό ενδιαφέρον και οι υλοποιήσεις των συναρτήσεων μας αποτελούν ένα εύχρηστο εργαλείο έτοιμο προς χρήση. Η επιστημονικότητα της μεθόδου εξασφαλίζεται εν μέρει και από τη διασύνδεση με τη βάση γνώσης για αυτό το πρόβλημα. Σε αυτό το υπόδειγμα, έχουμε 5 βήματα που αφορούν τη μέθοδο αν και εργασίες μπορούν να επικεντρωθούν σε μερικά από τα βήματα: 'επεξήγηση του προβλήματος, καθορισμός των απαιτήσεων, σχεδίαση αντικειμένου, επίδειξη αντικειμένου, αξιολόγηση αντικειμένου' (Johannesson, Perjons, 2014). Η παρούσα εργασία αξιοποιεί

για τα 2 πρώτα βήματα την ακαδημαϊκή βιβλιογραφία πάνω στο πρόβλημα, σχεδιάζει τα αντικείμενα προς επίλυση του προβλήματος και δίνει έμφαση στην επίδειξη των νέων αντικειμένων / μετρικών καθώς και την αξιολόγηση της χρησιμότητας τους σε επίπεδο αποτελεσματικότητας.

Συγκεκριμένα, έχουμε πως το αντικείμενο που δημιουργούμε έχει ως στόχο μία μέτρηση της ομοιότητας μεταξύ 2 χρονοσειρών. Σύμφωνα με την κατηγοριοποίηση των Johannessson, Perjons (2014) για τα αντικείμενα, σχεδιάζουμε όρους/ορισμούς που δεν παράγουν περιγραφική γνώση για τον κόσμο από μόνα τους. Σε αυτή την κατηγορία εμπίπτουν ο αλγόριθμος ομαλής ορατότητας που υλοποιείται καθώς και οι 2 μετρικές ομοιότητας χρονοσειράς με βάση τους δείκτες SMC και Jaccard. Ο στόχος της έρευνας όμως είναι να παράξει περιγραφική γνώση καθώς και γνώση πρόβλεψης. Κοινό στοιχείο και των 2 είναι ότι δεν προϋποθέτουν την ύπαρξη αιτιακού μηχανισμού μεταξύ των μεταβλητών και δεν επιδιώκουμε μία επεξήγηση σε αυτό το στάδιο.

3.1.1 Ερευνητικά Ερωτήματα

Με βάση την ως άνω πλαισίωση της έρευνας, προχωρούμε στη διατύπωση των ερευνητικών ερωτημάτων. Τα ερευνητικά ερωτήματα μπορούν να καταμεριστούν σε 2 κύριες κατηγορίες. Η πρώτη οικογένεια ερωτημάτων αφορά τη σχέση των δεικτών που προκύπτουν από την ανάλυση και τη σχέση της αιτιότητας κατά Granger με αυτούς τους δείκτες που μπορεί να ελεγχθεί στις αρχικές χρονοσειρές των μετοχών. Η δεύτερη κατηγορία ερωτημάτων σχετίζεται με τη χρήση των μετρικών για συσταδοποίηση και τη σύγκριση του αποτελέσματος της συσταδοποίησης με άλλες μεθόδους, όπως η μέθοδος των συσχετίσεων των επιστροφών των μετοχών (Mantegna, 1999). Συνεπώς, περιεκτικά:

1) Μπορούμε μόνο από το γράφο ορατότητας 2 χρονοσειρών να προβλέψουμε αν υπάρχει αιτιότητα κατά Granger μεταξύ 2 χρονοσειρών με τη χρήση τυχαίων δασών. Ποια μέτρα και παράμετροι είναι πιο σημαντικά στην κατάρτιση της πρόβλεψης.

2) Πόσο καλή συσταδοποίηση των υποκείμενων μετοχών μπορεί να επιτευχθεί με τη χρήση των μετρικών απόστασης που υλοποιήσαμε. Ποια είναι τα αποτελέσματα της σύγκρισης μας με άλλες μεθόδους συσταδοποίησης.

Για το πρώτο ερώτημα θα είναι κρίσιμα τα εργαλεία αξιολόγησης για δυαδικούς ταξινομητές καθώς και οι έννοιες της ανάλυσης VAR, της αιτιότητας κατά Granger καθώς και οι εκτιμητές ομοιότητας που ορίσαμε.

Για το δεύτερο ερώτημα, η επιλογή των χρονοσειρών που επιτρέπει την ετικετοποίηση καθώς και τα κριτήρια αξιολόγησης συσταδοποίησης θα είναι κρίσιμοι παράγοντες για τη σύγκριση ανάμεσα στα 3 πρότυπα συσταδοποίησης: με συσχέτιση και με κριτήρια ομοιότητας από τους γράφους ορατότητας.

3.1.2 Ερευνητική Στρατηγική

Η κύρια ερευνητική στρατηγική που προκύπτει είναι η χρήση του πειράματος. Σύμφωνα με τον ορισμό του Denscombe, (2010), τα πειράματα είναι κατάλληλα για την παρατήρηση της επιρροής συγκεκριμένων παραγόντων πάνω στο αποτέλεσμα. Στην παρούσα εργασία δεν είναι δυνατός ο έλεγχος των αρχικών δεδομένων ή των ενδιάμεσων μεταβλητών, αλλά χρησιμοποιούνται χρονοσειρές για τον έλεγχο υποθέσεων σε διαφορετικά χρονικά διαστήματα, θεωρείται συνεπώς ένα πλαίσιο ημι-πειράματος (Latham, 2022). Συνεπώς, δεν υπάρχει έλεγχος των πειραματικών μεταβλητών από μέρους του ερευνητή παρά μόνο στο επίπεδο της χρονικής περιόδου και η πιθανότητα ύπαρξης ιδιοσυγκρατικών φαινομένων μπορούν να περιορίσουν τη γενικευσιμότητα των συμπερασμάτων. Παρόλα αυτά, η διαδικασία σύγκριση μέτρων έχει χρησιμοποιήσει χρονοσειρές περιορισμένων χρονικών διαστημάτων για την αξιολόγηση μεθόδων ομοιότητας χρονοσειρών (Χαλκίδη, Βαζιργιάννη, 2005), (Wang et al., 2012) χρησιμοποιούν το συγκεκριμένο πλαίσιο για την αξιολόγηση μετρικών. Μπορεί να χαρακτηριστεί και σα βασική εφαρμογή τριγωνοποίησης (Latham, 2022), όπου για τη μεγαλύτερη γενικευσιμότητα των αποτελεσμάτων χρησιμοποιούνται στοιχεία από

διαφορετικές πηγές και με διαφορετικά μέσα. Η εγκυρότητα των αποτελεσμάτων εξασφαλίζεται εν μέρει από την αξιολόγηση των μεθόδων σε δεδομένα διαφορετικών χρονικών στιγμών. Στην παρούσα εργασία, η αξιολόγηση της καταλληλότητας των μέτρων περιορίζεται στο πεδίο των οικονομικών, παρόλο που η μέθοδος ορατότητας έχει χρησιμοποιηθεί και σε άλλα πεδία. Η γενικευσιμότητα των αποτελεσμάτων βασίζεται στη χρήση πολλαπλών χρονικών παραθύρων σε μη επικαλυπτόμενες χρονικές περιόδους.

Μερικά από τα πλεονεκτήματα που μας οδήγησαν στη συγκεκριμένη επιλογή είναι η επαναληψιμότητα της διαδικασίας καθώς και η ακρίβεια που μπορεί να επιτευχθεί στην ανάλυση μέσω της αλλαγής των χρονικών περιόδων από τον ερευνητή Denscombe, (2010). Παράλληλα θα πρέπει να σημειωθεί ότι όπως σε κάθε πείραμα, έχει μεγάλη σημασία η αντιπροσωπευτικότητα των στοιχείων που χρησιμοποιούνται Denscombe, (2010). Στην παρούσα εργασία, που είναι σε ένα βαθμό εξερευνητική, έχουμε πως η επιλογή του δείγματος έγινε με βάση την ευκολία πρόσβασης στο δείγμα, καθώς και τη διαδικασία διαχωρισμού. Η επιλογή των τριών κατηγοριών, υγεία, τεχνολογία και καταναλωτικά προϊόντα είναι εν μέρει αυθαίρετη ανάμεσα στις 11 κατηγορίες του δείκτη. Κριτήρια επιλογής ήταν η ύπαρξη 6 μετοχών στην κάθε κατηγορία καθώς και η διαφοροποίηση των τομέων.

Η αξιολόγηση των μετρικών έγινε επίσης με "φυσικά πειράματα". Μία σημαντική εναλλακτική είναι η χρήση προσομοίωσης για τον ίδιο σκοπό. Συγκεκριμένα στα οικονομικά, στοχαστικές προσομοιώσεις μπορούν να χρησιμοποιηθούν για την παραγωγή τεχνητών δεδομένων με τη μέθοδο Monte Carlo. Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί για τη διερεύνηση ασυμπτωτικών συμπεριφορών μοντέλων για πεπερασμένα δείγματα ή ακόμη για την εύρεση κρίσιμων τιμών για στατιστικές μεθόδους (Cuthbertson, Nitzsche, 2004). Όπως δείχνουν οι εργασίες των Wang et al., 2012, Serre, Arcos, 2024, Dolphin et al., 2021 η διαδικασία αυτή δεν είναι διαδεδομένη στην πρακτική των χρονοσειρών, όπου προτιμάται η χρήση έτοιμων χρονοσειρών. Ένα ση-

μαντικό αποθετήριο τέτοιων χρονοσειρών είναι το UCR που χρησιμοποιήθηκε στην έρευνα των Wang κ.ά (2012). Σε άλλες περιπτώσεις όμως και για διερευνητικούς σκοπούς, έχουν χρησιμοποιηθεί τεχνητά δεδομένα για την αξιολόγηση νέων εργαλείων όπως η μέθοδος τοπολογικής ανάλυσης ballmapper (Dlotko et al., 2022). Οι ερευνητές δεν ανέφεραν τους λόγους για την επιλογή τους, αλλά είναι λογικό να αναδειχθούν εδώ τα μειονεκτήματα και πλεονεκτήματα της κάθε μεθόδου.

Η αξιολόγηση μπορεί να χαρακτηριστεί ως νατουραλιστική ή τεχνητή κατά τους Johannesson, Perjons (2014). Η πειραματική μέθοδος αποτελεί μία νατουραλιστική προσέγγιση που εξασφαλίζει ότι υπάρχει υψηλός βαθμός εσωτερικής εγκυρότητας αλλά περιορισμένη εξωτερική εγκυρότητα (Johannesson, Perjons, 2014), δηλαδή γενικευσιμότητα των αποτελεσμάτων. Ο χαρακτηρισμός αυτός ισχύει επίσης για την προσομοίωση, όπου υπάρχει μεγάλη εσωτερική εγκυρότητα, αλλά η χρήση του εργαλείου για ρεαλιστικές χρονοσειρές δεν έχει ελεγχθεί σε σημαντικό βαθμό.

3.1.3 Συλλογή δεδομένων και υπολογιστικά εργαλεία

Η υλοποίηση της έρευνας σε δεδομένα μεγάλης κλίμακας με τη χρήση σύνθετων στατιστικών τεχνικών προϋποθέτει τη χρήση των κατάλληλων εργαλείων. Η ανάλυση έκανε χρήση των εργαλείων Python, Matlab και Rstudio. Η χρήση πολλαπλών εργαλείων κρίθηκε απαραίτητη λόγω της συνθετότητας της ανάλυσης και των διαφορετικών εξειδικεύσεων των εργαλείων.

Το πρώτο εργαλείο που χρησιμοποιήθηκε για την επεξεργασία των δεδομένων είναι η γλώσσα προγραμματισμού python. Η γλώσσα αυτή είναι μια γενική γλώσσα προγραμματισμού, υψηλής αφάιρεσης και διαθέτει πολλές διαφορετικές δομές δεδομένων που χρησιμοποιήθηκαν στη διαδικασία, όπως πίνακες, λίστες και λεξικά. Αποτελεί ταυτόχρονα ευανάγνωστη γλώσσα και διαθέτει ευκολίες για την ανίχνευση σφαλμάτων (Python, 2024). Για το συγκεκριμένο πρότζεκτ, αξιοποιήθηκε στη διαμόρφωση και στην προεπεξεργασία του συνόλου δεδομένων, στην υλοποίηση συ-

ναρτήσεων, μετρικών, στην ομαδοποίηση, στην υλοποίηση των τυχαίων δασών καθώς και στην αξιολόγηση των αποτελεσμάτων.

Κρίσιμο ρόλο στην ανάλυση αυτή έπαιξαν τα πολυάριθμα πακέτα που διαθέτει η γλώσσα. Τα κυριότερα είναι τα `numpy`, `pandas`, `sklearn`. Εν συντομία, το `numpy` μάς επέτρεψε το μετασχηματισμό πινάκων ή τη μείωση μέσω του μέσου όρου και τη χρήση της άλγεβρας πινάκων, το πακέτο `pandas` βοήθησε στην προεπεξεργασία των δεδομένων και στο μετασχηματισμό τους στην κατάλληλη δομή δεδομένων. Το `scikit-learn` αποτελεί ένα πακέτο που προσφέρει δυνατότητες για υλοποίηση μεθόδων επιβλεπόμενης μάθησης, ανεπίβλεπτης μάθησης, αξιολόγηση μοντέλων καθώς και απεικόνισης αποτελεσμάτων. Χρησιμοποιήθηκαν οι μέθοδοι για τα τυχαία δάση, τη συσσωρευτική συσταδοποίηση με βάση τις μετρικές που υλοποιήσαμε καθώς και τα μέτρα αξιολόγησης όπως `Rand score`, `silhouette` για την αξιολόγηση των συσταδοποιήσεων.

Ένα δεύτερο σημαντικό εργαλείο για την ανάλυση μας αποτέλεσε το `MATLAB`. Το `MATLAB` αποτελεί ένα περιβάλλον προγραμματισμού και αριθμητικών υπολογισμών που διαθέτει πακέτα και εφαρμογές για πολλές επιστημονικές εξειδικεύσεις, συμπεριλαμβανομένων των οικονομικών. Η πλατφόρμα αυτή χρησιμοποιήθηκε λόγω της ευκολίας υλοποίησης ορισμένων προχωρημένων στατιστικών ελέγχων, όπως του `Dickey-Fuller` και του `Ljungberg-Q` για τον έλεγχο στασιμότητας χρονοσειρών. Παράλληλα, χρησιμοποιήθηκε για τον υπολογισμό των τελικών μοντέλων για την εκτίμηση της αιτιότητας κατά `Granger` ανά ζεύγη. Σημαντικό παράγοντα σε αυτή την επιλογή έπαιξε η υλοποίηση της συνάρτησης `getest` που επιτρέπει στον ερευνητή την επιλογή προσαρμοσμένου επιπέδου σημαντικότητας καθώς και η ευκολία εκτέλεσης επαναληπτικών ελέγχων.

Τέλος, χρησιμοποιήθηκε και η γλώσσα προγραμματισμού `R`. Αποτελεί κατά βάση ένα εργαλείο στατιστικής ανάλυσης με περιορισμένες γενικευμένες δυνατότητες (υλοποίηση εφαρμογών). Η χρήση του εργαλείου οφείλεται στη δυνατότητα επαναλη-

πτικού υπολογισμού μοντέλων Var τάξης έως δοθέντος βαθμού και η επιλογή της κατάλληλης τάξης με βάση τα κριτήρια AIC, BIC μεταξύ άλλων. Τα αποτελέσματα αυτής της ανάλυσης χρησιμοποιήθηκαν ως είσοδος στην ανάλυση μας με MATLAB.

Η χρήση πολλών διαφορετικών προγραμμάτων θέτει ζητήματα αλληλεπίδρασης των προγραμμάτων και μετασχηματισμού των εξόδων του κάθε προγράμματος ώστε να είναι επεξεργάσιμο από το πρόγραμμα που το απαιτεί. Η επιλογή της πλατφόρμας Anaconda για τη χρήση των python και R εξασφάλισε την ομαλή εισαγωγή των απαιτούμενων πακέτων καθώς και ένα πλαίσιο για την παράλληλη επεξεργασία προγραμμάτων. Τα αποτελέσματα του προγραμματισμού βρίσκονται σε 2 κύριες μορφές. Η πρώτη μορφή είναι αρχεία csv και η δεύτερη είναι η μορφή SPYdata, μία μορφή αρχείου ειδικά για υπολογιστικά περιβάλλοντα που χρησιμοποιούν γλώσσα python και όπου αποθηκεύτηκαν οι δομές δεδομένων με τα αποτελέσματα.

Η επιλογή των εργαλείων στην παρούσα εργασία δεν ήταν μονοσήμαντη. Για τις τεχνικές της συσταδοποίησης και των τυχαίων δασών θα μπορούσε να έχει χρησιμοποιηθεί η πλατφόρμα του MATLAB. Η επιλογή έγινε με βάση τις δυνατότητες προσαρμογής των μοντέλων από το πακέτο scikit-learn. Επίσης, ο στατιστικός έλεγχος της ισχύος της αιτιότητας ανά Granger θα μπορούσε να έχει γίνει εξ ολοκλήρου στην R. Η γλώσσα Matlab προσέφερε ευκολότερη διαχείριση του τελικού αντικειμένου υπολογισμού και τη δυνατότητα επιλογής για διαφορετικά επίπεδα σημαντικότητας.

3.1.4 Περιορισμοί της έρευνας και ηθικά ζητήματα

Η παρούσα εργασία αποτελεί μία διπλωματική εργασία για την εκπλήρωση των υποχρεώσεων μεταπτυχιακού προγράμματος και ως εκ τούτου υπόκειται στους περιορισμούς έρευνας μικρής κλίμακας (Denscombe, 2010). Η πρόσβαση στα δεδομένα δεν αποτέλεσε πρόβλημα στην υλοποίηση της έρευνας, από τη στιγμή που υπάρχει η ιστοσελίδα yahoo finance για τη δωρεάν πρόσβαση σε στοιχεία μετοχών που καλύπτουν δεκαετίες καθώς και η σελίδα του χρηματιστηρίου NASDAQ που προσφέρει στοιχε-

ία για τις μετοχές του της τελευταίας δεκαετίας (NASDAQ, 2024). Τα οικονομικά κόστη επίσης δεν αποτέλεσαν σημαντικό περιορισμό για την υλοποίηση της εργασίας επειδή η πρόσβαση στα δεδομένα ήταν δωρεάν και τα προγραμματιστικά εργαλεία ήταν είτε δωρεάν διαθέσιμα, είτε διαθέσιμα σε δοκιμαστική μορφή (Matlab). Τέλος, ο περιορισμός του χρόνου διαμέσου των προθεσμιών αποτελεί σημαντική παράμετρο σε κάθε πρότζεκτ (Denscombe, 2010) και η επιλογή του εύρους του πεδίου της έρευνας, της ερευνητικής στρατηγικής και της μεθόδου έγιναν λαμβάνοντας υπόψη τους εγγενείς χρονικούς περιορισμούς του πλαισίου της έρευνας.

Όσον αφορά την ηθική διάσταση της εργασίας, η έρευνα αυτή δε διαθέτει επιβλαβείς έμμεσες επιπτώσεις σε 3ους, σημαντική παράμετρος για την αξιολόγηση της ηθικότητας μιας εργασίας (Latham, 2022). Ο στόχος της εργασίας είναι η διερεύνηση της διασύνδεσης εννοιών σε θεωρητικό επίπεδο. Οι πρακτικές εφαρμογές υλοποιούνται ήδη σχετικά ικανοποιητικά χωρίς επιβλαβείς συνέπειες, συνεπώς ο κίνδυνος για ηθικές αστοχίες είναι αμελητέος.

3.2 Εφαρμογή Μεθοδολογίας

3.2.1 Προεπεξεργασία δεδομένων

Ελέγχουμε πρώτα από όλα για την ύπαρξη ελλειπουσών τιμών, ή για την ύπαρξη στιγμών στις οποίες από τα δεδομένα λείπουν. Ελέγχουμε επίσης για το αν όλες οι τιμές είναι θετικές. Έπειτα, μετασχηματίζουμε την κάθε χρονοσειρά Q μας με βάση τον τύπο $Y_i = \ln X_i + 1 - \ln X_i$ (Mantegna, 1999), (Mimkes et al., 2013). Υπολογίζει το λογάριθμο του λόγου των τιμών της μετοχής X μεταξύ των 2 χρονικών στιγμών και για μικρές διαφορές στις 2 τιμές αποτελεί ικανοποιητική προσέγγιση της σχετικής απόδοσης (Hudson, Gregoriou, 2010). Παράλληλα αποτελεί σημαντική βοήθεια για την εξάλειψη της τάσης σε κάθε χρονοσειρά (Neusser, 2016), ενώ επίσης μειώνονται οι δεξιές ουρές των κατανομών (Diebold, 2006). Η πρακτική αυτή είναι

συνήθης και σε άλλες εργασίες του τομέα των σύνθετων δικτύων όπου αναλύεται η ύπαρξη αιτιότητας κατά Granger ανάμεσα σε μετοχές. Ενδεικτικά αναφέρουμε τις εργασίες των (Vyrost et al., 2015) όπου διερευνάται η ύπαρξη αιτιότητας ανάμεσα σε 20 διεθνείς δείκτες αγορών, η εργασία των Παπάνα κ.ά που συγκρίνει την αιτιότητα με τη μέθοδο PMIME με βάση δεδομένα 21 χρηματαγορών καθώς και των Shirokikh et al. (2022) που επιχειρούν την ανάλυση σε επίπεδο μετοχών για χρηματιστήριο των ΗΠΑ. Μία διαφοροποίηση σε σχέση με την τελευταία εργασία είναι πως επιλέγουμε να χρησιμοποιήσουμε τις τιμές κλεισίματος των μετοχών και όχι τις προσαρμοσμένες τιμές κλεισίματος. Μία άλλη σημαντική παράμετρος των δεδομένων προς επεξεργασία είναι το μήκος των χρονοσειρών που θα χρησιμοποιηθούν στην ανάλυση. Πρέπει να τονιστεί πως η κατηγοριοποίηση με βάση τη μεθοδολογία του Mantegna (1999) είναι ευαίσθητη στη χρονική κλίμακα και στο συγκεκριμένο παράδειγμα μπορεί να έχουμε μεταβολές των συσχετίσεων σε κλίμακα χρόνων. Στην ανάλυση των Marti et al., (2021) για το ίδιο πρόβλημα, η επιλογή παραθύρου παραμένει αυθαίρετη στη βιβλιογραφία. Συχνά επιλέγεται με βάση τα χρονικά περιθώρια για επένδυση, καθώς και τη σταθερότητα των συστάδων (Marti et al., 2021). Παράλληλα, η επιλογή των παραθύρων θα πρέπει να γίνει έτσι ώστε να μπορούμε να έχουμε ανάλυση σε επίπεδο αιτιότητας Granger μεταξύ των χρονοσειρών.

Στις προηγούμενες εργασίες κατά κανόνα χρησιμοποιείται ένα ημερολογιακό έτος ημερών συναλλαγής, δηλαδή 250-252 μέρες ως το παράθυρο των χρονοσειρών. Παράλληλα, οι Παπάνα κ.ά. υλοποιούν την ανάλυση τους για αιτιότητα με παράθυρο 250 ημερών, όπως και στις εργασίες για δίκτυα με αιτιότητα κατά Granger. Μία υλοποίηση για έλεγχο αιτιότητας κατά Granger μεταξύ μετόχων του χρηματοπιστωτικού συστήματος πραγματοποιήθηκε από τους (Billio et al., 2010). Οι συγγραφείς επέλεξαν να λάβουν τιμές σε μηνιαίο επίπεδο, αλλά χρησιμοποίησαν στις αναλύσεις τους διαστήματα 7-8 ετών, γεγονός που υποδηλώνει πως ίσως μίας πιο κατάλληλη επιλογή είναι κοντά στις 100 παρατηρήσεις για κάθε χρονοσειρά. Στο έργο του Tang (2019)

έχουμε τη χρήση παραθύρου 50 ημερών. Συνεπώς, η επιλογή του παραθύρου των 250 ημερών είναι σύμφωνη με την έως τώρα πρακτική και δεδομένου των απαιτήσεων για στασιμότητα των υποκείμενων χρονοσειρών που διαθέτει η εφαρμογή της αιτιότητας Granger (Shojaie, 2022), δικαιολογείται η επιλογή παραθύρων μικρότερων του έτους.

Παράλληλα, για την αξιολόγηση των μέτρων ομοιότητας που υλοποιήθηκαν με μεθόδους visibility, χρησιμοποιούνται και 3 μη επικαλυπτόμενα παράθυρα 100 ημερών από 01-02-2021 έως 07-04-2021 καθώς και άλλα 3 μη επικαλυπτόμενα παράθυρα 30 ημερών από 08-04-2022 έως 18-08-2022. Οι επιλογές αυτές επιτρέπουν τη σύγκριση των μετρικών για διαφορετικά χρονικά διαστήματα και συνεισφέρουν στη μείωση της αβεβαιότητας του αποτελέσματος, ενώ ταυτόχρονα εξομαλύνουν την επίδραση ειδικών παραγόντων κατά τη διάρκεια συγκεκριμένου χρονικού διαστήματος.

Σε αυτό το σημείο, διαθέτουμε αρκετά στοιχεία για την υλοποίηση των μεθόδων τη συσχέτισης, της ανίχνευσης αιτιότητας κατά Granger ανά ζεύγη καθώς και την εύρεση των γράφων ορατότητας για κάθε μετοχή σε κάθε χρονικό διάστημα.

3.2.2 Τοπικά διανύσματα ορατότητας

Ορίζουμε για περιττό M και δοθέν γράφο ορατότητας V που προκύπτει από χρονοσειρά μήκους $M * N$ την ακόλουθη χρονοσειρά για $t_i, i = 1 \bmod M$ με μήκος N .

$$w(t_i) = \left(\sum_{x,y=0}^{x,y=M-1} V(t_{i+x}, t_{i+y}) - 3M + 2 \right) / (M^2 - 3M + 2)$$

Η λογική πίσω από τα τοπικά διανύσματα ορατότητας είναι πως μάς ενδιαφέρει η συμπεριφορά του γράφου ορατότητας συγκεκριμένα σε διάστημα M χρονικών στιγμών. Στην πράξη, απομονώνουμε μία περιοχή διάρκειας M χρονικών στιγμών και υπολογίζουμε το άθροισμα των ακμών του υπογράφου ορατότητας. Είναι σημαντικό εδώ να σημειωθεί ότι λόγω των ιδιοτήτων του γράφου, η υπολογισθείσα τιμή κανονικοποιείται. Δεν ορίζεται ακμή μεταξύ των δύο ίδιων χρονικών στιγμών, οπότε

M συνδέσεις είναι αδιευκρίνιστες στον υπολογισμό. Επίσης, λόγω του ορισμού του γράφου ορατότητας (Λασασα et al., 2008), έχουμε πως διαδοχικές χρονικές στιγμές πάντα συνδέονται μεταξύ τους τετριμμένα. Σε ένα σύνολο με M χρονικές στιγμές, υπάρχουν $M - 1$ μη διατεταγμένα ζεύγη. Εισάγονται έτσι στο άθροισμα $2 * (M - 1)$ μονάδες. Για να κανονικοποιήσουμε αυτή τη μετρική, αφαιρούμε το σύνολο των μονάδων που απαιτούνται από τον ορισμό του γραφήματος ορατότητας και καταλήγουμε με μία χρονοσειρά με τιμές που κυμαίνονται από 0 έως 1. Στην παρούσα υλοποίηση, αν το μέγεθος της αρχικής χρονοσειράς δεν είναι ακέραιο πολλαπλάσιο του παραθύρου M , αγνοούνται τελικές τιμές μέχρι να ισχύει η συνθήκη διαιρετότητας.

Ο ορισμός αυτός αποφεύγει προβλήματα όπως του κυλιόμενου μέσου, με χαρακτηριστικό το ότι κάθε τιμή της χρονοσειράς w επηρεάζεται από άλλες τιμές της χρονοσειράς μόνο έμμεσα, μέσω των τιμών του υποκείμενου γράφου ορατότητας V . Παράλληλα, αποτελεί ένα μέτρο της κυρτότητας της χρονοσειράς για εκείνη τη χρονική περίοδο. Αποτελεί ενδιαφέρον θέμα ανάλυσης να βρεθεί το κατά πόσο αυτή η έμμεση επιρροή αποτυπώνεται στις τιμές της χρονοσειράς w . Σε συνδυασμό με το γεγονός ότι χρηματοπιστωτικά προϊόντα διαθέτουν κατά μέσο όρο μηδενική αυτοσυσχέτιση (Marti et al., 2021) αναμένουμε η αυτοσυσχέτιση των χρονοσειρών αυτών να είναι κοντά στο 0.

Για την ανάλυση των χρονικών διαστημάτων που έχουμε επιλέξει (30, 100 και 250 ημέρες), χρησιμοποιούμε στη συνέχεια παράθυρα για τα τοπικά διανύσματα ορατότητας μήκους 5, 11 και 22 ημερών. Οι επιλογές αυτές είναι αυθαίρετες αλλά δικαιολογούνται εν μέρει από τη βιβλιογραφία. Πρώτα από όλα, οι Stephen et al. (2015) στην εργασία τους χτίζουν ένα δίκτυο από δίκτυα ορατότητας. Η βασική μονάδα των δικτύων ορατότητας είναι για παράθυρο 5 ή 6 στιγμών, οπότε αξίζει να συνεχιστεί η μελέτη του συγκεκριμένου παραθύρου. Πιο μικρά παράθυρα δεν περιέχουν αρκετή πληροφορία (Stephen et al., 2015). Βέβαια στην εργασία διατηρούνται όλα τα πιθανά μοτίβα του γράφου ορατότητας ενώ στη δική μας εργασία αυτή η

πληροφορία συμπυκνώνεται σε ένα μέτρο.

Παράλληλα, η επιλογή ανάμεσα σε αλληλοεπικαλυπτόμενα ή μη διαστήματα είναι ιδιαίτερα ακανθώδης, όπως φαίνεται και από το παράδειγμα της συσχέτισης. Υπάρχει μία σαφής επιλογή διότι αν το παράθυρο είναι αλληλεπικαλυπτόμενο και η διαφορά ανάμεσα σε κάθε παράθυρο μικρή, τότε κάθε νέο παράθυρο δεν προσφέρει πολλή πληροφορία (Sensoy, Tabak, 2014). Όμως, η χρήση μη αλληλεπικαλυπτόμενων παραθύρων οδηγεί στην υπερβολική μείωση των διαθέσιμων παραθύρων, αναγκάζοντας την εξέταση μεγάλων διαστημάτων. Λαμβάνοντας υπόψη αυτό το επιχείρημα στη δική μας εργασία, δεδομένου ότι ο υπολογισμός γράφων ορατότητας είναι ιδιαίτερα απαιτητικός υπολογιστικά, καταλήξαμε στη χρήση μη επικαλυπτόμενων παραθύρων για τη μείωση των απαιτούμενων υπολογισμών.

Η επιλογή για το μήκος του διαστήματος είναι λιγότερο αυθαίρετη. Οι περισσότερες μελέτες που έχουμε συναντήσει έχουν αριθμό παρατηρήσεων από 50 Tang 2019 έως ετήσια βάση (Birch, 2015), με την εργασία των (Billio et al., 2010) και τα εμπλεκόμενα μακροοικονομικά μεγέθη να δικαιολογούν την επιλογή του τριμήνου. Υπάρχουν και έρευνες που χρησιμοποιούν χρονοσειρές μετοχών για πολλά χρόνια (Onnela et al., 2003) αλλά όπως αναφέρουν και οι Mantegna και Stanley (2000), σε τέτοια διαστήματα μπορεί να έχουμε αλλαγή της σχέσης μεταξύ 2 μετοχών.

Σε σύγκριση με τις κύριες προσεγγίσεις στο γράφο ορατότητας, όπου κάθε στοιχείο/ χρονοσειρά μετατρέπεται σε ένα γράφο με κορυφές χρονικές στιγμές, η προσέγγισή μας διατηρεί ως βασική μονάδα τη χρονοσειρά, αφού λάβει υπόψη την τοπική πληροφορία του γράφου ορατότητας. Η κύρια προσέγγιση οδηγεί στο διαχωρισμό χρονοσειρών με βάση ασυμπτωτικές συμπεριφορές μεγεθών των γράφων, όπως για παράδειγμα για τη διάκριση μεταξύ τυχαίων και φράκταλ σειρών (Silva et al., 2021). Δεν μπορεί όμως να χρησιμοποιηθεί για το διαχωρισμό μεταξύ σειρών της ίδιας κατηγορίας. Το βασικό στοιχείο της προτεινόμενης προσέγγισης είναι πως διατηρούμε τη δυνατότητα ανάλυσης σε επίπεδο χρονοσειράς, ενσωματώνοντας την πληροφορία

του γράφου ορατότητας.

3.2.3 Συσταδοποίηση με Συσχέτιση

Υλοποιήθηκε ο αλγόριθμος που περιγράφεται στο 13ο κεφάλαιο του Mantegna 1999. Από τις 18 διαθέσιμες χρονοσειρές, σε 8 χρονικά διαστήματα, υπολογίστηκαν οι λογαριθμικές αποδόσεις με τον τύπο $R_i = \log P_i + 1 - \log P_{i-1} \in [2, L]$ όπου L το μήκος του παραθύρου. Έπειτα, υπολογίστηκε ο πίνακας των συσχετίσεων. Η συσχέτιση μεταξύ κάθε ζεύγους χρονοσειρών x, y δίνεται από:

$$r_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 * \sum (y_i - \bar{y})^2}}$$

Έπειτα, χρησιμοποιήθηκε ο πίνακας των συσχετίσεων για να βρεθεί ένα μέτρο απόστασης με το μετασχηματισμό $d_{x,y} = \sqrt{2 - 2 * r_{x,y}}$. Στη συνέχεια υλοποιήθηκε ένας ιεραρχικός συσσωρευτικός αλγόριθμος συσταδοποίησης, με κριτήριο την υπολογισμένη απόσταση και πλήρη δεσμό. Η επιλογή αυτή έγινε λόγω του μικρού αριθμού χρονοσειρών και του κινδύνου για αλυσίδωση των συστάδων (Marti, 2021), σε αντίθεση με την υλοποίηση στο Mantegna (1999). Επίσης, δεν υλοποιείται το ελάχιστο δέντρο από τη στιγμή που δίνεται έμφαση στη συσταδοποίηση.

Μερικά από τα πλεονεκτήματα της μεθόδου αποτελεί η απλότητα της καθώς και η γρήγορη εποπτεία των αποτελεσμάτων. Επίσης, η μέθοδος σύμφωνα με αποτελέσματα δίνει διαισθητικά ικανοποιητικές απαντήσεις, δημιουργώντας ομάδες που αντιστοιχούν σε εμπορικούς τομείς (Mantegna, 1998), (Onnela et al., 2003). Η μεθοδολογία αυτή όμως διαθέτει και πολλά προβλήματα, τα κυριότερα από τα οποία είναι πιθανώς η ευαισθησία του αλγορίθμου συσταδοποίησης σε μικρές αλλαγές των δεδομένων καθώς και η άγνοια θεωρητικών αποτελεσμάτων για τις στατιστικές κατανομές της συσταδοποίησης (Marti, 2021). Η αστάθεια των αποτελεσμάτων μπορεί να οφείλεται εν μέρει στη συνάρτηση εγγύτητας 2 συνόλων που χρησιμοποιείται για την ομαδοπο-

ίση ή στη φύση της συσχέτισης κατά Pearson (Marti,1999). Θα πρέπει να σημειωθεί πως η μεθοδολογία μας διαφοροποιείται σε αυτά τα 2 σημεία.

3.2.4 Δημιουργία πίνακα αποστάσεων γράφων ορατότητας

Υλοποιούμε με βάση τα μέτρα SMC και Jaccard. Συγκεκριμένα, για κάθε χρονοσειρά υπολογίζουμε το γράφο ορατότητας της για κάθε ένα από τα 8 εξεταζόμενα χρονικά διαστήματα. Στη συνέχεια, εφαρμόζουμε τις υλοποιημένες συναρτήσεις μας σε κάθε γράφο ορατότητας και διατάσσουμε τα αποτελέσματα σε 2 πίνακες, έναν πίνακα με ομοιότητα κατά SMC και κατά Jaccard.

Για την υλοποίηση των γράφων ορατότητας υπάρχουν αρκετοί τρόποι, που αναλύονται στην εργασία του (Varela, 2020). Εμείς υλοποιήσαμε τον πιο απλό, τον 'αφελή' τρόπο που έχει θεωρητικό όριο συμπεριφοράς $O(n^3)$. Υπάρχει επίσης ένας άλλος αλγόριθμος που εξετάσει κάθε ζεύγος χρονικών στιγμών και δίνει πολυπλοκότητα $O(n^2)$ καθώς και ο αλγόριθμος διαίρει και βασίλευε των Lan et al. (2017) με πολυπλοκότητα $O(n \log(n))$.

Στη σύντομη αναφορά μας τα μέτρα ορίστηκαν για 2 διανύσματα ίσου μήκους, αλλά η γενίκευση για πίνακες ίδιων διαστάσεων είναι απλή. Κάθε θέση σημείου του γράφου ορατότητας καθορίζεται από μία διατεταγμένη δυάδα χρονικών στιγμών t_i, t_j . Ανάλογα με την παρουσίαση στο Βλαχάβας κ.ά., 2020, για 2 πίνακες γειτνίασης N_1, N_2 γράφων ορατότητας υπάρχουν 3 ενδεχόμενα για κάθε ζεύγος στιγμών t, t_j . Είτε οι τιμές γειτνίασης διαφέρουν, είτε οι δύο τιμές είναι ίσες με 0, είτε οι δύο τιμές είναι ίσες με 1. Αντίστοιχα ορίζουμε τον αριθμό των διατεταγμένων χρονικών ζευγών για τις οποίες ισχύει η κάθε συνθήκη ως N_{01}, N_{00}, N_{11} . Ορίζουμε λοιπόν το δείκτη Jaccard για 2 γράφους ορατότητας στο χρονοσειρών που ορίζονται στο ίδιο χρονικό διάστημα. Είναι ίσος με τον αριθμό των χρονικών δυάδων (υποσύνολο του (TxT) όπου T το σύνολο των χρονικών στιγμών που ορίζεται η χρονοσειρά) που και τα 2 στοιχεία των πινάκων γειτνίασης είναι ίσα με 1 για το δοθέν ζεύγος στιγμών προς

τον αριθμό των ζευγών που έστω ένα από τα 2 στοιχεία των 2 πινάκων γειτνίασης είναι ίσο με 1. Αριθμητικά έχουμε:

$$V_{Jacc} = \frac{N_{11}}{N_{11} + N_{01}}$$

. Με αντίστοιχη διαδικασία για καταλήγουμε και στον ορισμό του SMC για 2 γράφους ορατότητας,

$$V_{SMC} = \frac{N_{11} + N_{00}}{N_{11} + N_{01} + N_{00}}$$

3.2.5 Ομαδοποίηση με χρήση των μέτρων

Στη συνέχεια υλοποιήθηκε η ομαδοποίηση με βάση τα μέτρα απόστασης μεταξύ 2 γράφων ορατότητας. Οι δύο πίνακες απόστασης χρησιμοποιήθηκαν από συσσωρευτικό αλγόριθμο συσταδοποίησης με πλήρη δεσμό απόστασης για 2 σύνολα και εξήχθησαν έτσι 16 συσταδοποιήσεις για τα 8 χρονικά διαστήματα, 8 για την ομοιότητα κατά SMC, 8 για την ομοιότητα κατά Jaccard.

3.2.6 Σύγκριση ομαδοποιήσεων

Ως πρώτο βήμα συγκρίνουμε κάθε συσταδοποίηση με βάση το σχήμα επιλογής των μετοχών μας. Δηλαδή 3 συστάδες των 6 μετοχών. Το σχήμα αξιολόγησης αποτελείται από 3 διαφορετικές κατηγορίες συσταδοποιήσεων, που οφείλονται σε διαφορετικούς πίνακες συσταδοποίησης, δηλαδή διαφορετικές μετρικές απόστασης. Αρχικά, απαιτείται ένα μέτρο της εγγύτητας των συσταδοποιήσεων. Συγκρίνονται όλες οι συσταδοποιήσεις με βάση την τελευταία περίοδο 250 ημερών της κάθε κατηγορίας μετρικές. Ο μέσος όρος του δείκτη Ρανδ για όλες τις χρονικές περιόδους δείχνει κατά πόσο προβλέπουν την ίδια δομή για κάθε χρονική στιγμή. Στη συνέχεια, υλοποιούνται 16 συγκρίσεις Rand με βάση την υποκείμενη συσταδοποίηση. Αυτή η προσέγγιση δίνει ένα μέτρο του κατά πόσο αντιστοιχούν οι ομαδοποιήσεις στο στοχευμένο διαχωρι-

σμό των μετοχών. Στη συνέχεια τα μέτρα αυτά ομαδοποιούνται ανά χρονική περίοδο (2 χρονοσειρές μήκους 252, 3 χρονοσειρές μήκους 100, 3 χρονοσειρές μήκους 30) και αναλύονται οι διαφορές στα ποσοστά λήψης σωστών αποφάσεων με βάση τις 3 μετρικές.

Επίσης, χρησιμοποιείται το εσωτερικό κριτήριο αξιολόγησης συσταδοποίησης του δείκτη σκιαγράφησης. Ως αδιάστατο μέγεθος, αποτελεί μέτρο σύγκρισης για συσταδοποίηση με διαφορετικές μετρικές. Επίσης δεν απαιτεί κανονικοποίηση των δεδομένων, πολλαπλασιασμός των χαρακτηριστικών (και συνεπώς των αποστάσεων των δειγμάτων) επί ένα σταθερό αριθμό δεν αλλάζει το λόγο. Υψηλοί δείκτες κοντά στο 1 είναι θεμιτοί και δείχνουν καλή συσταδοποίηση (Βλαχάβας κ.ά., 2020). Βρίσκουμε το μέσο όρο του δείκτη για τις συσταδοποιήσεις ανά μετρική απόστασης και χρονικό διάστημα και αξιολογούμε την αποτελεσματικότητα των μετρικών ως προς αυτό τον παράγοντα.

Τέλος, έχει ενδιαφέρον για τους σκοπούς της προβλεπτικής ικανότητας των μετρικών η αξιολόγηση της εγγύτητας των συσταδοποιήσεων με βάση την ομοιότητα κατά SMC και την ομοιότητα κατά Jaccard. Υλοποιούνται 8 συγκρίσεις με βάση το δείκτη Rand ανά ζεύγη για τις 2 συσταδοποιήσεις, μία για κάθε χρονική περίοδο.

3.2.7 Μετρική ομοιότητας ομαλής ορατότητας

Η υλοποίηση της μεθόδου γίνεται με έναν αλγόριθμο που υπολογιστικά απαιτεί χρόνο $O(L^3)$, οπότε είναι σχετικά αργή. Στη βιβλιογραφία, περιγράφονται πιο περίπλοκες υλοποιήσεις όπου η αποθήκευση ήδη υπολογισμένων τιμών οδηγεί σε εξάρτηση $O(L * \log l)$ (Lan, 2015). Είναι σημαντικό να σημειωθεί εδώ ότι η από την υλοποίηση για κάθε παράμετρο ομαλότητας M , δημιουργείται ένας διακριτός δείκτης. Σύμφωνα με τους υπολογισμούς, ένας δείκτης ομαλής ορατότητας με $m = 3$ λαμβάνει υπόψη μόνο μία τιμή, με στάθμες στο 0 και στο 1. Λόγω του μειωμένου μεγέθους των χρονοσειρών που προκύπτουν από την εφαρμογή, οι τεχνική αυτή εφαρμόζεται για

$m = 5$, $m = 11$ και $m = 22$ στις δύο χρονικές περιόδους διάρκειας 1 έτους. Προκύπτουν έτσι 3 χρονοσειρές για κάθε χρονοσειρά ανά περίοδο, με μήκος 50, 22, 11. Από τις μετρικές αυτές μπορούμε να υλοποιήσουμε μία μετρική ομοιότητας. Στη συγκεκριμένη περίπτωση, επιλέγουμε την ευκλείδια απόσταση μεταξύ των 2 παραγόμενων χρονοσειρών ομαλής ορατότητας ως μέτρο απόστασης των 2 υποκείμενων χρονοσειρών. Με βάση αυτό το μέτρο, υλοποιούμε 3 πίνακες αποστάσεων και εκτελούμε συσταδοποίηση ως έλεγχο της διαδικασίας μας. Παράλληλα έχουμε δημιουργήσει και τις εισόδους μας για τον αλγόριθμο των τυχαίων δασών.

3.2.8 Εκτίμηση αιτιότητας

Η ανάλυση αιτιότητας κατά Granger προϋποθέτει την εύρεση του σωστού αριθμού όρων υστέρησης (λαγκ) (Shojaie, Fox, 2022). Με τη χρήση του πακέτου statsmodel, έγινε η εκτίμηση του βέλτιστου όρου αυτού. Έπειτα, επειδή στις αντίστοιχες εφαρμογές του προγράμματος python δεν υπήρχε η δυνατότητα αυτοματοποιημένης ανάκτησης των δεδομένων, υλοποιήσαμε τον έλεγχο για αιτιότητα με την εντολή gctest και την εισαγωγή του αριθμού των υστερήσεων. Η εντολή αυτή υπολογίζει μία στατιστική τιμή ελέγχου με τη μέθοδο OLS όπως περιγράφηκε στην ανάλυση για τα μοντέλα VAR. Οι Shojaie, Fox (2022) προτείνουν ως μεθοδολογικό πλαίσιο τον έλεγχο αιτιότητας μεταξύ 2 μεταβλητών ως ενδογενείς μεταβλητές του δικτύου, χωρίς τη χρήση εξωγενών παραγόντων στην περίπτωση ενός συνόλου μετοχών. Για περιορισμό των όρων υστέρησης, προτείνουν τη χρήση μίας ποινής Lasso που οδηγεί αρκετούς όρους υστέρησης να εκτιμώνται ως μηδενικοί.

Λόγω του πλήθους των ελέγχων, χρησιμοποιούμε μία συντηρητική προσέγγιση για την αποτίμηση της στατιστικής σημαντικότητας των αποτελεσμάτων μας, τη μέθοδο Bonferroni. Κατά βάση θέτουμε ως νέο οριακή τιμή το επίπεδο σημαντικότητας διαιρεμένο κατά τον αριθμό των ελέγχων. Η μεταβολή αυτή εξασφαλίζει ότι η πιθανότητα σφάλματος τύπου α για έστω έναν από τους πολλαπλούς ελέγχους παραμένει

μικρότερη από το αρχικό επίπεδο σημαντικότητας (Birch et al., 2015), (Shorokikh et al., 2022) για πολλαπλούς ελέγχους Granger αιτιότητας.

3.2.9 Δέντρα αποφάσεων και τυχαίο Δάσος

Τέλος, δημιουργήθηκαν μοντέλα τυχαίων δασών για τον έλεγχο της επιρροής των μετρικών μας στη συσταδοποίησης και στην αιτιότητα κατά Granger των λογαριθμικών επιστροφών των μετοχών. Εκτελέστηκαν 5 διαφορετικά τυχαία δάση, με αριθμό εκτιμητών $n = 1000$ και μέθοδο υπολογισμού κέρδους το δείκτη gini. Τα πρώτα 3 τυχαία δάση αποτελούν έναν έλεγχο των διαδικασιών συσταδοποίησης. Οι προηγούμενες διαδικασίες οδήγησαν στον υπολογισμό 10 διαφορετικών μετρικών ομοιότητας μεταξύ χρονοσειρών. Οι μετρικές αυτές είναι της απόστασης κατά SMC των γράφων ορατότητας, της απόστασης κατά Jaccard των γράφων ορατότητας καθώς και τη μέση ευκλείδεια απόσταση των χρονοσειρών ομαλής ορατότητας για $m = 5, 11, 22$, κάθε μέγεθος υπολογισμένο για τις 2 χρονοσειρές μήκους 1 έτους.

Υλοποιήθηκε ένας πίνακας που διαθέτει τις μετρικές για κάθε ζεύγος μετοχών καθώς και ένα δυαδικό διάνυσμα με τιμή 1 για τα ζεύγη μετοχών που ανήκουν στον ίδιο κλάδο, το διάνυσμα στόχων συσταδοποίησης. Ο πίνακας έχει δείκτη από το 1 έως το 324, που αντιπροσωπεύουν τα διατεταγμένα ζεύγη μετοχών. Χρησιμοποιήθηκε επίσης ένα δυαδικό διάνυσμα όπου για κάθε ζεύγος μετοχών x, y , η τιμή του διανύσματος είναι 1 αν ανιχνεύτηκε αιτιότητα κατά Granger από τη μετοχή x στη μετοχή y . Συνεπώς, ο αλγόριθμος των τυχαίων δασών χρησιμοποιείται σε αυτή την περίπτωση για να διαχωρίσει το χώρο των ζευγών μετοχών σε περιοχές όπου οι μετοχές συνδέονται μέσω αιτιότητας κατά Granger ή μη. Για κάθε πείραμα, αξιοποιήθηκε το 80% των παρατηρήσεων (259) για την εκπαίδευση του μοντέλου και το 20% (65) για τον έλεγχο. Σε κάθε τυχαίο δάσος, ο αριθμός των βασικών ταξινομητών ορίστηκε στα 1000 δέντρα και σύμφωνα με την εκτίμηση των Βλαχάβα κ.ά., (2020) τα χαρακτηριστικά που λήφθηκαν υπόψη σε κάθε βήμα είναι 2 (αν ο αριθμός των

χαρακτηριστικών είναι μικρότερος του 9) ή 3 (στα 10 χαρακτηριστικά). Η επιλογή του κριτηρίου ανομοιογένειας έγινε με βάση το μειωμένο υπολογιστικό κόστος και επιλέχθηκε το κόστος Gini.

Η πρώτη οικογένεια ελέγχων αφορά τη συσταδοποίηση με βάση τις μετρικές. Με βάση το διάνυσμα στόχων της συσταδοποίησης διερευνήθηκαν 3 μοντέλα τυχαίων δασών με διαφορετικά δεδομένα. Στην πρώτη περίπτωση, χρησιμοποιήθηκαν οι 4 δείκτες μόνο για τα μέτρα ομοιότητας από τους γράφους ορατότητας, για τη δεύτερη περίπτωση χρησιμοποιήθηκαν μόνο οι 6 μετρικές για τα μέτρα ομοιότητας με βάση την ομαλή ορατότητα και τέλος συνδυάστηκαν και τα 10 μέτρα.

Στη δεύτερη οικογένεια ελέγχων ανήκουν οι έλεγχοι με διάνυσμα στόχων το διάνυσμα των αιτιοτήτων ανά ζεύγος. Υλοποιήθηκαν 2 έλεγχοι, στον πρώτο χρησιμοποιήθηκαν οι 5 μετρικές της πρώτης περιόδου για την πρόβλεψη της αιτιότητας κατά Granger στην πρώτη περίοδο. Ο δεύτερος έλεγχος έγινε με την ίδια διαδικασία, αλλά αφορούσε στοιχεία της 2ης περιόδου.

Η επιτυχία των μοντέλων κρίθηκε με βάση το σκορ του αλγορίθμου, το ποσοστό των σωστών ταξινομήσεων. Συγκεκριμένα χρησιμοποιήθηκε ο μέσος όρος του σκορ. Υπολογίζονται επίσης οι σημαντικότητες των χαρακτηριστικών σε κάθε περίπτωση για την αξιολόγηση της επιρροής της κάθε μεταβλητής στο αποτέλεσμα.

4 Αποτελέσματα

4.1 Αποτελέσματα Ταξινόμησης με Ιεραρχική Συσταδοποίηση

Ακολουθούν τα αποτελέσματα της συσταδοποίησης με βάση τη συσχέτιση. Πρώτα από όλα, παρατηρούμε ότι υπάρχουν διαφορές στη συσταδοποίηση ανά το χρόνο. Με σημείο αναφοράς την Περίοδο 5, έχουμε πως οι αποφάσεις συσταδοποίησης στην περίοδο 0 είναι ίδιες μόνο από 50.3% έως 58.8%. Συνεπώς υπάρχει σημαντική διαφοροποίηση στη δυναμική των συστάδων χρονικά. Παραθέτουμε τους πίνακες με τα

αποτελέσματα της σύγκρισης μετρικών ανά κάθε χρονικό διάστημα. Παρατηρούμε πως κατά μέσο όρο, η συσχέτιση αποδίδει καλύτερα. Ανά χρονικές περιόδους, σε 4 περιόδους υπερτερεί η συσχέτιση, σε 2 περιόδους υπερτερεί η μετρική Jaccard, σε 1 περίοδο ισοβαθμούν η συσχέτιση με τη μετρική Jaccard και σε 1 περίοδο υπερτερεί το κριτήριο SMC. Παίρνοντας το μέσο όρο των περιόδων με βάση τη διάρκεια της περιόδου βλέπουμε ότι το μέγιστο σκορ επιτυγχάνεται για διαστήματα 252 ημερών με τη μέθοδο της συσχέτισης. Και για τις 3 διάρκειες, το υψηλότερο σκορ το διαθέτει η μετρική της συσχέτισης, με το επόμενο υψηλότερο σκορ να επιτυγχάνεται με τη χρήση της μετρικής SMC. Παρατηρούμε ωστόσο ότι οι μετρικές βασισμένες στους γράφους ορατότητας αποδίδουν καλύτερα από τη συσχέτιση στην περίοδο 0.

	Jaccard	SMC	Correlation
Period_0	0.863	0.732	0.667
Period_5	0.497	0.471	0.869
Period_1	0.693	0.458	0.732
Period_2	0.627	0.392	0.771
Period_3	0.771	0.614	0.771
Period_41	0.732	0.614	0.667
Period_42	0.588	0.693	0.667
Period_43	0.627	0.627	0.654

Σχήμα 1: Πίνακας 1. Σκορ Rand της κάθε μεθόδου ανά περίοδο με βάση την κατανομή σε τομείς

Παραθέτουμε επίσης και τους δείκτες σχιαγράφησης στον πίνακα 3. Οι δείκτες σχιαγράφησης κατά μέσο όρο σε όλες τις περιόδους είναι ίσοι με 0.083 για το δείκτη Jaccard, 0.091 για το δείκτη SMC και 0.177 για τη συσχέτιση, παίρνοντας το μέσο όρο σε όλες τις περιόδους.

	Jaccard	SMC	Correlation
252-days	0.680	0.601	0.768
100-days	0.697	0.488	0.758
30-days	0.649	0.645	0.662

Σχήμα 2: Πίνακας 2. Μέσος όρος σκορ Rand της κάθε μεθόδου ανά διάρκεια περιόδου με βάση την κατανομή σε τομείς

	Jaccard	SMC	Correlation
Period_0	0.069	0.096	0.153
Period_5	0.070	0.107	0.209
Period_1	0.086	0.113	0.161
Period_2	0.069	0.070	0.173
Period_3	0.095	0.078	0.207
Period_41	0.124	0.129	0.147
Period_42	0.062	0.057	0.169
Period_43	0.091	0.083	0.193

Σχήμα 3: Πίνακας 3. Δείκτες σκιαγράφησης ανά μέθοδο και περίοδο

4.2 Αποτελέσματα Συσταδοποίησης με μεθόδους Τυχαίου Δάσους

Παρουσιάζουμε τα αποτελέσματα για τα μοντέλα με τη μέθοδο των δασών, για την πρόβλεψη του αν ένα ζεύγος μετοχών σχετίζονται αιτιακά κατά Granger με βάση τις μετρικές που εξετάζουμε. Το πρώτο μοντέλο που εξετάζουμε είναι το μοντέλο για την επιρροή των 2 μετρικών βασισμένες στις μεθόδους γράφων ορατότητας, όπως μετρήθηκε για την περίοδο 0. Η περίοδος 5 λόγω της αυστηρότητας του κριτηρίου Bonferroni είχε μόνο 7 δείγματα αιτιότητας κατά Granger μεταξύ μετοχών και η κλάση αυτή υποαντιπροσωπευόταν. Δηλαδή χρησιμοποιήθηκαν 2 μεταβλητές. Στη συγκεκριμένη περίπτωση, έχουμε πως η ακρίβεια του ταξινομητή είναι ίση με 0.831. Οι τιμές σημαντικότητας των μεταβλητών είναι ως ακολούθως, 0.479 για το δείκτη

1η μέθοδος ▾	2η μέθοδος ▾	Σκορ ▾
Jaccard	SMC	0.877
Jaccard	m_5	0.831
Jaccard	m_11	0.769
Jaccard	m_22	0.831
SMC	m_5	0.815
SMC	m_11	0.815
SMC	m_22	0.785
m_5	m_11	0.707
m_5	m_22	0.831
m_11	m_22	0.769

Σχήμα 4: Πίνακας 4. Σκορ όλων των μοντέλων ανά 2 μεταβλητές

Jaccard και 0.521 για το δείκτη m_5 .

Στη συνέχεια, εισήγαμε περισσότερους συνδυασμούς μεταβλητών. Ελέγξαμε κάθε συνδυασμό των μεταβλητών ανά 2 και εξετάσαμε την προβλεπτική τους ικανότητα. Στο δείγμα ελέγχου, ο συνδυασμός των 2 μετρικών Jaccard και SMC με σκορ 0.877. 3 από τους 4 συνδυασμούς με την καλύτερη επίδοση περιείχαν τη μετρική Jaccard. Αντίστοιχα, 3 από τους συνδυασμούς με σκορ κάτω από 0.8 περιείχαν τη μεταβλητή m_{11} . Η χαμηλότερη επίδοση επιτεύχθηκε από το συνδυασμό των 2 μετρικών m_{11} και m_5 .

Το πρώτο μοντέλο με περισσότερες μεταβλητές που εξετάζεται περιλαμβάνει τις 3 μεταβλητές που ορίστηκαν για ζεύγη μετοχών με βάση την ευκλείδεια απόσταση της ομαλής ορατότητας τους. Το μοντέλο έχει ακρίβεια 0.665. Παρατίθενται στον πίνακα 1.4 τα αποτελέσματα για τις επιμέρους σημαντικότητες των αποτελεσμάτων. Οι σημαντικότητες είναι πολύ κοντά, με πιο σημαντικό χαρακτηριστικό το δείκτη SMC με μέσο βάρος στις 10 επαναλήψεις των δοκιμών ίσο με 0.345. Το 3ο μοντέλο που εξετάζεται περιλαμβάνει και τις 5 προαναφερθείσες μεταβλητές και έχει αυξημένη ακρίβεια, κατά μικρό βαθμό ωστόσο, με ακρίβεια 0.831. Περιλαμβάνου-

	Jaccard	SMC	m_5
Run 1	0.325	0.333	0.343
Run 2	0.36	0.34	0.3
Run 3	0.342	0.352	0.306
Run 4	0.356	0.343	0.302
Run 5	0.338	0.353	0.309
Run 6	0.331	0.339	0.33
Run 7	0.351	0.332	0.318
Run 8	0.334	0.359	0.307
Run 9	0.334	0.344	0.321
Run 10	0.311	0.359	0.331
Mean	0.3382	0.3454	0.3167
Variance	0.015	0.010	0.014

Σχήμα 5: Πίνακας 5. Σημαντικότητες χαρακτηριστικών στο μοντέλο 2, με 6 μεταβλητές

με τις σημαντικότητες των μεταβλητών στον πίνακα 5. Σε αυτή την περίπτωση οι σημαντικές μετρικές έρχονται από την περίοδο 0 και είναι κατά σειρά η $m_5=0.232$, Jaccard=0.2134. Οι υπόλοιπες 3 μεταβλητές έχουν σημαντικότητα γύρω από 0.185, κάτω από το 1/5. Τη μεγαλύτερη διακύμανση στη σημαντικότητα των γνωρισμάτων παρουσιάζει ο δείκτης Jaccard με τυπική απόκλιση 0.035.

4.3 Ανάλυση ομαλών χρονοσειρών ορατότητας με βάση αυτοπαλινδρόμηση

Οι χρονοσειρές ομαλής ορατότητας των περιόδων 0 και 5 αναλύθηκαν για την ύπαρξη αυτοπαλινδρόμησης με το τεστ Ljung-Box Test για την ύπαρξη στατιστικά σημαντικού συντελεστή υστέρησης βαθμού 1. Παρατηρήσαμε πως στην περίοδο 0, μόνο οι MRNA, WBA έχουν στατιστικά σημαντικές ενδείξεις ενάντι στη μηδενική υπόθεση, με καμία τιμή της στατιστικής του τεστ να μην είναι κάτω από το όριο του επιπέδου αξιοπιστίας 0.1/18 όπου προσαρμόζεται με βάση τον κανόνα του Bonferroni. Η αντίστοιχη ανάλυση για την περίοδο 5 καταδεικνύει ότι πάλι δεν υπάρχει αυτο-

	Jaccard	SMC	m_5	m_11	m_22
Run 1	0.242	0.166	0.203	0.203	0.186
Run 2	0.248	0.212	0.211	0.157	0.172
Run 3	0.188	0.21	0.233	0.19	0.179
Run 4	0.257	0.204	0.226	0.163	0.15
Run 5	0.171	0.185	0.225	0.204	0.215
Run 6	0.22	0.171	0.262	0.179	0.168
Run 7	0.192	0.147	0.229	0.191	0.241
Run 8	0.254	0.173	0.206	0.173	0.193
Run 9	0.197	0.191	0.261	0.181	0.171
Run 10	0.165	0.186	0.264	0.203	0.183
Mean	0.2134	0.1845	0.232	0.1844	0.1858
Variance	0.035	0.021	0.023	0.017	0.026

Σχήμα 6: Πίνακας 6. Σημαντικότητες χαρακτηριστικών στο 3ο μοντέλο, με 5 μεταβλητές

παλινδρόμηση για κάθε μετοχή, αλλά σε σημαντικότερο βαθμό. Καμία μετοχή δεν ξεπερνά το επίπεδο σημαντικότητας 10%, με όλες τις χρονοσειρές να είναι συνεπείς με τη μηδενική υπόθεση της μη αυτοπαλινδρόμησης.

4.4 Αποτελέσματα αιτιότητας κατά Granger

Τα αποτελέσματα της ανάλυσης με μοντέλα VAR έδωσε αποτελέσματα για τις τάξεις των μοντέλων από 1 έως 4. Για ορισμένες μετοχές όπως αυτή της KHC ο καλύτερος προσδιορισμός τάξης είναι 4 για 15 από τις 17 διαθέσιμες επιλογές για τη δημιουργία ζεύγους και για μετοχές όπως η MRNA η πλειοψηφία αποτελεί μοντέλα με υστέρηση 1. Οι υστερήσεις των μοντέλων είναι κατανομημένες ως εξής για την 1η περίοδο (επιλέχθηκαν με βάση το BIC). $p = 1, 96$, $p = 2, 44$, $p = 3, 66$, $p = 4, 100$.

Η αντίστοιχη ανάλυση για την περίοδο 5 δίνει διαφορετικά αποτελέσματα, με την υστέρηση 1 και 2 να έχουν τη συντριπτική πλειοψηφία των ζευγών ($p = 1, 200$, $p = 2, 54$, $p = 3, 40$, $p = 4, 12$). Συνολικά, η διάμεση υστέρηση στην περίοδο 0 είναι 3 ενώ στην περίοδο 5 είναι 1, δείχνοντας μία σημαντική διαφορά στην κατανομή μεταξύ

των 2 περιόδων.

4.5 Πρόβλεψη αιτιότητας κατά Granger

Παραθέτουμε πρώτα το πλήθος των στατιστικά σημαντικών ελέγχων κατά Granger στις περιόδους 0 και 5 όπου έχουμε 42 και 7 ζεύγη αντίστοιχα. Τα δείγματα με σχέση αιτιότητας για την περίοδο 5 είναι ελάχιστα και η ακρίβεια του μοντέλου μεγάλη, 0.955 λόγω της ανισορροπίας του δείγματος. Ο έλεγχος που έγινε με τις 5 μεταβλητές για την περίοδο 0 μάς έδωσε ακρίβεια 0.809. Τα γνωρίσματα των μοντέλων κατά μέσο όρο είχαν σημασία 0.2, με τα πιο σημαντικά να είναι οι δείκτες Jaccard και τα τοπικά διανύσματα ορατότητας με βάρος 0.213 και 0.232. Παρατίθενται στον πίνακα 7 αναλυτικά.

Jaccard	0.213
SMC	0.184
m_5	0.232
m_11	0.184
m_22	0.186

Σχήμα 7: Πίνακας 7. Σημαντικότητες χαρακτηριστικών στα μοντέλα πρόβλεψης της αιτιότητας Granger

5 Συζήτηση

Προχωρούμε στο σχολιασμό των αποτελεσμάτων. Σε πρώτο βαθμό, αναλύουμε τα αποτελέσματα για την ιεραρχική συσταδοποίηση με τις 3 μετρικές, κατά Jaccard, κατά SMC και κατά συσχέτιση. Παρατηρούμε ότι το σκορ Rand είναι πιο ψηλό κατά μέσο όρο για τις συσχετίσεις, έπεται ο δείκτης Jaccard και έπεται ο δείκτης SMC

κατά μέσο όρο για τις δοκιμές στα 3 διαστήματα. Ο δείκτης SMC διαθέτει μερικές επιδόσεις που είναι πολύ χαμηλές όπως στην περίοδο 2, περίοδο 1 και περίοδο 5. Οι χαμηλότερες βαθμολογίες της συσχέτισης ανά περίοδο είναι μεγαλύτερη από τις χαμηλότερες των άλλων δύο δεικτών, καταδεικνύοντας ότι η συσταδοποίηση με βάση τη συσχέτιση συμφωνεί περισσότερο με τη συσταδοποίηση με βάση τον τομέα της μετοχής.

Έχουμε παράλληλα πως η αξιοπιστία του δείκτη της συσχέτισης με βάση περίοδο φθίνει, ενώ για το δείκτη Jaccard φαίνεται να είναι σταθερό (0.65-0.7). Ο δείκτης SMC αποδίδει πολύ χαμηλά για το διάστημα των περιόδων 100 ημερών. Η επίδοση των 3 δεικτών είναι πολύ κοντά στις 30 ημέρες. Συνεπώς, μπορούμε να πούμε ότι η επίδοση των δεικτών μας δεν είναι τόσο ευαίσθητη στο μήκος του χρονικού παραθύρου όσο η μέθοδος της συσχέτισης, ή πρέπει επιπρόσθετα να ελεγχθούν παράθυρα μεγαλύτερα του 1 έτους.

Όσον αφορά τις μεταβλητές για πρόβλεψη, τα μοντέλα με 2 μεταβλητές περιέχουν αρκετή πληροφορία για την πρόβλεψη μεταβλητών. Σα μέτρο σύγκρισης μπορούμε να θεωρήσουμε τον απλό μηχανισμό πρόβλεψης που προβλέπει καμία αιτιότητα για κάθε ζεύγος μετοχών που επιτυγχάνει ακρίβεια 0.86 στο δείγμα μας. Τα τυχαία δάση επιλέχθηκαν ζυγισμένα και συνεπώς η αντίστοιχη στάθμη αναφοράς είναι 0.5.

Σε κάθε περίπτωση βρίσκουμε πως τα μέτρα Jaccard και $m5$ αποδίδουν καλύτερα από τα υπόλοιπα μέτρα. Επίσης με βάση τα αποτελέσματα στα ζυγισμένα δάση, αυτές οι μετρικές έχουν και μεγαλύτερη μεταβλητότητα στα μοντέλα 3 και 5 μεταβλητών. Η ερμηνεία των αποτελεσμάτων είναι δύσκολη διότι συνδυαστικά μπορεί οι μεταβλητές να μην περιέχουν αρκετή πληροφορία για το διαχωρισμό των περιπτώσεων ή μπορεί σε κάποιες περιπτώσεις οι 2 μελετούμενες μετρικές να περιέχουν την ίδια πληροφορία και να μην έχουν επιπλέον συνεισφορά στην πρόβλεψη. Σε κάθε περίπτωση, μπορούμε να διακρίνουμε πως η μετρική $m1$ δε διαθέτει μεγάλη προβλεπτική ικανότητα, όπως και η $m2$, αν και με μεγαλύτερη μεταβλητότητα.

Πρέπει να σημειωθεί εδώ πως οι δείκτες που προτείνονται έχουν σημαντικά υψηλότερη επίδοση στην περίοδο 0, που είναι η περίοδος που καλύπτει την έλευση του κορωνοϊού και όπου η μέθοδος της συσχέτισης έχει σημαντικά χαμηλότερη επίδοση σε αντίθεση με την άλλη περίοδο 1 έτους, αλλά και τις 3 περιόδους 100 ημερών. Ένα χαρακτηριστικό της περιόδου αυτής είναι η ύπαρξη πολλών σχέσεων αιτιότητας κατά Granger (42 ζεύγη στα 306). Επίσης, στους στατιστικούς ελέγχους, η τάξη των μοντέλων VAR που προτείνονται σε αυτή την περίοδο είναι υψηλότερη από την περίοδο 5. Έχουμε αντίστοιχα για το συντελεστή σχιαγράφησης πως η επίδοση της συσχέτισης είναι υψηλότερη από τις επιδόσεις των άλλων 2 δεικτών.

Συνολικά, τα αποτελέσματα δείχνουν ότι οι δείκτες από μόνοι τους υστερούν της συσχέτισης όσον αφορά τη συσταδοποίηση και στις 2 μετρικές, τα σκορ Rand όπου η βάση σύγκρισης είναι η αληθής ταξινόμηση των μετοχών με βάση τον τομέα τους καθώς και στη σχιαγράφηση που αποτελεί γενικό μέτρο ποιότητας, αν και δεν ενδείκνυται για σύγκριση μεταξύ διαφορετικών μετρικών. Παρουσιάζονται μικρές αποκλίσεις για την περίοδο των 30 ημερών καθώς και υπεραπόδοση των προτεινόμενων μέτρων για την κρίσιμη περίοδο του κορωνοϊού. Μία ερμηνεία συμβατή με τα δεδομένα είναι πως σε καιρούς ηρεμίας, ο γράφος ορατότητας δίνει τοπική πληροφορία. Αυτό είναι συμβατό με τη μικρή πιθανότητα σύνδεσης απομακρυσμένων χρονικά κόμβων (Luque et al., 2008) στην εξιδανικευμένη περίπτωση που οι μετοχές ακολουθούν τυχαίο περίπατο (Shirokikh et al, 2022). Η υπεροχή της μεθόδου για την περίοδο 0 δείχνει πως πιθανώς οι δείκτες να μπορούν να κατοπτρίσουν με ακρίβεια μακροπρόθεσμες/μη γραμμικές αλληλεπιδράσεις, μία αδυναμία της μεθόδου της συσχέτισης (Marti et al., 2021).

Τα μέτρα της ομαλής ορατότητας με παράθυρο 5, 11 και 22 χρονικών στιγμών αντίθετα δίνουν ικανοποιητικά αποτελέσματα για την περίοδο 5 (και τα τρία αποτελέσματα είναι ψηλότερα από τα σκορ για τις μετρικές Jaccard, SMC αλλά πάλι υστερούν σε σχέση με τη συσχέτιση. Γενικά, οι μετρικές αυτές εξόρισμού έχουν

τοπικό χαρακτήρα και αυτό μειώνει την απόδοση τους. Η αύξηση όμως του παραθύρου οδηγεί σε μείωση του μήκους του διανύσματος στο οποίο μετασχηματίζεται η χρονοσειρά. Χαρακτηριστικά, η ετήσια χρονοσειρά μας συμπυκνώνεται σε 11 στιγμές στην περίπτωση που θέσουμε το παράθυρο ίσο με 22. Έχουμε λοιπόν ένα trade-off που επηρεάζει σημαντικά την επιλογή του παραθύρου. Παράλληλα, έχουμε το γεγονός ότι κάποιες από τις μετρικές που υπολογίσαμε παρουσιάζουν σημαντικά καλύτερη απόδοση σε ένα έτος έναντι του άλλου. Θα μπορούσε να αποδοθεί στην τύχη αλλά οι μεταβολές στα σκορ Rand των υπολογισμών είναι μεγάλες σε σχέση με τη μεταβλητότητα των σκορ Rand για την ίδια κατηγορία μετρικής.

Παράλληλα, δείξαμε ότι δεν υπάρχει αυτοσυσχέτιση μεταξύ των μετρικών με τη βοήθεια στατιστικών ελέγχων. Αυτό σημαίνει πως η στατιστική τους ερμηνεία είναι συχνά απλούστερη και μπορούν να υπόκεινται σε στατιστικές και θεωρητικές αναλύσεις με απλούστερες υποθέσεις.

Όσον αφορά τα μοντέλα τυχαίων δασών έχουμε πως για την περίοδο 0 που διαθέτει τις περισσότερες σχέσεις αιτιότητας κατά Granger, η ανάλυση μας καταδεικνύει πως η εισαγωγή της μετρικής SMC μείωσε την ακρίβεια του μοντέλου και συνεπώς οι δείκτες που φαίνονται πιο κρίσιμοι για την εκτίμηση είναι οι αρχικοί 2, Jaccard και m_5 . Η ακρίβεια που επιτυγχάνεται είναι σημαντική (0.83 με 5 μεταβλητές, 0.66 με 3 μεταβλητές) και υπάρχει περιθώριο βελτίωσης.

6 Μελλοντική Έρευνα

Η διερεύνηση των μετρικών αυτών δεν έχει εξαντληθεί σε αυτή την εργασία. Πρώτα από όλα, κάποιες απλές κατευθύνσεις έρευνας που μπορούν να ακολουθηθούν είναι η δοκιμή σε άλλες χρονιές των δεδομένων καθώς και η επιλογή διαφορετικών χρονικών παραθύρων. Όπως αναφέρεται και στο Marti et al. (2021) η επιλογή του μήκους των παραθύρων είναι εν μέρει αυθαίρετη σε αυτό το πεδίο και η ανάλυση των παραθύρων

που δίνουν σταθερές συσταδοποιήσεις με τη χρήση των νεοεισαχθέντων μετρικών θα δώσει κάποιες θεωρητικές απαντήσεις σε αυτό το ερώτημα. Παράλληλα, η μελέτη της συμπεριφοράς των συστάδων με διαφορετικούς αλγορίθμους αποτελεί σημαντικό ερώτημα και υποδεικνύεται από τα χαμηλά σκορ σκιαγράφησης των μεθόδων Jaccard, SMC. Η ίδια η μέθοδος της σκιαγράφησης αποδίδει καλύτερα σε συστάδες που είναι κυρτές και η διερεύνηση του αν υπάρχουν αλγόριθμοι συσταδοποίησης που να αποδίδουν καλύτερα αποτελεί πιθανό μελλοντικό ερώτημα.

Ένας ακόμη σημαντικός έλεγχος που είναι από τον τομέα της ομοιότητας χρονοσειρών είναι η υλοποίηση ενός αλγορίθμου πλησιέστερου γείτονα για την κατηγοριοποίηση των χρονοσειρών σε συστάδες. Υλοποιήθηκε στην εργασία του Wang και των συνεργατών του (2012) και θα μπορούσε να αποτελέσει τη βάση για τη διαλογή της κατάλληλης μετρικής για προβλέψεις αιτιότητας, ειδικά στο πλαίσιο περισσότερων χρονοσειρών προς σύγκριση.

Παρατηρήθηκε επίσης πως υπάρχουν ριζικά διαφορετικά αποτελέσματα στα δύο έτη που μελετήθηκαν, με το ένα έτος να περιλαμβάνει την περίοδο του σοκ του κορωνοϊού. Σε μια τέτοια περίπτωση είναι σημαντικό να μελετηθεί άμεσα η σχέση των μετρικών με το δημιουργούμενο δίκτυο. Συγκεκριμένα, θα μπορούσε να μελετηθεί το δίκτυο που δημιουργείται αν χρησιμοποιήσουμε την τιμή Φ του ελέγχου κρισιμότητας για την αιτιότητα κατά Granger και να συγκριθεί άμεσα με τη συσταδοποίηση που δημιουργείται από τις μετρικές μας. Σε γενικότερο πλαίσιο, μία σύγκριση των δικτύων που δημιουργούνται από το κριτήριο της αιτιότητας κατά Granger και από το κριτήριο της αιτιότητας μπορεί να διαλευκάνει το πεδίο της ομοιότητας χρονοσειρών και να αποτελέσει τη βάση και για θεωρητικές εξελίξεις στο αντικείμενο.

Επίσης, έχει ενδιαφέρον η χρήση ενός άλλου αλγορίθμου ορατότητας στον ορισμό της ομαλής ορατότητας. Συγκεκριμένα η χρήση του Limited Penetrability Visibility μπορεί να άρει εν μέρει τον τοπικό χαρακτήρα των ομαλής ορατότητας. Επίσης, ένας μετασχηματισμός όπου η ομαλή ορατότητα κάθε σημείου ορίζεται με αλληλε-

πικαλυπτόμενα διαστήματα θα μπορούσε να αποφύγει το πρόβλημα της μείωσης της πληροφορίας που υφίσταται στην παρούσα υλοποίηση.

Τέλος, στο θεωρητικό κομμάτι, έχουμε δει ήδη το tradeoff μεταξύ της επιλογής μεγάλου ή μικρού μήκους παραθύρου στη μετρική της ομαλής ορατότητας. Έχουμε τότε ότι με μεθόδους υπολογισμού της εντροπίας (από το θεωρητικό ορισμό των ιδιοτήτων των γράφων οριζόντιας πιθανότητας) μπορεί να επιλεγεί το κατάλληλο παράθυρο. Συγκεκριμένα, για χρονοσειρά μήκους N (μεγάλη τιμή) κάθε παράθυρο διαθέτει $M^2 - 3M + 2$ μη ισοπίθανες διαφορετικές τιμές και η χρονοσειρά συμπυκνώνεται σε N/M τιμές.

7 Αναφορές

Laopodis, Nikiforos (2021). Understanding Investments: Theories and Strategies (2nd Edition). Routledge, New York.

Marti G. & Nielsen F. & Bińkowski M. & Donnat P., 2017. "A review of two decades of correlations, hierarchies, networks and clustering in financial markets," Papers 1703.00485, arXiv.org αναθεωρήθηκε 11/2020.

Markowitz, H. (1952). Portfolio Selection. The Journal of Finance, 7(1), 77–91. <https://doi.org/10.2307/2975974>

Tola V., Lillo F., Gallegati M., Mantegna R. N. (2008). Cluster analysis for portfolio optimization. J. Econ. Dyn. Control 32, 235–258. 10.1016/j.jedc.2007.01.034

Wu L, Ahmad M, Qureshi SA, Raza K, Khan YA (2022) An analysis of machine learning risk factors and risk parity portfolio optimization. PLoS ONE 17(9): e0272521. <https://doi.org/10.1371/journal.pone.0272521>

Nanda, S.R., Mahanty, B., & Tiwari, M.K. (2010). Clustering Indian stock market data for portfolio management. Expert Syst. Appl., 37, 8793-8798.

Lacasa, L., Luque, B., Ballesteros, F.J., Luque, J., & Nuño, J.C. (2008). From time series to complex networks: The visibility graph. Proceedings of the National Academy of Sciences, 105, 4972 - 4975.

Murphy, K. P. (2022). Probabilistic Machine Learning: An Introduction, 1st Edition. MIT PRESS, Massachusetts USA.

Bouchaud JP. Econophysics: still fringe after 30 years? Europhysics News 50 (1) 24-27 (2019) DOI: 10.1051/epn/2019103

Simeone, A. (2023), Predictive Methods in Economics: The Link Between Econophysics and Artificial Intelligence, p. 107-122, Springer

Yee, A. K. (2021). Econophysics: making sense of a chimera. European Journal for Philosophy of Science 11 (4):1-34.

Kutner, Ryszard & Ausloos, Marcel & Grech, Dariusz & Di Matteo, Tiziana

& Schinckus, Christophe & Stanley, H.. (2018). Econophysics and sociophysics: Their milestones & challenges. *Physica A: Statistical Mechanics and its Applications*. 516. 10.1016/j.physa.2018.10.019.

Chakraborti, A. & Toke, I. & Patriarca, M. & Abergel, F. (2011). Econophysics review: I. Empirical facts. *Quantitative Finance*. 11. 991-1012. 10.1080/14697688.2010.539248.

Mantegna, R.N. (1999) Hierarchical Structure in Financial Markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11, 193-197.

M. Raddant & T. Di Matteo, 2023. "A look at financial dependencies by means of econophysics and financial economics," *Journal of Economic Interaction and Coordination*, Springer; Society for Economic Science with Heterogeneous Interacting Agents, vol. 18(4), pages 701-734, October.

Ruiz Real, Jose & Uribe-Toril, Juan & Torres, José & De Pablo, Jaime. (2021). Artificial intelligence in business and economics research: Trends and future. *Journal of Business Economics and Management*. 22. 98-117. 10.3846/j-bem.2020.13641.

Hoang, Daniel and Wiegratz, Kevin, *Machine Learning Methods in Finance: Recent Applications and Prospects* (January 24, 2023). Available at SSRN: <https://ssrn.com/abstract=4293977> or <http://dx.doi.org/10.2139/ssrn.4293977>

Διαμαντάρας Κ., Μπότσης Δ. (2019). *Μηχανική Μάθηση*. Κλειδάριθμος, Αθήνα
Kelly, Bryan T. and Xiu, Dacheng, *Financial Machine Learning* (July 1, 2023). Available at SSRN: <https://ssrn.com/abstract=4501707>

Latham, J. R. (2022). *The research canvas: Framework for designing and aligning the “DNA” of your research study* (5th ed.). Organization Design Studio® Ltd. <https://www.drjohnlatham.com/books/research-canvas/>

Johannesson, Paul & Perjons, Erik. (2014). *An Introduction to Design Science*. 10.1007/978-3-319-10632-8.

J.-P. Onnela, A. Chakraborti, K. Kaski and J. Kertész. *Physica A: Statistical Mechanics and its Applications*, 2003, vol. 324, issue 1, 247-252

Mantegna, Rosario & Stanley, H.. (2000). An Introduction to Econophysics: Correlations and Complexity in Finance. 10.1063/1.1341926.

So, Mike & Chu, Amanda & Chan, Thomas. (2020). Impacts of the COVID-19 pandemic on financial market connectedness. *Finance Research Letters*. 38. 101864. 10.1016/j.frl.2020.101864.

Saha, S., Gao, J., & Gerlach, R.H. (2022). A survey of the application of graph-based approaches in stock market analysis and prediction. *International Journal of Data Science and Analytics*, 14, 1 - 15.

Hartman, David & Hlinka, Jaroslav. (2018). Nonlinearity in stock networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*. 28. 083127. 10.1063/1.5023309.

Macmahon, Mel & Garlaschelli, Diego. (2013). Community Detection for Correlation Matrices. *Physical Review X*. 5. 10.1103/PhysRevX.5.021006.

Tumminello, Michele & Aste, Tomaso & Di Matteo, Tiziana & Mantegna, Rosario. (2005). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*. 102. 10421-6. 10.1073/pnas.0500298102.

Jenna Birch & Athanasios A. Pantelous & Kimmo Soramäki, 2016. "Analysis of Correlation Based Networks Representing DAX 30 Stock Price Returns," *Computational Economics*, Springer;Society for Computational Economics, vol. 47(4), pages 501-525, April.

Guo, Xue & Zhang, Hu & Tian, Tianhai. (2018). Development of stock correlation networks using mutual information and financial big data. *PloS one*. 13. e0195941. 10.1371/journal.pone.0195941.

Basu S, Shojaie A, Michailidis G. Network Granger Causality with Inherent

Grouping Structure. *J Mach Learn Res.* 2015;16(13):417-453. PMID: 34267606; PMCID: PMC8278320.

Papana, Angeliki & Kyrtsov, Catherine & Kugiumtzis, Dimitris & Diks, Cees, 2017. "Financial networks based on Granger causality: A case study," *Physica A: Statistical Mechanics and its Applications*, Elsevier, vol. 482(C), pages 65-73.

Siggiridou, Elsa & Koutlis, Christos & Tsimpiris, Alkiviadis & Kugiumtzis, Dimitris. (2019). Evaluation of Granger causality measures for constructing networks from multivariate time series.

Shirokikh, Oleg & Pastukhov, Grigory & Semenov, Alexander & Butenko, Sergiy & Veremyev, Alexander & Pasiliao, Eduardo & Boginski, Vladimir. (2022). Networks of causal relationships in the U.S. stock market. *Dependence Modeling*. 10. 177-190. 10.1515/demo-2022-0110.

Zou, Y., Donner, R.V., Marwan, N., Donges, J.F., & Kurths, J. (2019). Co-mplex network approaches to nonlinear time series analysis. *Physics Reports* 21

Yang, Yue & Yang, Huijie & Yang, Huijie & Mang, Jingshi. (2009). Visibility graph approach to exchange rate series. *Physica A: Statistical Mechanics and its Applications*. 388. 4431-4437. 10.1016/j.physa.2009.07.016.

Long, Yu. (2013). Visibility graph network analysis of gold price time series. *Physica A: Statistical Mechanics and its Applications*. 392. 3374-3384. 10.1016/j.physa.2013.03.063.

Luque, Bartolo & Lacasa, Lucas & Ballesteros, Fernando & Luque, Jordi. (2009). Horizontal visibility graphs: Exact results for random time series. *Physical review. E, Statistical, nonlinear, and soft matter physics*. 80. 046103. 10.1103/PhysRevE.80.046103.

Ting-Ting, Zhou & Ning-De, Tin & Zhong-Ke, Gao & Yue-Bin, Luo. (2012). Limited penetrable visibility graph for establishing complex network from time series. *Acta Physica Sinica -Chinese Edition-*. 61. 10.7498/aps.61.030506.

- Núñez, Angel & Lacasa, Lucas & Gomez, Jose & Luque, Bartolo. (2012). Visibility Algorithms: A Short Review. 10.5772/34810
- Azizi, Hadis & Sulaimany, Sadegh. (2024). A review of visibility graph analysis. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3401485.
- Gutin, Gregory & Mansour, Toufik & Severini, Simone. (2011). A characterization of horizontal visibility graphs and combinatorics on words. Physica A Statistical and Theoretical Physics. 390. 2421-. 10.1016/j.physa.2011.02.031.
- Gutin, Gregory & Mansour, Toufik & Severini, Simone. (2011). A characterization of horizontal visibility graphs and combinatorics on words. Physica A Statistical and Theoretical Physics. 390. 2421-. 10.1016/j.physa.2011.02.031.
- Serafino, Matteo & Gabrielli, Andrea & Caldarelli, Guido & Cimini, Giulio. (2017). Statistical validation of financial time series via visibility graph.
- Mutua, Stephen & Gu, Changgui & Yang, Huijie. (2015). Visibility Graph Based Time Series Analysis. PloS one. 10. e0143015. 10.1371/journal.pone.0143015.
- τεξτλανιλι, Κεσηι & Ωενγ, Τονγφενγ & Γυ, ηανγγυι & Ψανγ, Ηυιθιε, 2020. 'ϊ-σιβιλιτψ γραπη αναλψσις οφ Βιτςοιν πρισε σεριεσ,' Πηψσιςα Α: Στατιστισαλ Μεσηανις ανδ ιτς Αππλιςατιονς, Ελσειερ, ολ. 538(°).
- Βλαχάβας Ι., Κεφαλάς Π., Βασιλειάδης Ν., Κόκκορας Φ., Σακελλαρίου Η. (2020). Τεχνητή Νοημοσύνη (4η Έκδοση). Εκδόσεις Πανεπιστημίου Μακεδονίας, Θεσσαλονίκη
- Θεοδωρίδης , Σέργιος Πικράκης , Άγγελος , Κουτρούμπας , Κωνσταντίνος Σ. , Κάβουρας , Διονύσης (2011). Εισαγωγή στην αναγνώριση προτύπων με MATLAB. Εκδόσεις Πασχαλίδης
- Alpaydin, E. (2020) Introduction to Machine Learning. 4th Edition, MIT Press Academic, Cambridge.
- Timm, Neil. (2002). Applied Multivariate Analysis. 10.1007/b98963.
- Podobnik, B. and Stanley, H.E. (2008) Detrended Cross-Correlation Analysis:

A New Method for Analyzing Two Nonstationary Time Series. *Physical Review Letters*, 100, 084102.

Θεοδωρίδης Σ., Κουτρούμπας Κ. (2012). *Αναγνώριση Προτύπων*. Broken Hill Publishers, Nicosia

Dolphin, Rian & Smyth, Barry & Xu, Yang & Dong, Ruihai. (2021). Measuring Financial Time Series Similarity with a View to Identifying Profitable Stock Market Opportunities. *arXiv:2107.03926v1 [q-fin.ST]* 7 Jul 2021

Evan Anderson, Eric Ghysels and Jennifer L. Juergens. The impact of risk and uncertainty on expected returns. *Journal of Financial Economics*, 2009, vol. 94, issue 2, 233-263

Franses, Philip & Wiemann, Thomas. (2020). Intertemporal Similarity of Economic Time Series: An Application of Dynamic Time Warping. *Computational Economics*. 56. 10.1007/s10614-020-09986-0.

Serrà, Joan & Arcos, Josep Lluís. (2014). An Empirical Evaluation of Similarity Measures for Time Series Classification. *Knowledge-Based Systems*. 67. 10.1016/j.knosys.2014.04.035.

Βαζιργιάννη Μ. - Χαλκίδη Μ., Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό. Τυπωθήτω, 2005.

Liu Y, Li Z, Xiong H, Gao X, Wu J (2010) Understanding of internal clustering validation measures. In: *IEEE international conference on data mining*, pp 911–916

D Tsiotas, A Charakopoulos. *Visibility in the topology of complex networks - Physica A: Statistical Mechanics and its Applications*, 2018

Asteriou, D. and Hall, S. (2019) *Applied Econometrics: A Modern Approach* (3rd. Edition). Palgrave Macmillan, New York.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, 48(1), 1–48. <https://doi.org/10.2307/1912017>

Diebold, F.X. (2019), *Econometric Data Science: A Predictive Modeling Approach*, Department of Economics, University of Pennsylvania, <http://www.ssc.upenn.edu/fdiebold/Textbooks.html>.

Denscombe, M. (2010). *The Good Research Guide for Small Scale Research Projects* (4th ed.). Buckingham: Open University Press.

Cuthbertson, K., & Nitzche, D. (2004). *Quantitative Financial Economics* (2nd ed.). Chichester: John Wiley and Sons.

Pawel Dlotko & Wanling Qiu & Simon Rudkin, 2022. "Topological Data Analysis Ball Mapper for Finance," Papers 2206.03622, arXiv.org.

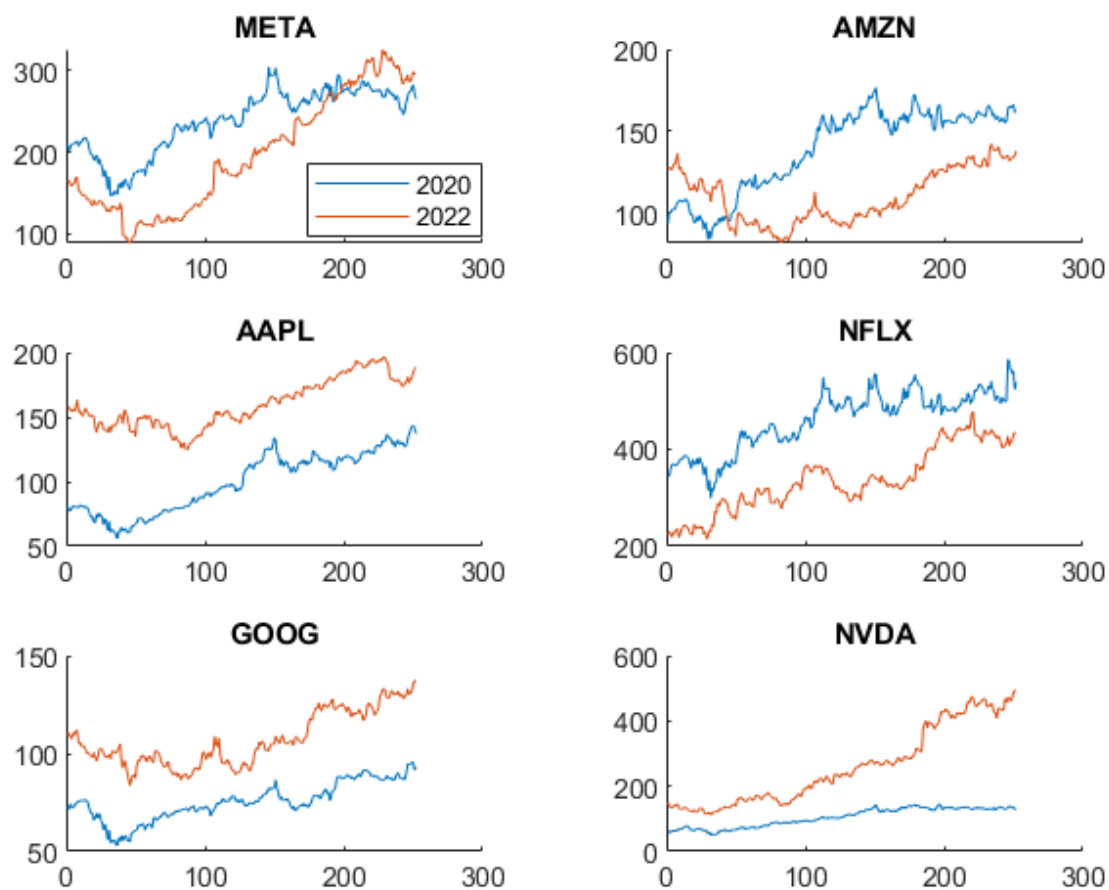
Richmond, Peter & Mimkes, Jürgen & Hutzler, Stefan. (2013). *Econophysics and Physical Economics*. 10.1093/acprof:oso/9780199674701.001.0001.

Hudson, Robert & Gregoriou, Andros. (2010). Calculating and Comparing Security Returns is Harder than you Think: A Comparison between Logarithmic and Simple Returns. *International Review of Financial Analysis*. 38. 10.2139/ssrn.1549328.

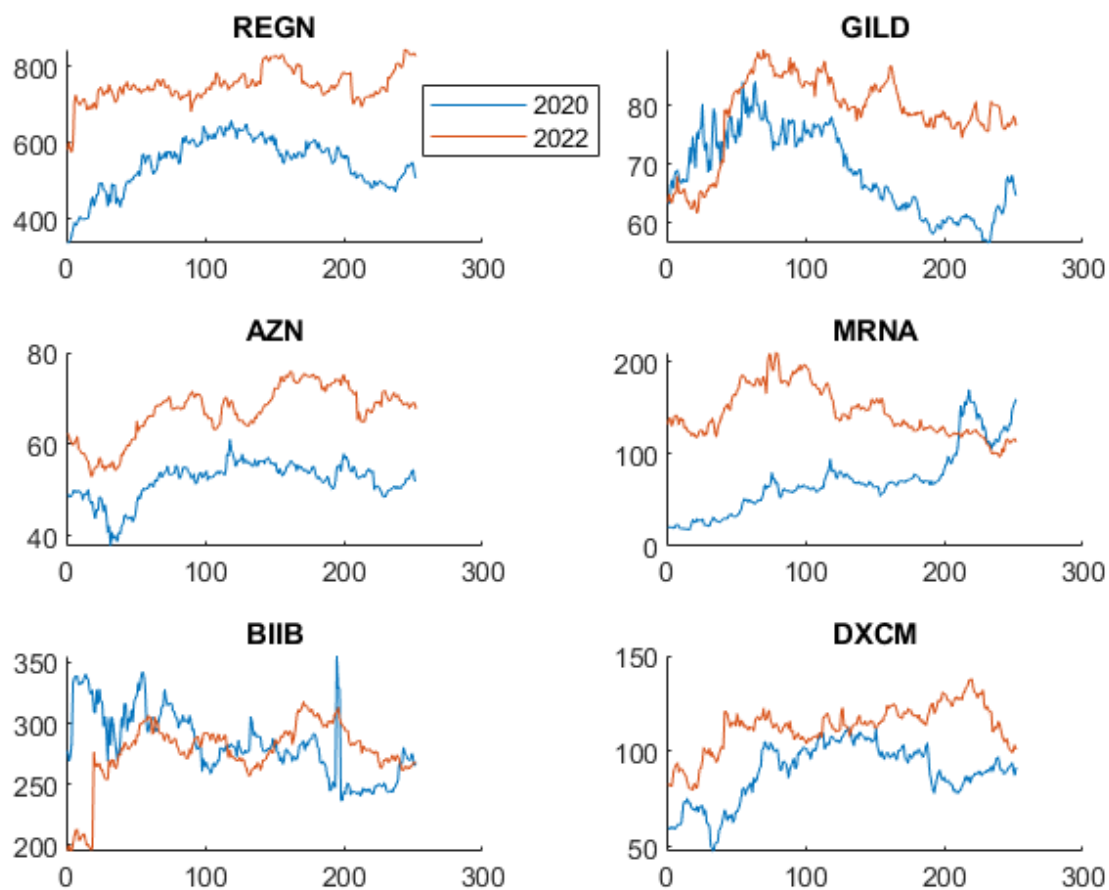
Výrost, Tomáš & Lyócsa, Štefan & Baumöhl, Eduard, 2015. "Granger causality stock market networks: Temporal proximity and preferential attachment," *Physica A: Statistical Mechanics and its Applications*, Elsevier, vol. 427(C), pages 262-276.

Neusser, Klaus, (2016), *Time Series Econometrics*, Springer.

8 Παραρτήματα



Σχήμα 8: Περιγραφική Στατιστική-Μετοχές Τεχνολογίας



Σχήμα 9: Περιγραφική Στατιστική-Μετοχές Υγείας



Σχήμα 10: Περιγραφική Στατιστική-Μετοχές Καταναλωτικών Αγαθών