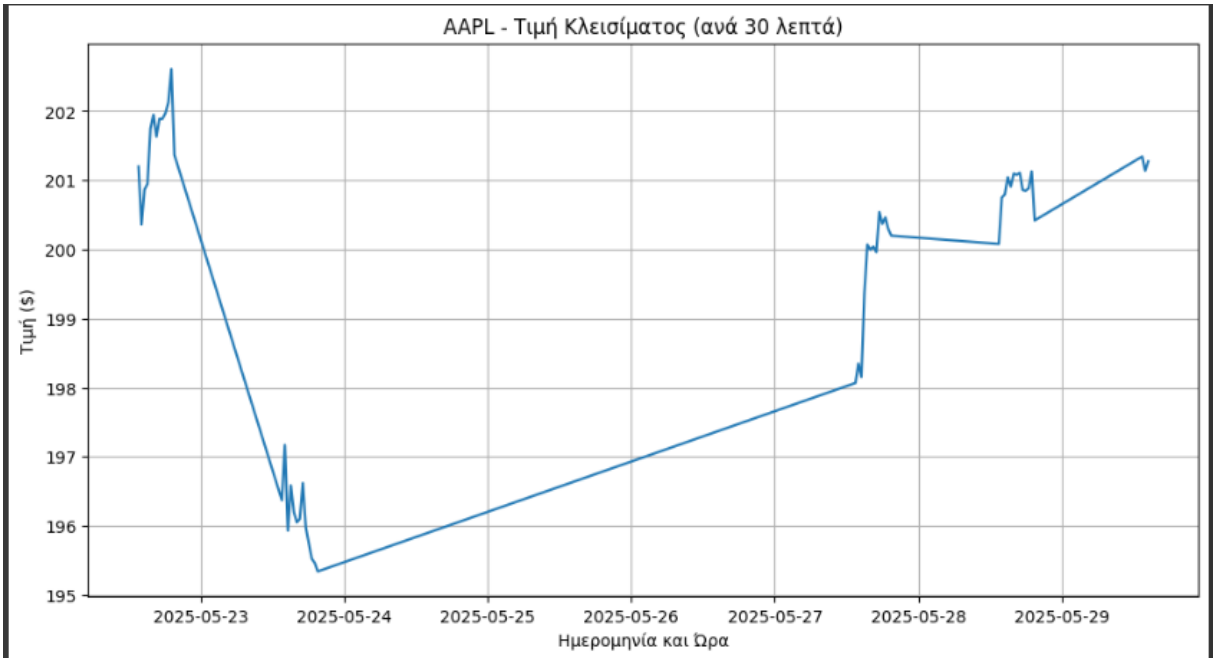
Αριθμητική Ανάλυση - Εργασία

Κουδούνης Βασίλειος ΑΕΜ:10739
Παπαδόπουλος Παναγιώτης ΑΕΜ: 10697

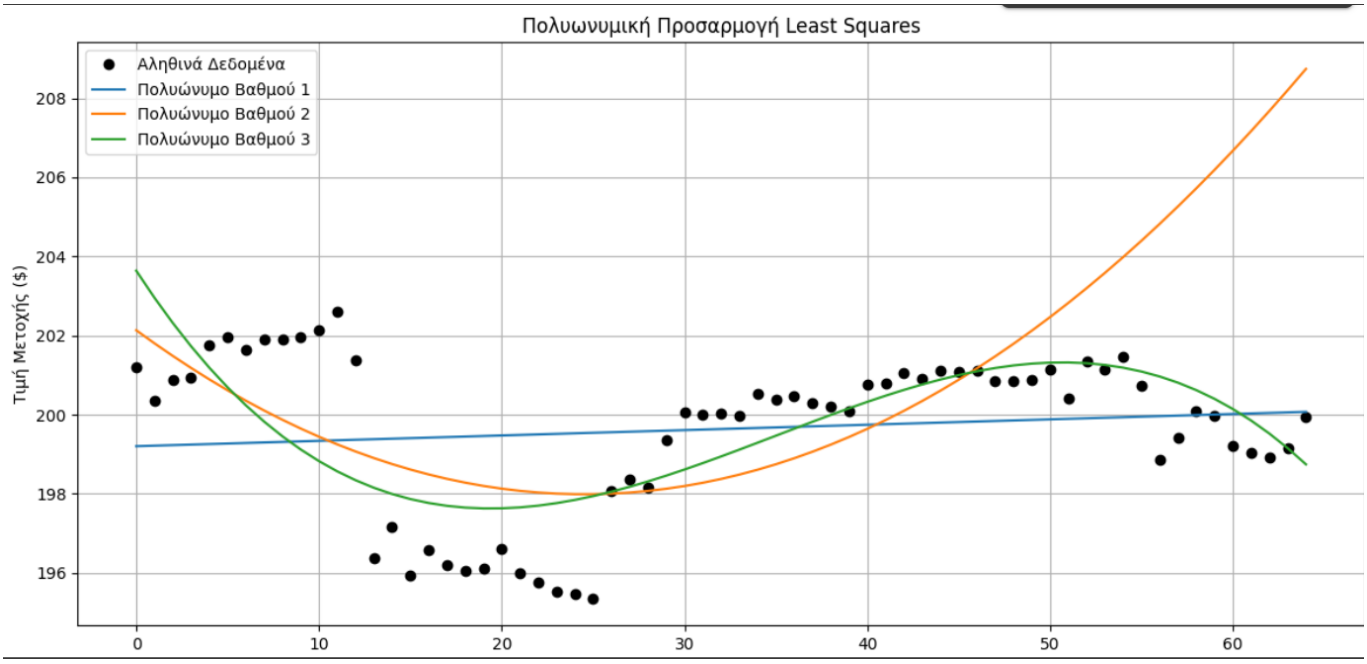
31 Μαΐου 2025

1. Συλλογή Δεδομένων

Για την εργασία μας επιλέχθηκε η μετοχή της εταιρείας Apple έπειτα από έντονη παρότρυνση του Παναγιώτη. Παρακάτω εμφανίζεται η πορεία της μετοχής για διάστημα 5 ημερών με δειγματοληψία κάθε 30 λεπτά.



2. Πολυωνυμική Προσαρμογή Ελαχίστων Τετραγώνων



Βαθμός	**MAE (Train)**	**MSE (Train)**	**MAE (Test)**	**MSE (Test)**
1	1.856544	4.776260	0.824576	0.905201
2	1.484002	2.846558	5.872632	40.554410
3	1.227826	2.438008	0.675183	0.795443

Πίνακας 1: Σφάλματα MAE και MSE για πολυωνυμικά μοντέλα βαθμού 1, 2 και 3 στο training και test set.

Ανάλυση Αποτελεσμάτων Μοντέλων και Υπερπροσαρμογή

Από τον Πίνακα 1 παρατηρούμε ότι:

- Η απόδοση στο training set βελτιώνεται όσο αυξάνεται ο βαθμός του πολυωνύμου. Το MAE μειώνεται από 1.85 (βαθμός 1) σε 1.23 (βαθμός 3), γεγονός που ανταναικλά τη μεγαλύτερη ευελιξία των πολυωνυμικών μοντέλων υψηλότερου βαθμού να προσαρμόζονται στα δεδομένα εκπαίδευσης.
- Ωστόσο, στο test set παρατηρούμε μια εντελώς διαφορετική συμπεριφορά. Το μοντέλο βαθμού 2, παρά το σχετικά χαμηλό σφάλμα στο training, παρουσιάζει πολύ μεγάλο σφάλμα στο τεστ (MAE = 5.87, ΜΣΕ = 40.55), υποδεικνύοντας ισχυρό φαινόμενο υπερπροσαρμογής (**overfitting**).
- Το μοντέλο βαθμού 3 επιτυγχάνει τη βέλτιστη ισορροπία: έχει το χαμηλότερο **MAE** και **MSE** στο **test set**, και παράλληλα μικρό σφάλμα στο training. Αυτό σημαίνει ότι καταφέρνει να προσαρμόζεται στα δεδομένα χωρίς να υπερπροσαρμόζει.

Συμπέρασμα: Αν και η αύξηση του βαθμού μειώνει το σφάλμα εκπαίδευσης, αυτό δεν συνεπάγεται απαραίτητα καλύτερη πρόβλεψη. Το μοντέλο βαθμού 2 μαθαίνει υπερβολικά καλά τα δεδομένα εκπαίδευσης, αλλά αποτυγχάνει να γενικεύσει σε νέα δεδομένα. Το μοντέλο βαθμού 3 είναι τελικά το πιο αξιόπιστο, επιτυγχάνοντας χαμηλό σφάλμα και στο **training** και στο **test set**. Ένα πολυώνυμο υψηλότερου βαθμού έχει περισσότερους βαθμούς ελευθερίας και μπορεί να προσαρμοστεί σχεδόν τέλεια στα σημεία του training set. Αυτό σημαίνει ότι μπορεί να «μάθει» όχι μόνο τη γενική τάση των δεδομένων, αλλά και τον θόρυβο ή τις τοπικές διακυμάνσεις.

Αυτό οδηγεί στο φαινόμενο υπερπροσαρμογής (overfitting): το μοντέλο έχει εξαιρετική απόδοση στα δεδομένα εκπαίδευσης, αλλά αποτυγχάνει να γενικεύσει σε νέα ή άγνωστα δεδομένα (π.χ. test set ή μελλοντική πρόβλεψη).

Πώς το εντοπίζουμε αριθμητικά:

Υπολογίζουμε τα σφάλματα στο training set και στο test set. Αν παρατηρήσουμε ότι το σφάλμα στο training set είναι πολύ μικρό, ενώ το σφάλμα στο test set είναι σημαντικά μεγαλύτερο, τότε υπάρχει έντονη ένδειξη για φαινόμενο overfitting.

Παράδειγμα:

- $\text{MAE}_{\text{Train}} = 1.2$ (πολύ καλό)
- $\text{MAE}_{\text{Test}} = 5.8$ (πολύ χειρότερο)

Αυτό σημαίνει ότι το μοντέλο ¨έμαθε' υπερβολικά καλά τα σημεία του training, αλλά αποτυγχάνει να γενικεύσει σε νέα δεδομένα.

Μέτρα για την αποφυγή του **overfitting**:

- Επιλογή απλούστερου μοντέλου (μικρότερος βαθμός πολυωνύμου).
- Χρήση cross-validation για πιο αξιόπιστη αξιολόγηση. Αυτή η μέθοδος χρησιμοποιείται συχνά σε μοντέλα μηχανικής μάθησης.
- Εφαρμογή τεχνικής L2 regularization. Αυτή η τεχνική αναγκάζει τους συντελεστές να παραμένουν μικροί και ομαλοποιημένοι. Αυτό έχει ως αποτέλεσμα:

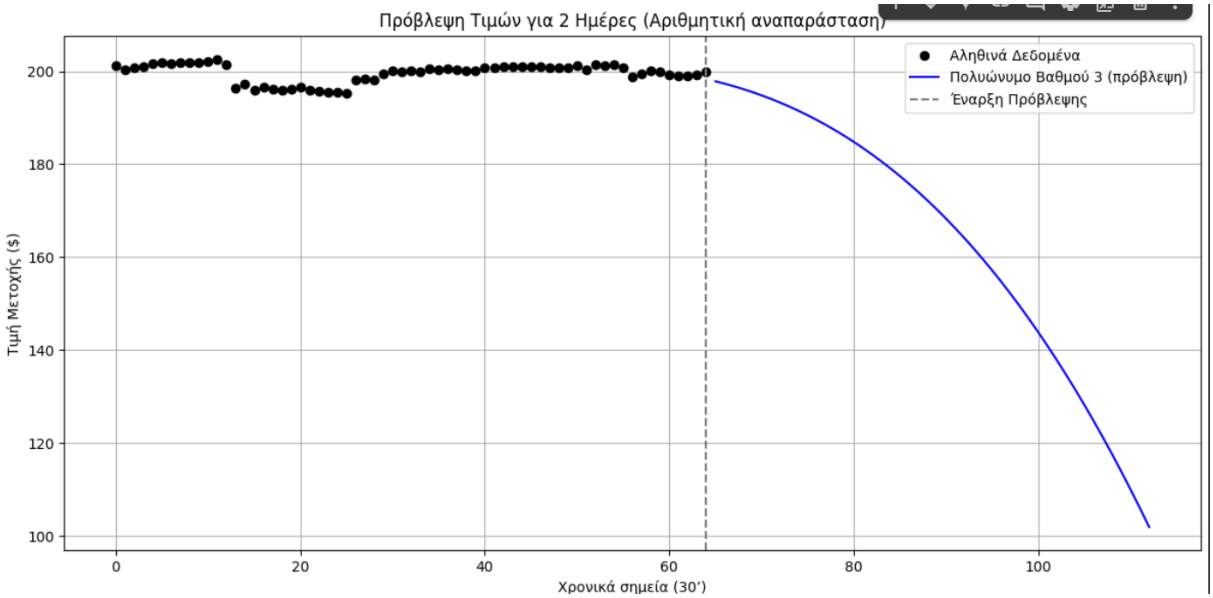
  1.Να αποφεύγεται η δημιουργία πολυωνύμων με υπερβολική καμπυλότητα,

  2.Να μειώνεται η ευαισθησία του μοντέλου στον θόρυβο του training set,

  3.Να ενισχύεται η γενίκευση σε νέα δεδομένα.

- Αύξηση του μεγέθους του training set για καλύτερη γενίκευση.

3. Πρόβλεψη Επόμενης μέρας



Το πολυώνυμο 3ου βαθμού είχε το χαμηλότερο σφάλμα στο test set, άρα φαινομενικά ήταν το "καλύτερο" μοντέλο.

Ωστόσο, στην εξωπραγματική πρόβλεψη για την επόμενη ημέρα:

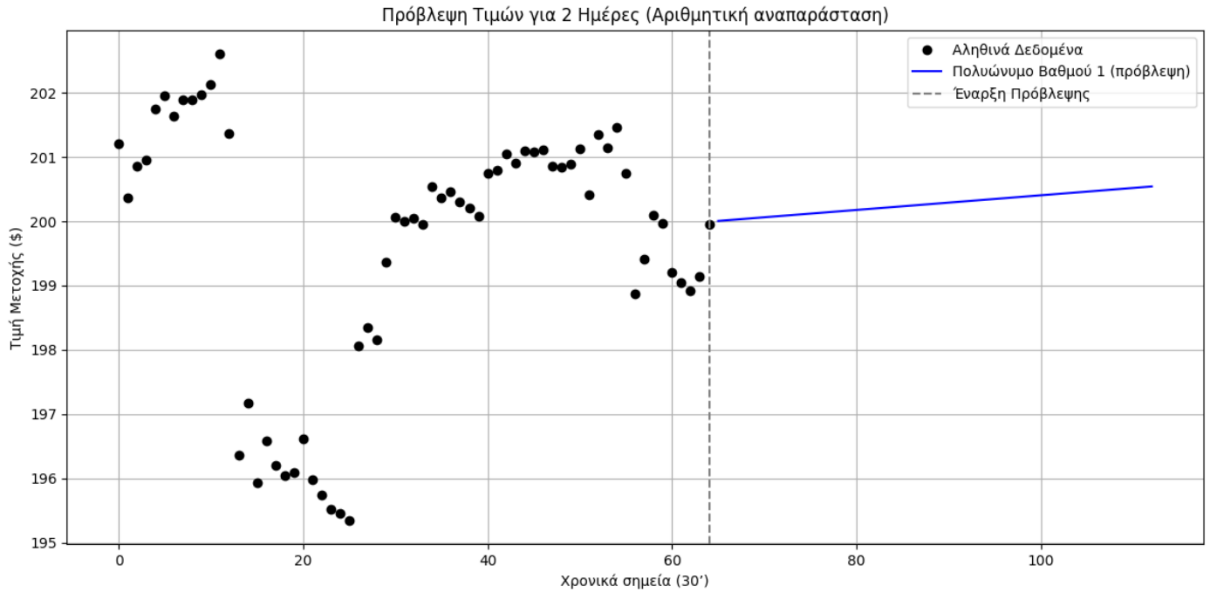
Η καμπύλη ¨βυθίζεται' απότομα.

Το μοντέλο επηρεάζεται έντονα από την τελευταία αρνητική κλίση των δεδομένων.

Το αποτέλεσμα είναι μία μη ρεαλιστική πρόβλεψη (π.χ. πτώση κάτω από 140 δολάρια σε 2 μέρες).

Το πολυώνυμο 3ου βαθμού έχει μεγαλύτερη ευκαμψία και προσαρμόζεται "υπερβολικά" στα δεδομένα. Αυτό είναι ένα τυπικό παράδειγμα του φαινομένου overfitting στην πρόβλεψη, ακόμα και αν δεν το δείχνει ο αριθμητικός δείκτης σφάλματος.

Το πολυώνυμο 1ου βαθμού, παρότι είχε ελαφρώς υψηλότερο MAE στο test set, παράγει πιο σταθερή, ρεαλιστική και "συντηρητική" πρόβλεψη, χωρίς απότομες καμπές ή τεχνητές "εκρήξεις".



Η αστάθεια της πρόβλεψης με πολυώνυμο 3ου βαθμού δεν εντοπίστηκε άμεσα μέσω του test set, πιθανώς λόγω του μικρού αριθμού δειγμάτων και της σχετικά ήρεμης περιοχής στα τελευταία σημεία. Με περιορισμένα δεδομένα, το μοντέλο δεν ¨προλαβαίνει' να αποκαλύψει την τάση του για υπερπροσαρμογή και ασταθή πρόβλεψη.

Αυτό καταδεικνύει τη σημασία ύπαρξης μεγαλύτερων και πιο αντιπροσωπευτικών συνόλων δεδομένων για την αξιόπιστη αξιολόγηση ενός μοντέλου.

Γενικά, η επιλογή βαθμού πολυωνύμου επηρεάζει άμεσα την ικανότητα του μοντέλου να προσαρμόζεται στα δεδομένα και να κάνει προβλέψεις.

Ένα πολυώνυμο χαμηλού βαθμού (π.χ. βαθμός 1) μπορεί να υποπροσαρμόζει τα δεδομένα, δηλαδή να μην ακολουθεί τις μεταβολές τους με ακρίβεια.

Ένα πολυώνυμο υψηλού βαθμού έχει μεγαλύτερη ευελιξία και μπορεί να προσαρμόζεται πολύ καλά στα σημεία του training set, αλλά συχνά υπερπροσαρμόζει (overfitting), με αποτέλεσμα:

να ¨μαθαίνει' τον θόρυβο των δεδομένων,

και να παράγει ασταθείς ή μη ρεαλιστικές προβλέψεις έξω από το γνωστό διάστημα (πρόβλεψη).

Η αστάθεια της πρόβλεψης μπορεί να εντοπιστεί με:

1.Οπτική παρατήρηση της καμπύλης πρόβλεψης: Αν η πολυωνυμική καμπύλη εκτρέπεται απότομα προς τα πάνω ή κάτω μετά το τελευταίο γνωστό σημείο, υπάρχει αστάθεια.

2. Αριθμητική σύγκριση σφαλμάτων: Αν το σφάλμα στο training είναι πολύ μικρό (π.χ. MAE $= 1.2$), αλλά στο τεστ πολύ μεγάλο (π.χ. MAE $= 5.8$), αυτό υποδεικνύει overfitting και πιθανή αστάθεια στην πρόβλεψη.

4. Ολοκλήρωση της Καμπύλης Τιμών για Μέση Τιμή

Μέθοδοι Αριθμητικής Ολοκλήρωσης: Τραπεζίου και **Simpson**

Κανόνας του Τραπεζίου: Η μέθοδος αυτή προσεγγίζει το ολοκλήρωμα χρησιμοποιώντας ευθύγραμμα τμήματα μεταξύ διαδοχικών σημείων.

$$\int_a^b f(x)\,dx \approx \frac{h}{2}\left[f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right] \tag{1}$$

Όπου:

- $h = \frac{b-a}{n}$ είναι το σταθερό βήμα,
- $x_0, x_1, \ldots, x_n$ τα διακριτά σημεία του διαστήματος,
- n το πλήθος των υποδιαστημάτων.

Κανόνας του **Simpson**: Ο κανόνας του Simpson προσεγγίζει την καμπύλη με παραβολές που διέρχονται από κάθε τρία διαδοχικά σημεία. Απαιτεί το πλήθος των διαστημάτων να είναι άρτιο (n άρτιο).

$$\int_a^b f(x)\,dx \approx \frac{h}{3}\left[f(x_0) + 4\sum_{\text{μονά } i} f(x_i) + 2\sum_{\text{ζυγά } i} f(x_i) + f(x_n)\right] \tag{2}$$

Η μέθοδος αυτή είναι συνήθως πιο ακριβής από τον Κανόνα του Τραπεζίου, ειδικά όταν η $f(x)$ είναι ομαλή και έχει καμπυλότητα που μπορεί να περιγραφεί με παραβολές.

Συμπέρασμα: Ο Κανόνας του Τραπεζίου είναι απλούστερος και λειτουργεί καλά για περίπου γραμμικές συναρτήσεις, ενώ ο Κανόνας του Simpson είναι καταλληλότερος για καμπύλες συναρτήσεις με ομαλή παραβολική συμπεριφορά. Και οι δύο μέθοδοι γίνονται πιο ακριβείς όσο μικραίνει το βήμα h.

Βαθμός	Βήμα h	Μέση Τιμή (Τραπεζίου)	Μέση Τιμή (**Simpson**)
1	1	199.628392	199.628392
1	2	199.753067	199.753067
1	3	199.640615	199.632086
1	4	199.741542	199.741542
1	5	199.732208	199.732208
1	6	199.892202	199.892202
2	1	199.607839	199.671512
2	2	199.714938	199.713708
2	3	199.582710	199.516277
2	4	199.663158	199.679832
2	5	199.602935	199.591182
2	6	199.759545	199.744806
3	1	199.607839	199.607839
3	2	199.714938	199.713708
3	3	199.582710	199.600904
3	4	199.663158	199.657932
3	5	199.602935	199.591182
3	6	199.759545	199.744806

Πίνακας 2: Μέση τιμή της τιμής της μετοχής για διαφορετικούς βαθμούς πολυωνύμου και τιμές βήματος h, με χρήση του κανόνα του Τραπεζίου και του κανόνα του Simpson.

Επίδραση του Βήματος στην Ακρίβεια της Αριθμητικής Ολοκλήρωσης

Η επιλογή του βήματος h επηρεάζει σημαντικά την ακρίβεια των αριθμητικών μεθόδων ολοκλήρωσης, όπως ο Κανόνας του Τραπεζίου και ο Κανόνας του Simpson. Όταν το h είναι μικρό, η καμπύλη της συνάρτησης προσεγγίζεται με μεγαλύτερη ακρίβεια, καθώς τα αριθμητικά σχήματα βασίζονται σε τοπικές γραμμικές ή παραβολικές προσπελάσεις της συνάρτησης.

Όσο μικρότερο το h, τόσο περισσότερα σημεία χρησιμοποιούνται, με αποτέλεσμα:

- καλύτερη προσαρμογή της αριθμητικής μεθόδου στο σχήμα της συνάρτησης,
- μείωση του τοπικού σφάλματος ανά υποδιάστημα,
- και βελτίωση της συνολικής ακρίβειας της ολοκλήρωσης.

Αντίθετα, όταν το h είναι μεγάλο, η αριθμητική μέθοδος «χάνει» τη λεπτομέρεια της συνάρτησης και το ολοκλήρωμα γίνεται πιο προσεγγιστικό, με πιθανώς μεγαλύτερο σφάλμα.

Παρατήρηση: Στο συγκεκριμένο παράδειγμα (με δεδομένα τιμών μετοχής), παρατηρείται ότι ανεξαρτήτως βαθμού πολυωνύμου (1 έως 3) και ανεξαρτήτως τιμής βήματος h (από 1 έως 6), η εκτιμώμενη μέση τιμή της συνάρτησης (δηλαδή το εμβαδόν δια του μήκους του διαστήματος) παραμένει σταθερά κοντά στην τιμή 199. Αυτό οφείλεται στην ομαλότητα της καμπύλης και στην απουσία απότομων μεταβολών. Συνεπώς, ακόμα και με διαφορετικές παραμετροποιήσεις, η ακρίβεια παραμένει ικανοποιητική.

Αυτό δείχνει ότι, σε ομαλές συναρτήσεις όπως αυτή της μετοχικής τιμής, οι αριθμητικές μέθοδοι ολοκλήρωσης μπορούν να δώσουν αξιόπιστα αποτελέσματα ακόμη και με σχετικά μεγάλα h.

5. Ανίχνευση Ανωμαλιών στα Ιστορικά Δεδομένα

Ανίχνευση Ανωμαλιών στα Ιστορικά Δεδομένα

Για την ανίχνευση ανωμαλιών στα ιστορικά δεδομένα της μετοχής, εφαρμόστηκαν δύο αριθμητικές μέθοδοι παρεμβολής: ο αλγόριθμος Aitken-Neville και η μέθοδος διηρημένων διαφορών του Newton. Οι μέθοδοι εφαρμόστηκαν σε διαστήματα πέντε συνεχόμενων χρονικών σημείων (δηλαδή κάθε 2,5 ώρες), τα οποία ορίστηκαν με ολίσθηση σε όλο το χρονικό εύρος των διαθέσιμων δεδομένων.

Για κάθε τέτοιο διάστημα, υπολογίστηκε με παρεμβολή η τιμή του επόμενου χρονικού σημείου και στη συνέχεια συγκρίθηκε με την πραγματική τιμή από τα δεδομένα. Η διαδικασία αυτή εφαρμόστηκε για όλα τα διαστήματα και τα αποτελέσματα απεικονίζονται στο γράφημα με τίτλο "AAPL: Interpolation Predictions with Input Points".

Στο γράφημα:

- Οι γκρι κουκκίδες αντιπροσωπεύουν τα σημεία εισόδου για κάθε διάστημα παρεμβολής.
- Η μπλε γραμμή απεικονίζει την πραγματική τιμή της μετοχής.
- Οι μαύροι σταυροί και οι πράσινες διακεκομμένες γραμμές δείχνουν τις προβλέψεις των μεθόδων Aitken και Newton, αντίστοιχα.
- Οι κόκκινοι και μωβ κύκλοι υποδεικνύουν τα σημεία στα οποία καταγράφηκαν ανωμαλίες.

Παρατηρείται ότι οι δύο μέθοδοι παρεμβολής δίνουν απολύτως ταυτόσημες τιμές πρόβλεψης για όλα τα διαστήματα. Αυτό είναι αναμενόμενο από μαθηματική σκοπιά, καθώς ο αλγόριθμος Aitken στην ουσία υλοποιεί αναδρομικά τον ίδιο παρεμβολικό πολυώνυμο που υπολογίζει και η Newton μέσω διαιρεμένων διαφορών.

Για την ανίχνευση των ανωμαλιών χρησιμοποιήθηκε δυναμικό κατώφλι που προσαρμόζεται σε κάθε χρονικό διάστημα σύμφωνα με τον τύπο:

$$\delta = 0.05 \cdot \text{Range} + 0.5 \cdot \sigma$$

όπου:

- Range είναι η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής της μετοχής στο διάστημα,
- σ είναι η τυπική απόκλιση των τιμών του ίδιου διαστήματος.

Στον κώδικα, το κατώφλι αυτό υπολογίζεται δυναμικά για κάθε διάστημα και συγκρίνεται με την απόλυτη διαφορά μεταξύ της προβλεπόμενης τιμής (μέσω παρεμβολής) και της πραγματικής τιμής από τα δεδομένα. Αν η διαφορά αυτή υπερβαίνει το κατώφλι δ, τότε το σημείο καταγράφεται ως ανωμαλία.

Ενδεικτικά, για το δεύτερο διάστημα με σημεία $[201.5, 202, 202.5, 203, 203.5, 204]$, η προβλεπόμενη τιμή από τις μεθόδους Aitken και Newton ήταν περίπου 209, ενώ η πραγματική από τα δεδομένα ήταν γύρω στην τιμή 201. Η απόλυτη διαφορά ξεπέρασε το δυναμικό κατώφλι και έτσι σημειώθηκε ανωμαλία. Ομοίως, το ίδιο συμβαίνει σε όλα τα διαστήματα σημείων. Σε άλλα η διαφορά είναι μικρότερη, όπως το τέταρτο και ένατο διάστημα, ή μεγαλύτερη σαν το πέμπτο και έβδομο.

Αυτή η προσέγγιση αποδείχθηκε αρκετά αποτελεσματική, γιατί το κατώφλι προσαρμόζεται αυτόματα ανάλογα με τη «νευρικότητα» των τιμών. Έτσι, δεν εντοπίζονται ψευδώς ανωμαλίες όταν οι τιμές μεταβάλλονται φυσιολογικά, αλλά μόνο όταν πραγματικά αποκλίνουν από τη γενική τάση.

Η μεταβλητότητα των τιμών παίζει καθοριστικό ρόλο στον καθορισμό του κατωφλίου ανωμαλίας, καθώς αυξάνει τόσο το εύρος (Range) όσο και τη διασπορά (σ) των τιμών. Αυτό οδηγεί σε μεγαλύτερο δ, άρα το σύστημα γίνεται πιο 'ανεκτικό' σε αποκλίσεις.

Σε ένα πιο ευμετάβλητο σύνολο δεδομένων, οι τιμές παρουσιάζουν έντονες διακυμάνσεις, γεγονός που μπορεί να οδηγήσει σε μεγάλες αποκλίσεις ακόμη και υπό κανονικές συνθήκες. Αν χρησιμοποιηθεί ένα σταθερό κατώφλι, τότε οι αποκλίσεις αυτές θα καταγράφονται εσφαλμένα ως ανωμαλίες. Για αυτόν τον λόγο, είναι απαραίτητο το κατώφλι να είναι δυναμικό και προσαρμοστικό στη μεταβλητότητα του διαστήματος, όπως συμβαίνει με τον τύπο $\delta = 0.05 \cdot \text{Range} + 0.5 \cdot \sigma$. Αντίθετα, σε πιο σταθερά σύνολα, ένα χαμηλότερο κατώφλι επιτρέπει την ανίχνευση ακόμα και λεπτών αποκλίσεων. Η διαφορετική αυτή προσέγγιση εξασφαλίζει ισορροπία μεταξύ ακρίβειας και ευαισθησίας στις ανωμαλίες.