ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ & ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μελέτη και εφαρμογή προηγμένων αλγορίθμων αναγνώρισης προτύπων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Λυδία Γ. Ιωάννου

Επιβλέπων: Κωνσταντίνος Παπαοδυσσεύς

Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Απρίλιος 2014

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ & ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μελέτη και εφαρμογή προηγμένων αλγορίθμων αναγνώρισης προτύπων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Λυδία Γ. Ιωάννου

Επιβλέπων: Κωνσταντίνος Παπαοδυσσεύς

Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 28^{η} Απριλίου 2014.

……………………………......	……………………..	……………………....
Κωνσταντίνος Παπαοδυσσεύς	Βασίλειος Λούμος	Ηλίας Κουκούτσης
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

....

Ιωάννου Λυδία

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

ΠΕΡΙΛΗΨΗ

Το ακαδημαϊκό ενδιαφέρον της παρούσας διπλωματικής εργασίας εστιάζεται στη μελέτη και εφαρμογή προηγμένων αλγορίθμων αναγνώρισης προτύπων με στόχο την αυτόματη αναγνώριση του γραφέα από τα κείμενά του. Η αυτόματη ταυτοποίηση του γραφέα ενός αρχαίου κειμένου με χρήση σύγχρονων μαθηματικών αλγορίθμων, μπορεί να οδηγήσει με ασφάλεια στην ακριβή χρονολόγηση ενός αρχαίου κειμένου, άρα και στην επίλυση των σχετικών σημαντικών επιστημονικών προβλημάτων. Στα πλαίσια αυτής της διπλωματικής γίνεται αρχικά μία πλήρης παρουσίαση του μαθηματικού υπόβαθρου που χρησιμοποιήθηκε και στη συνέχεια γίνεται μία εκτενής περιγραφή της μεθόδου που εφαρμόστηκε. Συνοπτικά αυτή περιλαμβάνει: (α) κατάτμηση εικόνας για την εξαγωγή του σώματος κάθε γράμματος κάθε κειμένου, (β) εξαγωγή του περιγράμματος του κάθε σώματος και (γ) χρήση πρωτότυπων τεχνικών επεξεργασίας εικόνας, καμπυλών και αναγνώρισης προτύπων. Τέλος, τα αποτελέσματα, που προέκυψαν από την εφαρμογή της εξεταζόμενης μεθόδου, μας οδηγούν σε σημαντικά συμπεράσματα γύρω από το αντικείμενο.

Λέξεις Κλειδιά:

Κατάτμηση εικόνας, αναγνώριση προτύπων, αυτόματη αναγνώριση γραφέα, αρχαία παπυρολογικά κείμενα.

ABSTRACT

The academic interest of this thesis focuses on the study and application of advanced pattern recognition algorithms in order to automatically identify the scribe from his writings. The automatic identification of the scribe of an ancient text using modern mathematical algorithms can safely lead to the precise dating of an ancient text, and thus to the settlement of the relevant scientific problems. In the context of this thesis we initially make a complete presentation of the mathematical background that is used and then an extensive description of the applied method. This method briefly includes: (a) image segmentation to extract the body of each letter of each text, (b) extraction of the contour of each body and (c) use of original methods of image and curve processing and pattern recognition. Finally, the results obtained from the application of the test method, lead us to important conclusions about the subject studied.

Key words:

Image segmentation, pattern recognition, automatic writer identification, ancient papyrus texts.

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο και σημασία της παρούσης διπλωματικής

Η παρούσα διπλωματική εργασία πραγματεύεται τη μελέτη και εφαρμογή προηγμένων αλγορίθμων αναγνώρισης προτύπων με στόχο την αυτόματη ταυτοποίηση του γραφέα από τα κείμενα του. Συγκεκριμένα, εφαρμόστηκαν νέες μέθοδοι ταυτοποίησης γραφέα σε παπυρολογικά κείμενα του έργου «Αθηναίων Πολιτεία» του Αριστοτέλη, με ειδική επεξεργασία του κάθε γράμματος του κάθε κειμένου ξεχωριστά, και ακολούθως παρήχθησαν κατάλληλες μορφές των συμβόλων της αλφαβήτου με σκοπό τη στατική επεξεργασία τους. Για το σκοπό αυτό πραγματοποιήθηκαν τα εξής βήματα:

(α) Χρησιμοποιήθηκε μία πρωτότυπη μέθοδος κατάτμησης εικόνας για την εξαγωγή του σώματος κάθε γράμματος του κάθε κειμένου.

(β) Χρησιμοποιήθηκαν άλλοι αλγόριθμοι για την εξαγωγή του περιγράμματος του σώματος κάθε γράμματος.

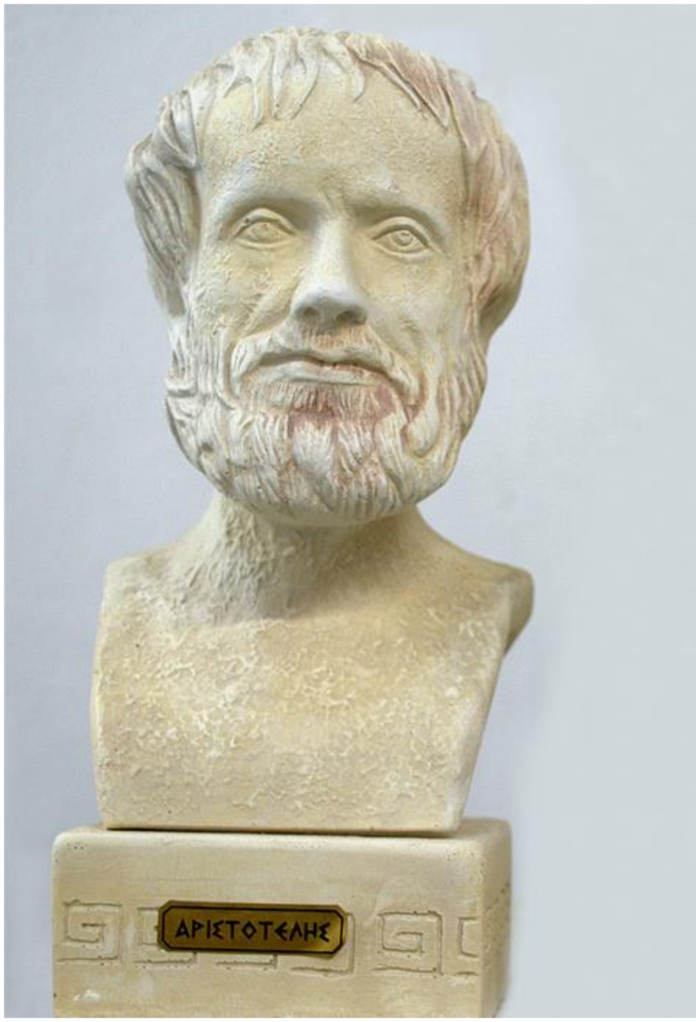
(γ) Χρησιμοποιήθηκαν πρωτότυπες τεχνικές επεξεργασίας εικόνας, καμπυλών και αναγνώρισης προτύπων για τη δημιουργία αντικειμένων, τα οποία βέλτιστα προσιδιάζουν τον αντιπρόσωπο κάθε συμβόλου της αλφαβήτου ανά κείμενο που μελετάται.

Σε αυτό το σημείο αξίζει να αναφέρουμε τη σπουδαιότητα της ταυτοποίησης των γραφέων των αρχαίων κειμένων αναφορικά με την ακριβή χρονολόγηση ενός αρχαίου κειμένου. Κατά την αρχαιότητα οι γραφείς κατά κανόνα δεν υπέγραφαν τα κείμενα τους, ούτε έβαζαν ημερομηνία με συνέπεια τη δημιουργία αμφιβολιών και ενδεχομένως επιστημονικών διαφωνιών αναφορικά με την ακριβή χρονολόγηση ενός αρχαίου κειμένου. Με την εφαρμογή μαθηματικοποιημένων κριτηρίων είναι δυνατή η επεξεργασία των γραμμάτων των αρχαίων κειμένων ώστε να προκύψουν συμπεράσματα για το γραφέα του κειμένου και επομένως για την χρονολογική περίοδο στην οποία ανήκει το κάθε κείμενο. Τα μαθηματικοποιημένα αυτά κριτήρια, σε αντίθεση με όσα εφαρμόζονται μέχρι σήμερα, δίνουν ποσοτικά αποτελέσματα, τα οποία με τη σειρά τους επιτρέπουν την ταυτοποίηση του γραφέα και συνεπώς τον διαχωρισμό και την κατάταξη των κειμένων. Συνεπώς, η αυτόματη ταυτοποίηση του γραφέα ενός αρχαίου κειμένου με χρήση σύγχρονων μαθηματικών αλγορίθμων, μπορεί να οδηγήσει με ασφάλεια στην ακριβή χρονολόγηση ενός αρχαίου κειμένου, άρα και στην επίλυση των σχετικών σημαντικών επιστημονικών προβλημάτων. Επιπλέον, η αυτόματη ταυτοποίηση γραφέων είναι καθοριστικής σημασίας για την ορθή καταγραφή της ιστορίας σε παγκόσμιο επίπεδο, αφού μπορεί να εφαρμοστεί όχι μόνο σε αρχαία ελληνικά κείμενα, αλλά σε οποιοδήποτε αρχαίο κείμενο. Εκτός από την αρχαιολογία που ασχολείται άμεσα με τα αρχαιολογικά κείμενα και τη ιστορική χρονολόγησή τους, παράλληλα επωφελούνται και οι υπόλοιπες ανθρωπιστικές επιστήμες, που έχουν πεδίο έρευνας στην αρχαία εποχή, από τις πληροφορίες που μπορούμε να συλλέξουμε μέσω της ταυτοποίησης του γραφέα.

Συνοψίζοντας, ο βασικός σκοπός του έργου που παρουσιάζεται στην παρούσα διπλωματική είναι : (α) να οδηγηθούμε με όσον το δυνατόν περισσότερη αντικειμενικότητα στην απόδοση του κάθε κειμένου στον σωστό γραφέα, (β) να δημιουργήσουμε τις βάσεις ώστε να επιτύχουμε την ταυτοποίηση του γραφέα σε περιπτώσεις που οι αρχαιολογικές πληροφορίες και αναλύσεις δεν μπορούν, και (γ) να συμβάλουμε προς την επίλυση αρχαιολογικών διαφορών σχετικά με τη σωστή χρονολόγηση των γεγονότων και (δ) να μελετήσουμε διεξοδικότερα και να εφαρμόσουμε σε βυζαντινούς κώδικες μια νέα μέθοδο γραφέων, η οποία πρωτοπαρουσιάστηκε ιδιαιτέρως στο [7].

1.2 Ιστορικά στοιχεία

Το συγγραφικό έργο του Αριστοτέλη. Ήδη στην αρχαιότητα τα έργα του Αριστοτέλη χωρίζονταν σε δύο κυρίως ομάδες. Την πρώτη την αποτελούσαν έργα με τα οποία ο Αριστοτέλης απευθυνόταν σε ένα πλατύτερο αναγνωστικό κοινό, ένα κοινό πέρα από το ακροατήριό του στην Ακαδημία και, αργότερα, στο Λύκειο. Οι αρχαίοι ονόμαζαν τα έργα της ομάδας αυτής, ακριβώς γι αυτόν τον λόγο, *εξωτερικά*. Σήμερα δεν έχουμε από τα έργα αυτά του Αριστοτέλη παρά μόνο αποσπάσματα, κομμάτια δηλαδή, μικρότερα ή μεγαλύτερα, που διασώθηκαν στα έργα άλλων συγγραφέων της αρχαιότητας. Τη δεύτερη ομάδα την αποτελούν τα έργα του που μας σώθηκαν ολόκληρα. Όλα τους σχεδόν είναι τα χειρόγραφα που είχε μαζί του ο Αριστοτέλης στα μαθήματά του˙ μερικά μάλιστα από αυτά δεν είναι παρά σημειώσεις στη συντομότερη δυνατή μορφή, πρέπει επομένως να είχαν γραφεί για αποκλειστικά προσωπική χρήση, για να βοηθήσουν δηλαδή τη μνήμη του Αριστοτέλη την ώρα της παράδοσης (οι αρχαίοι ονόμαζαν τα έργα αυτά *ακροαματικά*). Η απορία τότε για το ποια έργα του σώθηκαν ως εμάς και ποιά έργα του χάθηκαν είναι πλέον δικαιολογημένη: Τα έργα του Αριστοτέλη που κυκλοφόρησαν στο ευρύ κοινό, τα έργα επομένως που πολλαπλασιάστηκαν σε πλήθος αντιτύπων, είναι ακριβώς τα έργα του που χάθηκαν˙ αντίθετα, τα έργα του από τα οποία δεν υπήρξε καν δεύτερο αντίτυπο πέρα από το προσωπικό του χειρόγραφο, αυτά έφτασαν τελικά ως εμάς. [1]



Εικόνα 1: προτομή Αριστοτέλη

Ύστερα από τον θάνατο του Αριστοτέλη τα βιβλία του πέρασαν στην ιδιοκτησία του Θεόφραστου. Με τη σειρά του ο Θεόφραστος τα κληροδότησε στον Νηλέα από τη Σκήψη, τον τελευταίο που ζούσε ακόμη από τον στενό κύκλο των φίλων του Αριστοτέλη. Λίγο ύστερα από τον θάνατο του Θεόφραστου, δηλαδή λίγο ύστερα από το έτος 287/6, ο Νηλέας έφυγε από την Αθήνα για να εγκατασταθεί στην ιδιαίτερη πατρίδα του. Μαζί του έφερε εκεί και τα προσωπικά χειρόγραφα του παλιού του φίλου, πιο πολύ από σεβασμό προς τη μνήμη του. Οι κληρονόμοι του, ανίδεοι άνθρωποι, δεν είχαν πια κανένα ενδιαφέρον για το περιεχόμενο των βιβλίων που κληρονόμησαν, πολύ θα ήθελαν όμως να τα σώσουν από τη βουλιμία των ηγεμόνων της Περγάμου, που τι δεν θα έκαναν για να αποκτήσουν την πολύτιμη συλλογή, αφού είχαν και αυτοί αποδυθεί στον ίδιο με τους Πτολεμαίους της Αιγύπτου αγώνα για τη δημιουργία μεγάλης βιβλιοθήκης. Το μόνο λοιπόν που τους έμενε να κάνουν ήταν να τα κρύψουν σε μία υπόγεια αποθήκη στη Σκήψη. Εκεί τα ανακάλυψε ένας περίφημος βιβλιόφιλος της Αθήνας, ο Απελλικών, στις αρχές του 1ου πια αιώνα π.Χ. Ο Απελλικών έφτασε στη Σκήψη στην κατάλληλη στιγμή. Τα βιβλία είχαν αρχίσει κιόλας να καταστρέφονται από την υγρασία και τα σκουλήκια. Τα αγόρασε λοιπόν και τα ξανάφερε πίσω στην Αθήνα. Όταν όμως το 86 π.Χ. ο Ρωμαίος Σύλλας κυρίεψε την Αθήνα, ανάμεσα στην πλούσια λεία που έστειλε από εκεί στη Ρώμη, ήταν και τα βιβλία του Αριστοτέλη. Νέες περιπέτειες άρχιζαν εκεί για τα βιβλία του μεγάλου φιλοσόφου. Αίτιοι τη φορά αυτή ήταν διάφοροι αντιγραφείς. Ώσπου στο δεύτερο μισό του πρώτου αιώνα π.Χ. ανέλαβε και έφερε σε πέρας την έκδοση των έργων του Αριστοτέλη ο Ροδίτης Ανδρόνικος. Η έκδοση αυτή έδωσε το έναυσμα για μία συστηματική πλέον απασχόληση με τα έργα και τη φιλοσοφία του Αριστοτέλη. Σύντομα άρχισε μάλιστα, με βάση τα αντίγραφα της έκδοσης αυτής, και ο σχολιασμός των έργων του Αριστοτέλη. Σε αυτό ακριβώς το ενδιαφέρον, που ήταν ο καρπός της συστηματικής εργασίας του Ανδρόνικου, χρωστούμε εμείς σήμερα όσα έργα του Αριστοτέλη έχουμε. Την ίδια όμως στιγμή το ενδιαφέρον για τα έργα που αποκάλυψε στον κόσμο ο Ανδρόνικος, έκανε να ξεχαστούν τα άλλα έργα του Αριστοτέλη, εκείνα που είχαν γνωρίσει τη μεγάλη δημοσιότητα. [1]

Μένει να τονίσουμε ότι ο Αριστοτέλης καλλιέργησε με εντυπωσιακή επιτυχία πολλούς επιμέρους φιλοσοφικούς και επιστημονικούς κλάδους (κάποιων μάλιστα από αυτούς υπήρξε και ο ιδρυτής και πρώτος διαμορφωτής τους). Μερικοί από αυτούς είναι η λογική, η φυσική, η κοσμολογία, η ζωολογία, η βιολογία, η φυσιολογία, η ψυχολογία, η οντολογία, η ηθική και πολιτική φιλοσοφία, η ρητορική, η λογοτεχνική κριτική.

Οι ιστορικές πραγματείες του Αριστοτέλη. Με το όνομα *Πολιτείαι* (πολιτεύματα) ήταν γνωστή στην αρχαιότητα μία συλλογή από πραγματείες του Αριστοτέλη στις οποίες περιγράφονταν οι πολιτικές αρχές ενός μεγάλου αριθμού ελληνικών κρατών. Οι εν λόγω πραγματείες έφταναν τις 158 σύμφωνα με τους καταλόγους του Ησυχίου και του Διογένους του Λαερτίου, και τις χρησιμοποιούσαν επανειλημμένα οι συγγραφείς, που έτσι μας διασώσανε πολυάριθμα αποσπάσματα. Φαίνεται όμως να είναι αλήθεια ότι ο Αριστοτέλης ασχολήθηκε και με την έρευνα και κριτική των πολιτευμάτων ορισμένων βαρβαρικών κρατών και ιδιαίτερα με το πολίτευμα των Καρχηδονίων. Γι αυτό μία δεύτερη πραγματεία, που έχει κατά την παράδοση το όνομα *Νόμιμα βαρβαρικά*, περιλάμβανε την ιστορία και κριτική ανάλυση των αξιολογότερων πολιτευμάτων της Ασίας, της Ιταλίας, της Σικελίας και της Β. Αφρικής. Όσο για την ιστορία του πολιτεύματος της Καρχηδόνος, γι αυτό ο φιλόσοφος έγραψε ξεχωριστή πραγματεία. [2]

Η «Αθηναίων Πολιτεία». Η θλιβερή διαπίστωση είναι ότι απ' όλη αυτή τη συλλογή, μόνο η *«Αθηναίων Πολιτεία»* έφτασε σε εμάς σχεδόν ακέραια. Το κείμενο της *«Αθηναίων Πολιτείας»* μας το αποκάλυψε τον Ιανουάριο του 1891 ο Άγγλος φιλόλογος Φρειδερίκος Κένυον, που το διάβασε και το δημοσίευσε για πρώτη φορά, όταν βρέθηκε μέσα σε μία δέσμη παπύρων του Βρετανικού Μουσείου- το 1885- είχαν διαβαστεί σ' έναν φθαρμένο πάπυρο του Αιγυπτιακού Μουσείου του Βερολίνου μερικές περικοπές, που εύκολα αναγνωρίστηκαν ότι ανήκαν σ' αυτό το σύγγραμμα.

Στα διάφορα αυτά έργα, που αποτέλεσαν τη λεγόμενη «Συναγωγή Πολιτειών», απλούστερα τις «Πολιτείες», χρησιμοποιήθηκε εξάπαντος το υλικό που είχε συλλέξει ο Αριστοτέλης για να συγγράψει τα «Πολιτικά» του και που το ξαναθεώρησε, το συμπλήρωσε και το τακτοποίησε με τη βοήθεια των μαθητών του. Προκειμένου για την τεράστια συγγραφική παραγωγή του Αριστοτέλη, δεν πρέπει να ξεχνάμε ότι η ομαδική εργασία, η συνεργασία, ήταν σχεδόν ο απόλυτος κανόνας που έβρισκε εφαρμογή στην Περιπατητική Σχολή. Αλλά αυτό το σύστημα εργασίας προϋποθέτει φυσικά προσεκτική επαγρύπνηση και επιθεώρηση εκ μέρους του Αριστοτέλη. Όσον αφορά μάλιστα τα κεφαλαιώδη έργα, μεταξύ των οποίων συμπεριλαμβάνεται και η «Αθηναίων Πολιτεία», μπορεί να θεωρηθεί βέβαιο ότι μόνο την καθαρά συλλεκτική εργασία έκαναν οι μαθητές, ενώ η κατάταξη και η σύνταξη είναι έργο προσωπικό του δασκάλου.

Όταν δημοσιεύτηκε για πρώτη φορά το κείμενο της «Αθηναίων Πολιτείας» οι νέες πληροφορίες και απόψεις που έδινε το έργο αυτό, οι διαφορές και αντιθέσεις του προς την ήδη γνωστή παράδοση, τέλος η σύνθεσή του που ήταν αρκετά διαφορετική από τα άλλα συγγράμματα του Αριστοτέλη, προξένησαν ισχυρή εντύπωση στους μελετητές του. Σε μερικούς μάλιστα απ' αυτούς η χαρά για την ανακάλυψη ήταν ανάμικτη με κάποια δυσάρεστη έκπληξη, σχεδόν απογοήτευση. Η πίστη που είχαν ορισμένοι απ' αυτούς στην ενότητα της ιστορικής παράδοσης της Ελλάδας και στο αλάθητο του Αριστοτέλη έγινε αφορμή για ζωηρές φιλονικίες, που κράτησαν χρόνια, γύρω από το έργο αυτό, που μερικοί σοφοί έφτασαν στο σημείο να του αμφισβητήσουν την αυθεντικότητα, και πολύ περισσότερο την ακεραιότητα.

Σήμερα ο σάλος αυτός της πολεμικής έχει κοπάσει, και τα διάφορα εξαγόμενα που αφορούν τη σύνθεση του έργου και την αξία του σαν ιστορικής μαρτυρίας μπορούν πλέον να θεωρηθούν ασφαλή. Το κείμενο συμπίπτει με τόση ακρίβεια με τα σχετικά αποσπάσματα που μας κληροδότησε η αρχαιότητα, ώστε να μην παρουσιάζεται σήμερα καμία σοβαρή αντιγνωμία όσον αφορά τη γνησιότητα της «Αθηναίων Πολιτείας», της οποίας δεχόμαστε ως χρόνο συγγραφής την περίοδο πριν από το 336 π.Χ. [2]

Κριτική του κειμένου της «Αθηναίων Πολιτείας». Όσον αφορά το κείμενο της «Αθηναίων Πολιτείας», μας παραδόθηκε σε πολύ καλή κατάσταση και μόνο ένα μέρος της αρχής του λείπει. Από την κριτική που έχει γίνει μέχρι σήμερα δεν απορρίπτονται τα κεφάλαια εκείνα που είχαν αρχικά θεωρηθεί πλαστά ή μεταγενέστερες προσθήκες. Από την εξέταση του έργου πειθόμαστε ότι δεν υπάρχει τίποτα που να μη μπορούσε να το γράψει ο ίδιος και ότι η πρωτοτυπία και η σημασία του έργου βρίσκεται ακριβώς στις νέες ειδήσεις και απόψεις που μας παρουσιάζει.

Το σύγγραμμα αυτό αποτελείται από δύο μέρη αρκετά διαφορετικά μεταξύ τους. Τα 41 πρώτα κεφάλαια του έργου, που απαρτίζουν το «ιστορικό» μέρος, μας εκθέτουν την εξέλιξη του αθηναϊκού πολιτεύματος μέχρι το έτος 473 π.Χ., τότε που επώνυμος άρχων των Αθηνών ήταν ο Ευκλείδης.

Στο δεύτερο μέρος, το «περιγραφικό», - από το κεφάλαιο 42 μέχρι το τέλος- εκτίθεται αναλυτικά το πολιτειακό σύστημα των Αθηνών κατά τους χρόνους του Αριστοτέλη. Η διαίρεση αυτή σε δύο ενότητες με διαφορετικά θέματα παρουσιάζει σπουδαίες διαφορές ανάμεσα στα δύο μέρη, όσον αφορά την ιστορική μέθοδο του Αριστοτέλη. [2]

2. ΧΡΗΣΙΜΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΘΕΩΡΗΜΑΤΑ ΑΠΟ ΤΗ ΘΕΩΡΙΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΕΠΑΓΩΓΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

2.1 Στατιστικές Κατανομές

2.1.1 Κανονική κατανομή

Ορισμός : Έστω X μία απόλυτα συνεχής τυχαία μεταβλητή. Το X ακολουθεί την *κανονική κατανομή* (normal distribution) με μέση τιμή μ και διασπορά σ^2, σ>0 (συμβολικά $X \sim N(\mu, \sigma^2)$) αν η συνάρτηση πυκνότητας πιθανότητας έχει τύπο :

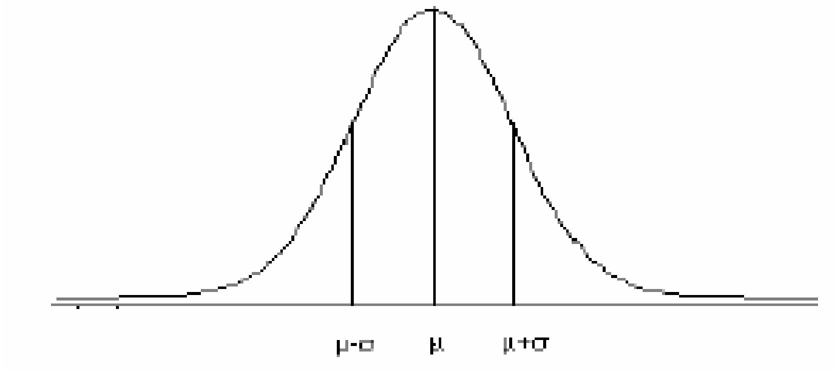
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

όπου $x \in \mathbb{R}$.

Οι παράμετροι μ και σ χαρακτηρίζουν την κανονική κατανομή, δηλαδή, μπορούμε να την προσδιορίσουμε πλήρως αν γνωρίζουμε μόνο τη μέση τιμή της, μ και την τυπική απόκλισή της, σ.

Ιδιότητες :

1. Η συνάρτηση πυκνότητας της κανονικής κατανομής δεν ορίζει μία συγκεκριμένη κανονική καμπύλη αλλά μία οικογένεια κανονικών καμπυλών. Για διαφορετικές τιμές των παραμέτρων μ και σ παίρνουμε διαφορετικές κανονικές καμπύλες.
2. Η κανονική κατανομή είναι συμμετρική γύρω από το σημείο x=μ.
3. Έχει κωδωνοειδή μορφή (σχήμα καμπάνας) με επικρατούσα τιμή στο σημείο x=μ. Επομένως στην θέση x=μ παρουσιάζει μέγιστη τιμή $f(x) = \frac{1}{\sigma\sqrt{2\pi}}$ και στις θέσεις x=μ-σ και x=μ+σ παρουσιάζει σημεία καμπής.



Εικόνα 1: η καμπύλη της συνάρτησης πυκνότητας της κανονικής κατανομής

4. Η κανονική κατανομή είναι μία καλά ορισμένη κατανομή. Αυτό αποδεικνύεται ως εξής :

$$\int_{-\infty}^{+\infty} f(x)dx = 2\int_{\mu}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

Θέτοντας $\frac{x-\mu}{\sigma} = y$ έχουμε $dy = \frac{1}{\sigma}dx$ και έτσι:

$$\int_{-\infty}^{+\infty} f(x)dx = 2\int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

Θέτοντας $\frac{y^2}{2} = u$ έχουμε $ydy = du \Rightarrow dy = \frac{1}{\sqrt{2u}}du$. Έτσι :

$$\int_{-\infty}^{+\infty} f(x)dx = 2\int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{2u} e^{-u} du = \frac{1}{\sqrt{\pi}} \int_{0}^{+\infty} e^{-u} u^{-\frac{1}{2}} du \Rightarrow$$

$$\int_{-\infty}^{+\infty} f(x)dx = \frac{\Gamma\left(\frac{1}{2}\right)}{\sqrt{\pi}} = \frac{\sqrt{\pi}}{\sqrt{\pi}} = 1$$

5. Η μέση τιμή της κανονικής κατανομής είναι **E(X)=μ**.

Απόδειξη :

$$E(X) = \int_{-\infty}^{+\infty} x\, f(x)dx = \int_{-\infty}^{+\infty} (x-\mu+\mu)\, f(x)dx =$$

$$2\int_{\mu}^{+\infty} (x-\mu)\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx + \mu$$

Θέτοντας $y = -\frac{(x-\mu)^2}{2\sigma^2}$ έχουμε $dy = -\frac{x-\mu}{\sigma^2}\ dx$ και έτσι :

$$E(X) = \frac{-2\sigma}{\sqrt{2\pi}}\int_{0}^{-\infty} e^{y} dy + \mu = \mu$$

6. Η διασπορά της κανονικής κατανομής είναι **Δ(X)=σ²**.

Απόδειξη :

$$\Delta(X) = \int_{-\infty}^{+\infty} (x-\mu)^2\, f(x)dx = \int_{-\infty}^{+\infty} (x-\mu)^2\ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

Θέτοντας $y = \frac{x-\mu}{\sigma}$ έχουμε $dy = \frac{1}{\sigma} dx$ και έτσι :

$$\Delta(X) = \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} y^2 e^{-\frac{y^2}{2}} dy = -\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} y(e^{-\frac{y^2}{2}})' dy =$$

$$-\frac{\sigma^2}{\sqrt{2\pi}}\left(0-\sqrt{2\pi}\right) = \sigma^2$$

Τυποποιημένη κανονική κατανομή

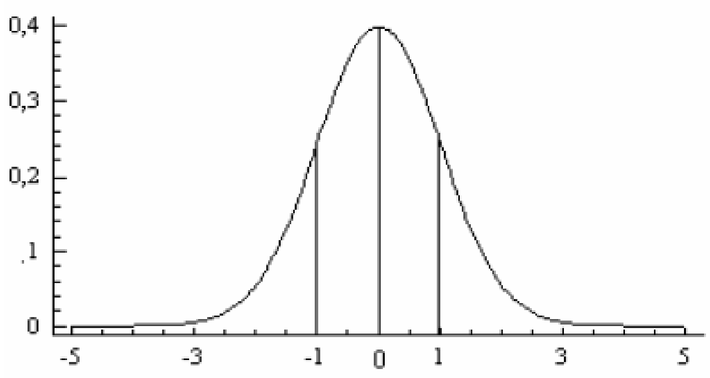
Ο απευθείας υπολογισμός πιθανοτήτων της κανονικής κατανομής είναι δύσκολος λόγω της μορφής της συνάρτησης πυκνότητας πιθανότητας. Ο υπολογισμός όμως αυτός διευκολύνεται με τον ορισμό της τυποποιημένης κανονικής κατανομής.

Ορισμός : Η κανονική κατανομή που έχει μέση τιμή μ=0 και τυπική απόκλιση σ=1, συμβολίζεται με N(0,1) και ονομάζεται τυποποιημένη κανονική κατανομή (standard normal distribution).

Μία τυχαία μεταβλητή που ακολουθεί την τυποποιημένη κανονική κατανομή, έχει επικρατήσει να συμβολίζεται με Z και η συνάρτηση πυκνότητας πιθανότητάς της είναι :

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \qquad z \in \mathbb{R}$$

Η γνώση των πιθανοτήτων της Z παρέχει αμέσως γνώση των πιθανοτήτων της X. Γι' αυτό το λόγο όλοι οι στατιστικοί πίνακες και οι περισσότερες βάσεις δεδομένων στους υπολογιστές υπολογίζουν και χρησιμοποιούν ως τυχαία μεταβλητή αναφοράς την Z.



Εικόνα 2: καμπύλη της συνάρτησης πυκνότητας της τυποποιημένης κανονικής κατανομής

2.1.2 Κατανομή x^2

Ορισμός : Έστω οι ανεξάρτητες τυχαίες μεταβλητές X_1, X_2, X_3,...,X_n οι οποίες ακολουθούν την τυποποιημένη κανονική κατανομή N(0,1). Το άθροισμα των τετραγώνων των τυχαίων μεταβλητών αυτών είναι μη αρνητικό μέγεθος και συμβολίζεται με το X^2, δηλαδή :

$$X^2 = X_1^2 + X_2^2 + X_3^2 + \cdots + X_n^2$$

Στην περίπτωση αυτή λέγεται ότι το τυχαίο μέγεθος X^2 ακολουθεί τη X^2-*κατανομή με n βαθμούς ελευθερίας* (chi-squared distribution) και συμβολίζεται με X_n^2. [3]

Είναι προφανές ότι πρόκειται για οικογένεια κατανομών. Για κάθε τιμή του *n* παίρνουμε και μια άλλη *κατανομή χι-τετράγωνο*.

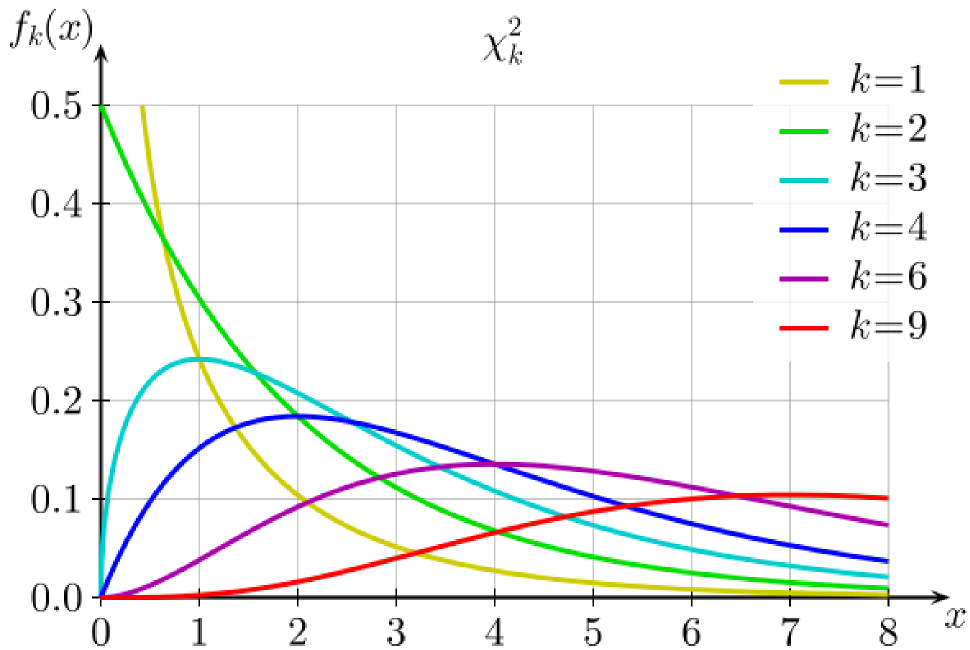
Για να γράψουμε τη συνάρτηση πυκνότητας πιθανότητας της *χι-τετράγωνο*, χρησιμοποιούμε τη συνάρτηση Γάμα, η οποία είναι η αναλυτική επέκταση του n! και δίνεται από τον τύπο :

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx, \qquad n > 0$$

Η συνάρτηση πυκνότητας πιθανότητας της κατανομής *χι-τετράγωνο* είναι :

$$f_{x^2}(x) = \begin{cases} 0, & x \leq 0 \\ \dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0 \end{cases}$$

Στο σχήμα που ακολουθεί φαίνεται η γραφική παράσταση της *συνάρτησης πυκνότητας* για διάφορες τιμές του k.



Εικόνα 3: καμπύλη της συνάρτησης πυκνότητας της κατανομής χ^2 *[4]*

2.1.3 Κατανομή t ή Student

Ορισμός : Έστω ότι η τυχαία μεταβλητή X ακολουθεί την τυποποιημένη κανονική κατανομή N(0,1) και η τυχαία μεταβλητή X^2 την αντίστοιχη X^2-κατανομή με ν βαθμούς ελευθερίας. Οι τυχαίες μεταβλητές X και X^2 θεωρούνται μεταξύ τους ανεξάρτητες. Η τυχαία μεταβλητή :

$$T = \frac{X}{\sqrt{X^2/\nu}}$$

λέγεται ότι ακολουθεί την t-κατανομή ή κατανομή Student. [3]

Η συνάρτηση πυκνότητας πιθανότητας έχει τη μορφή :

$$f_T(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\cdot\Gamma\left(\frac{1}{2}\right)\cdot\sqrt{\nu}}(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}, \qquad -\infty < x < +\infty$$

γεωμετρικά δε είναι καμπύλη συμμετρική ως προς την ευθεία x=0.

Η κατανομή Student είναι ασυμπτωτικά κανονική με μέσο

$$M(T) = 0$$

και διασπορά

$$\Delta(T) = \frac{\nu}{\nu-2}. \qquad [5]$$

2.2 Έλεγχοι στατιστικών υποθέσεων

Ο στατιστικός έλεγχος υποθέσεων (hypothesis testing) είναι μια συμπερασματική διαδικασία/μέθοδος που προσφέρει η Στατιστική Συμπερασματολογία και βρίσκει εφαρμογή σε στοχαστικά προβλήματα απόφασης μεταξύ δύο εναλλακτικών υποθέσεων. Η μία υπόθεση έχει επικρατήσει να συμβολίζεται με H_0 και ονομάζεται μηδενική υπόθεση (null hypothesis), και η άλλη με H_1 και ονομάζεται εναλλακτική υπόθεση (alternative hypothesis).

2.2.1 Στατιστικοί έλεγχοι υποθέσεων για τη μέση τιμή, μ, ενός πληθυσμού

Η γενική θεωρία στατιστικών ελέγχων γίνεται κυρίως με βάση την κανονική κατανομή. Αυτό συμβαίνει διότι, για μεγάλο μέγεθος δείγματος ν ($\nu \to \infty$), η δειγματική κατανομή του τυποποιημένου δειγματικού μέσου προσεγγίζει την τυποποιημένη κανονική. Έτσι, η εφαρμοσιμότητα των αποτελεσμάτων είναι αρκετά ευρεία και η κανονική κατανομή καλύπτει τις περισσότερες ενδιαφέρουσες περιπτώσεις.

<u>Βασικές έννοιες</u>

1) Επίπεδο σημαντικότητας

Ο έλεγχος μίας στατιστικής υπόθεσης στηρίζεται στην αρχή σύμφωνα με την οποία πολύ απίθανα ενδεχόμενα θεωρούνται αδύνατα και πολύ πιθανά ενδεχόμενα θεωρούνται βέβαια. Η μέγιστη τιμή του σφάλματος να απορρίψουμε την H_0 όταν η H_0 είναι αληθής (Σφάλμα τύπου 1) είναι γνωστή ως επίπεδο σημαντικότητας και συμβολίζεται με α.

Αν ορίσουμε επίπεδο σημαντικότητας 1%, αυτό σημαίνει ότι σε ένα δείγμα από τα 100 έχουμε πιθανότητα να απορρίψουμε την ορθή υπόθεση H_0. Με άλλα λόγια αυτό σημαίνει ότι είμαστε 99% βέβαιοι ότι η απόφαση μας να απορρίψουμε την H_0 είναι ορθή.

Το επίπεδο σημαντικότητας πάντοτε ορίζεται πριν τη συλλογή των πληροφοριών του δείγματος.

2) Κρίσιμη τιμή

Έστω z_1, z_2, …, $z_κ$ οι τιμές μίας στατιστικής, που χρησιμοποιούνται για να ελέγξουμε κάποια υπόθεση H_0. Μερικές τιμές οδηγούν στην απόρριψη της H_0, ενώ άλλες οδηγούν στη μη απόρριψη της H_0.
Οι τιμές που οδηγούν στην απόρριψη της H_0 μας δίνουν μία περιοχή η οποία ονομάζεται κρίσιμη περιοχή.
Η τιμή της στατιστικής που χωρίζει την κρίσιμη περιοχή από την περιοχή μη απόρριψης ονομάζεται κρίσιμη τιμή.

- **Έλεγχος με γνωστή διασπορά**

Έστω $X_1, X_2, \dots, X_\nu$ ένα τυχαίο δείγμα από έναν πληθυσμό που ακολουθεί την κανονική κατανομή $N(\mu,\sigma^2)$ με γνωστή διασπορά σ^2 και μέση τιμή $\mu=\mu_0$. Από τη θεωρία πιθανοτήτων γνωρίζουμε ότι στην περίπτωση αυτή, η κατανομή της τυχαίας μεταβλητής $\bar{X} = \frac{X_1+X_2+\cdots+X_\nu}{\nu}$, ανεξαρτήτως του μεγέθους του δείγματος, είναι κανονική κατανομή με $\bar{X} \sim N(\mu_0, \frac{\sigma^2)}{\nu}$ και επομένως $\frac{(\bar{X}-\mu_0)\sqrt{\nu}}{\sigma} = Z \sim N(0,1)$.

(α) Για τον μονόπλευρο έλεγχο της

$$H_0: \mu \leq \mu_0 \text{ έναντι της } H_1: \mu > \mu_0,$$

απορρίπτουμε την H_0 όταν

$$\bar{X} \geq \mu_0 + \frac{\sigma}{\sqrt{\nu}} z_\alpha$$

όπου z_α το α-άνω ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

(β) Για τον μονόπλευρο έλεγχο της

$$H_0: \mu \geq \mu_0 \text{ έναντι της } H_1: \mu < \mu_0,$$

απορρίπτουμε την H_0 όταν

$$\bar{X} \leq \mu_0 - \frac{\sigma}{\sqrt{\nu}} z_\alpha$$

(γ) Για τον δίπλευρο έλεγχο της

$$H_0: \mu = \mu_0 \text{ έναντι της } H_1: \mu \neq \mu_0,$$

απορρίπτουμε την H_0 όταν

$$|\bar{X} - \mu_0| \geq \frac{\sigma}{\sqrt{\nu}} z_{\alpha/2}$$

όπου $z_{\alpha/2}$ το α/2-άνω ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής. [6]

- **Έλεγχος με άγνωστη διασπορά**

Έστω $X_1, X_2, \dots, X_\nu$ ένα τυχαίο δείγμα από έναν πληθυσμό που ακολουθεί την κανονική κατανομή N(μ,σ^2) με γνωστή διασπορά σ^2 και μέση τιμή $\mu=\mu_0$. Επειδή η διασπορά του πληθυσμού, σ^2, είναι άγνωστη, δε μπορούμε ως στατιστική συνάρτηση ελέγχου να χρησιμοποιήσουμε την $Z = \frac{(\bar{X}-\mu_0)\sqrt{\nu}}{\sigma}$ γιατί δε μπορούμε να υπολογίσουμε την τιμή της, z, από το δείγμα. Γι αυτό, εκτιμάμε την άγνωστη διασπορά, σ^2, από τη δειγματική διασπορά $S^2 = \frac{1}{\nu-1}\sum_{i=1}^{\nu}(X_i - \bar{X})^2$ και ως στατιστική συνάρτηση ελέγχου, χρησιμοποιούμε την $T = \frac{(\bar{X}-\mu_0)\sqrt{\nu}}{S}$, η οποία, είναι γνωστό ότι όταν $X_i \sim N(\mu_0, \sigma^2), i = 1,2, \dots \nu,$ και ανεξαρτήτως του μεγέθους του δείγματος, ακολουθεί την κατανομή Student με ν-1 βαθμούς ελευθερίας. Δηλαδή $T = \frac{(\bar{X}-\mu_0)\sqrt{\nu}}{S} \sim t_{\nu-1}$.

(α) Για τον μονόπλευρο έλεγχο της

$$H_0: \mu \leq \mu_0 \text{ έναντι της } H_1: \mu > \mu_0,$$

απορρίπτουμε την H_0 όταν

$$\bar{X} \geq \mu_0 + \frac{S}{\sqrt{\nu}} t_{\nu-1;\alpha}$$

όπου $t_{\nu-1;\alpha}$ το α-άνω ποσοστιαίο σημείο της $t_{\nu-1}$ και

$S = \sqrt{\frac{1}{\nu-1}\sum_{i=1}^{\nu}(X_i - \bar{X})^2}$ η δειγματική διασπορά.

(β) Για τον μονόπλευρο έλεγχο της

$$H_0: \mu \geq \mu_0 \text{ έναντι της } H_1: \mu < \mu_0,$$

απορρίπτουμε την H_0 όταν

$$\bar{X} \leq \mu_0 - \frac{S}{\sqrt{\nu}} t_{\nu-1;\alpha}$$

(γ) Για τον δίπλευρο έλεγχο της

$$H_0: \mu = \mu_0 \text{ έναντι της } H_1: \mu \neq \mu_0,$$

απορρίπτουμε την H_0 όταν

$$|\bar{X} - \mu_0| \geq \frac{\sigma}{\sqrt{\nu}} t_{\nu-1;\alpha/2}$$

όπου $t_{\nu-1;\alpha/2}$ το α/2-άνω ποσοστιαίο σημείο της $t_{\nu-1}$. [6]

2.2.2 Στατιστικοί έλεγχοι υποθέσεων για τη διαφορά των μέσων δύο πληθυσμών

Μία άλλη σημαντική εφαρμογή στους ελέγχους υπόθεσης έχουμε όταν πρόκειται να συγκρίνουμε δύο ανεξάρτητους πληθυσμούς Π_1, Π_2 με βάση δύο αντίστοιχα ανεξάρτητα δείγματα

$$X_{1,}X_2, \dots, X_{\nu 1} \sim F_1 \text{ και } Y_{1,}Y_2, \dots, Y_{\nu 2} \sim F_2$$

όπου F_1 και F_2 οι σ.κ. των Π_1 και Π_2.

- **Έλεγχος για τη διαφορά δύο μέσων με γνωστές διασπορές**

Έστω τα ανεξάρτητα δείγματα

$$X_{1,}X_2, \dots, X_{\nu 1} \sim N(\mu_1, \sigma_1^2) \text{ και } Y_{1,}Y_2, \dots, Y_{\nu 2} \sim N(\mu_2, \sigma_2^2)$$

όπου $\sigma_1^2 > 0,\ \sigma_2^2 > 0, \sigma_1^2, \sigma_2^2$ γνωστά.

(α) Για τον έλεγχο της

$$H_0: \mu_1 - \mu_2 \leq \delta_0 \text{ έναντι της } H_1: \mu_1 - \mu_2 > \delta_0$$

απορρίπτουμε την H_0 όταν

$$\bar{X} - \bar{Y} \geq \delta_0 + \sqrt{\frac{\sigma_1^2}{\nu_1} + \frac{\sigma_2^2}{\nu_2} z_\alpha}.$$

(β) Για τον έλεγχο της

$$H_0: \mu_1 - \mu_2 \geq \delta_0 \text{ έναντι της } H_1: \mu_1 - \mu_2 < \delta_0$$

απορρίπτουμε την H_0 όταν

$$\bar{X} - \bar{Y} \leq \delta_0 - \sqrt{\frac{\sigma_1^2}{\nu_1} + \frac{\sigma_2^2}{\nu_2} z_\alpha}.$$

(γ) Για τον δίπλευρο έλεγχο της

$$H_0\colon \mu_1 - \mu_2 = \delta_0 \text{ έναντι της } H_1\colon \mu_1 - \mu_2 \neq \delta_0$$

απορρίπτουμε την H_0 όταν

$$|(\bar{X} - \bar{Y}) - \delta_0| \geq \sqrt{\frac{\sigma_1^2}{\nu_1} + \frac{\sigma_2^2}{\nu_2}}\, z_{\alpha/2}.$$

- **Έλεγχος για τη διαφορά δύο μέσων με άγνωστες διασπορές**

Έστω τα ανεξάρτητα δείγματα

$$X_{1,}X_2, \dots, X_{\nu 1} \sim N(\mu_1, \sigma^2) \text{ και } Y_{1,}Y_2, \dots, Y_{\nu 2} \sim N(\mu_2, \sigma^2)$$

όπου $\sigma^2 > 0$ άγνωστη. Θέτουμε

$$S_1^2 = \frac{1}{\nu_1 - 1}\sum_{i=1}^{\nu_1}(X_i - \bar{X})^2\,,\ S_2^2 = \frac{1}{\nu_2 - 1}\sum_{j=1}^{\nu_2}(Y_j - \bar{Y})^2$$

και

$$S_p^2 = \frac{1}{\nu_1 + \nu_2 - 2}\left(\sum_{i=1}^{\nu_1}(X_i - \bar{X})^2 + \sum_{j=1}^{\nu_2}\left(Y_j - \bar{Y}\right)^2\right)$$

(α) Για τον έλεγχο της

$$H_0\colon \mu_1 - \mu_2 \leq \delta_0 \text{ έναντι της } H_1\colon \mu_1 - \mu_2 > \delta_0$$

απορρίπτουμε την H_0 όταν

$$\bar{X} - \bar{Y} \geq \delta_0 + S_p\sqrt{\frac{1}{\nu_1} + \frac{1}{\nu_2}}\, t_{\nu_1 + \nu_2 - 2;a}.$$

(β) Για τον έλεγχο της

$$H_0: \mu_1 - \mu_2 \geq \delta_0 \text{ έναντι της } H_1: \mu_1 - \mu_2 < \delta_0$$

απορρίπτουμε την H_0 όταν

$$\bar{X} - \bar{Y} \leq \delta_0 - S_p \sqrt{\frac{1}{\nu_1} + \frac{1}{\nu_2} t_{\nu_1 + \nu_2 - 2; a}}.$$

(γ) Για τον δίπλευρο έλεγχο της

$$H_0: \mu_1 - \mu_2 = \delta_0 \text{ έναντι της } H_1: \mu_1 - \mu_2 \neq \delta_0$$

απορρίπτουμε την H_0 όταν

$$|(\bar{X} - \bar{Y}) - \delta_0| \geq S_p \sqrt{\frac{1}{\nu_1} + \frac{1}{\nu_2} t_{\nu_1 + \nu_2 - 2; a/2}}.$$

2.3 Έλεγχος υποθέσεων κατανομής με τεστ Kolmogorov-Smirnov

Δίνεται ένα δείγμα $X_1, X_2, \ldots, X_\nu$ από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με συνάρτηση κατανομής F. Θεωρούμε το πρόβλημα ελέγχου :

$$H_0\colon F = F_0 \text{ έναντι } H_1\colon F \neq F_0$$

όπου F_0 είναι κάποια συγκεκριμένη συνάρτηση κατανομής.

Η H_0 μπορεί να ελεγχθεί χρησιμοποιώντας το στατιστικό Kolmogorov-Smirnov:

$$D_\nu = sup|F_\nu(x) - F(x)|$$

όπου F_ν είναι η εμπειρική συνάρτηση κατανομής του δείγματος.

- Αν $D_\nu < D_{\nu,\alpha}$, όπου $D_{\nu,\alpha}$ είναι σταθερές, οι ακριβείς τιμές των οποίων καθορίζονται από το πλήθος του δείγματος και από το επίπεδο σημαντικότητας α, τότε αποδεχόμαστε την H_0.
- Αν $D_\nu \geq D_{\nu,\alpha}$, τότε απορρίπτεται η H_0 και γίνεται δεκτή η εναλλακτική της H_1.

Ο έλεγχος Kolmogorov-Smirnov είναι μία μη παραμετρική δοκιμή, για την ισότητα συνεχών, μονοδιάστατων κατανομών πιθανότητας. Η στατιστική Kolmogorov-Smirnov ποσοτικοποιεί την απόσταση μεταξύ της εμπειρικής συνάρτησης κατανομής του δείγματος και της αθροιστικής συνάρτησης κατανομής της κατανομής αναφοράς.

2.4 Έλεγχος του Bonferroni

Ο έλεγχος ή αλλιώς διόρθωση του Bonferroni βασίζεται σε απλά t tests, τα οποία γίνονται μεταξύ όλων των πληθυσμιακών μέσων τιμών αλλά με τροποποιημένο επίπεδο σημαντικότητας. Πραγματοποιώντας πολλαπλά t tests μεταξύ k πληθυσμιακών μέσων τιμών, αυξάνουμε σημαντικά τη συνολική πιθανότητα για εσφαλμένη απόρριψη της μηδενικής υπόθεσης σε έναν από τους ελέγχους. Με τη διόρθωση του Bonferroni μπορούμε να παρακάμψουμε το συγκεκριμένο πρόβλημα, καθιστώντας τους ελέγχους αυτούς πιο συντηρητικούς. Η λογική της διόρθωσης στοχεύει στην ελάττωση του επιπέδου σημαντικότητας των επιμέρους ελέγχων σε τέτοιο βαθμό, ώστε το συνολικό επίπεδο σημαντικότητας να διατηρείται κάτω από μία προκαθορισμένη τιμή.

Το επίπεδο σημαντικότητας που πρέπει να διασφαλιστεί σε κάθε επιμέρους έλεγχο, ώστε να «προστατεύεται» το συνολικό επίπεδο σημαντικότητας όλων των ελέγχων εξαρτάται από τον αριθμό των δυνατών συγκρίσεων που μπορούν να γίνουν. Όσο πιο μεγάλος είναι ο αριθμός των συγκρίσεων τόσο μικρότερο πρέπει να είναι το επίπεδο σημαντικότητας κάθε ελέγχου. Για να θέσουμε τη συνολική πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης π.χ. στο 0,05, πρέπει να χρησιμοποιήσουμε ως επίπεδο σημαντικότητας του κάθε επιμέρους ελέγχου τουλάχιστον το

$$\alpha^* = \frac{0{,}05}{m}$$

όπου m είναι ο αριθμός των δυνατών συγκρίσεων μεταξύ των k μέσων τιμών.

Το πρόβλημα με τη συγκεκριμένη μέθοδο είναι ότι, όταν ο αριθμός των συγκρίσεων αυξάνει, ελαττώνεται σημαντικά η ισχύς του συνολικού ελέγχου, με αποτέλεσμα να μην μπορεί να τεκμηριώσει σημαντικές διαφορές και σε περιπτώσεις όπου αυτές πράγματι υπάρχουν. Και αυτό διότι, σε μία τέτοια περίπτωση, το συνολικό επίπεδο σημαντικότητας με βάση το οποίο θα γίνουν όλοι οι έλεγχοι, γίνεται πλέον πολύ μικρό και επομένως δύσκολα μπορεί να απορριφθούν οι μηδενικές υποθέσεις των επιμέρους ελέγχων.

3. ΜΙΑ ΝΕΑ ΜΕΘΟΔΟΣ ΑΝΑΓΝΩΡΙΣΗΣ ΓΡΑΦΕΩΝ

3.1 Εισαγωγή

Τα αρχαία ελληνικά κείμενα σε πάπυρο αποτελούν μία από τις πιο σημαντικές πηγές πληροφορίας για τον αρχαίο κόσμο και ιδιαίτερα για την αρχαία Ελλάδα. Υπάρχει μία τεράστια συλλογή παπυρολογικών κειμένων που έχουν βρεθεί στην Αθήνα και στην Αττική και η οποία μας δίνει άμεση πρόσβαση στο παρελθόν μας. Οι κυριότερες δυσκολίες που καλούμαστε να αντιμετωπίσουμε στην μελέτη των κειμένων αυτών είναι ότι, κατά κανόνα, είναι ανυπόγραφα, χωρίς ημερομηνία, καθώς οι γραφείς τους, ούτε υπέγραφαν ούτε έβαζαν ημερομηνία στα κείμενά τους. Συνεπώς, οι παραπάνω δυσκολίες καθιστούν τη χρονολογική κατάταξη των κειμένων εξαιρετικά δύσκολη και συχνά αδύνατη. Παρ' όλα αυτά, η δυνατότητα να δώσουμε στα παπυρολογικά κείμενα το σωστό ιστορικό πλαίσιο, είναι κρίσιμης σημασίας ώστε να ξεκλειδώσουμε τις χρήσιμες πληροφορίες που περιέχουν.

Η αναγνώριση του γραφέα που έγραψε το κείμενο επιτρέπει τη σωστή και σαφή χρονολόγησή του. Για παράδειγμα, κατά τη διάρκεια της κλασικής εποχής στην Αττική, ο κάθε γραφέας είχε μία επαγγελματική σταδιοδρομία περιορισμένης διάρκειας. Ας υποθέσουμε ότι μπορούμε, μέσω της επεξεργασίας εικόνας και αναγνώρισης προτύπων, να προσδιορίσουμε τον γραφικό χαρακτήρα ενός συγκεκριμένου γραφέα και ότι τουλάχιστον ένα από τα κείμενα αυτού του γραφέα μπορεί να χρονολογηθεί με βάση το περιεχόμενό του ή/και αρχαιολογικά επιχειρήματα. Τότε κάθε άλλο κείμενο που μπορεί να εξακριβωθεί ότι είναι του ίδιου γραφέα, αποκτά χρονολογία.

Μέχρι σήμερα η χρονολογική κατάταξη των κειμένων γινόταν με βάση συγκριτικές χρονολογήσεις, το «ένστικτο» και την εν γένει εμπειρία των αρχαιολόγων. Οι τρόποι αυτοί προσφέρουν αποτελέσματα, πλην όμως κανείς εξ αυτών δε δίνει με συνέπεια, αντικειμενικά και αδιαμφισβήτητα συμπεράσματα. Επομένως, η ανάπτυξη ενός πληροφοριακού συστήματος που θα εκτελεί την αυτόματη αναγνώριση των γραφέων και την κατάταξη των κειμένων με βάση το χέρι που τα έγραψε, είναι ιδιαιτέρως σημαντική για την ιστορία και την αρχαιολογία με τελικό στόχο την αντικειμενική κατάταξη, ομαδοποίηση και χρονολόγηση των επιγραφών.

Από τη στιγμή που θα επιτευχθεί η ομαδοποίηση των επιγραφών ανά γραφέα είναι πολύ πιο εύκολο, πρακτικά σχεδόν σίγουρο, να μπορεί να συναχθεί η περίοδος που έζησε ο γραφέας από το συνδυασμό του περιεχομένου των κειμένων που έγραψε. Στο πλαίσιο αυτής της προσπάθειας, έχουν εξαχθεί μαθηματικοποιημένα κριτήρια τα οποία, σε αντίθεση με τα όσα εφαρμόζονται μέχρι σήμερα, δίνουν ποσοτικά αποτελέσματα, τα οποία με τη σειρά τους οδηγούν σε ταυτοποίηση του γραφέα και κατά συνέπεια ομαδοποίηση των επιγραφών. Τα κριτήρια αυτά οδήγησαν στην ανάπτυξη πρωτότυπων αλγορίθμων αναγνώρισης προτύπων, μέσω των οποίων έχει επετευχθεί ταυτοποίηση των γραφέων ικανού αριθμού παπυρολογικών κειμένων με πλήρη επιτυχία. Μερικές από τις βασικότερες δυσκολίες που μπορεί να αντιμετωπίσει κάποιος στην ανάπτυξη του συστήματος της αυτόματης αναγνώρισης γραφέων είναι :

(1) Η ουσιώδης μεταβλητότητα του σχήματος του κάθε γράμματος μέσα στο ίδιο κείμενο, του ίδιου γραφέα.

(2) Είναι αρκετά σύνηθες, η ομοιότητα μεταξύ δύο κειμένων του ίδιου γραφέα να είναι μικρότερη από την ομοιότητα μεταξύ διαφορετικών κειμένων από διαφορετικούς γραφείς.

(3) Τα κείμενα έχουν υποστεί σημαντική φθορά.

(4) Συμπληρωματικά, επισημαίνουμε ότι υπάρχει σημαντική μεταβλητότητα στην αντίληψη της μορφής του κάθε συμβόλου της αλφαβήτου στον τρόπο με τον οποίο παράγεται από τον ίδιο γραφέα. Για παράδειγμα, πολύ συχνά, ο ίδιος γραφέας ενώνει την μεσαία γραμμή του «Α» μόνο με το αριστερό πόδι, άλλες φορές μόνο με το δεξί, ενώ άλλες φορές δεν την ενώνει με κανένα πόδι. Η μεταβλητότητα αυτή δεν αφορά μόνο τη μεσαία γραμμή του «Α», αλλά επεκτείνεται και στη σχετική θέση των ποδιών, στη μορφή της κορυφής του «Α», κλπ. Ανάλογες διαφορές προκύπτουν και για τα υπόλοιπα γράμματα της αλφαβήτου.

3.2 Σύντομη περιγραφή της εφαρμοζόμενης μεθόδου

Απώτερος σκοπός του εγχειρήματος της ανάπτυξης συστήματος αυτόματης αναγνώρισης γραφέα είναι να επιτευχθεί η ταυτοποίηση του γραφέα στηριζόμενη μόνο σε μαθηματική επεξεργασία και μεθόδους αναγνώρισης προτύπων στα γράμματα του κάθε κειμένου. Η μεθοδολογία που ακολουθείται αποτελείται από τα παρακάτω βήματα :

1. Τα εδάφια της Αθηναίων Πολιτείας φωτογραφήθηκαν σύμφωνα με ένα αυστηρό πρωτόκολλο.
2. Γίνεται κατάτμηση των ληφθέντων εικόνων και εξάγονται τα περιγράμματα καθενός γράμματος του κάθε κειμένου.
3. Για κάθε γράμμα εντοπίζεται ένας αριθμός από κρίσιμα σημεία.
4. Υποθέτουμε ότι ο γραφέας είχε ένα ιδανικό πρότυπο για κάθε γράμμα του αλφαβήτου, του οποίου το περίγραμμα αντιστοιχεί στην καμπύλη $\overrightarrow{r^M}(t)$. Λόγω της αστάθειας του χεριού του γραφέα, της τυχαίας αλληλεπίδρασης μεταξύ πάπυρου και πένας, των διαφορετικών στάσεων του γραφέα και των εναλλαγών της διάθεσής του, της ποικιλίας των χρησιμοποιούμενων εργαλείων γραφής και του παπύρου, ο γραφέας τελικά έγραψε την καμπύλη $\overrightarrow{r^M}(t) + \vec{n}(t)$, όπου $\vec{n}(t)$ αντιπροσωπεύει τον συσχετιζόμενο θόρυβο. Έχει αναπτυχθεί μία καινούρια μέθοδος καταστολής του θορύβου $\vec{n}(t)$, από διαφορετικές πραγματοποιήσεις του ίδιου γράμματος σε ένα κείμενο. Αυτή η μέθοδος συνίσταται στην περιστροφή, μετάφραση και προσαρμογή του μεγέθους όλων των εμφανίσεων του κάθε γράμματος μέσα στο κείμενο, ώστε να ταιριάξουν βέλτιστα σύμφωνα με ένα εφαρμοζόμενο κριτήριο. Υπολογίζοντας καταλλήλως τον μέσο όρο των ταιριασμένων περιγραμμάτων, η παράγεται η υλοποίηση του ιδανικού προτύπου για κάθε γράμμα και για κάθε κείμενο χωριστά. Αυτή η υλοποίηση ονομάζεται ιδανική ή «πλατωνική» υλοποίηση του γράμματος της αλφαβήτου.
5. Όλα τα κείμενα συγκρίνονται ανά ζεύγη χρησιμοποιώντας αυτές τις ιδανικές υλοποιήσεις και τις μεμονωμένες υλοποιήσεις του κάθε γράμματος. Υποθέτουμε ότι αν δύο κείμενα έχουν γραφτεί από

διαφορετικούς γραφείς, τότε θα υπάρχει ένας σημαντικός αριθμός γραμμάτων, των οποίων οι πρότυπες υλοποιήσεις θα διαφέρουν σημαντικά. Από την άλλη πλευρά, οι πλατωνικές πρότυπες υλοποιήσεις του ίδιου γράμματος διαφορετικών κειμένων του ίδιου γραφέα θα πρέπει να παρουσιάζουν ουσιώδεις ομοιότητες. Πρωτότυπα κριτήρια έχουν αναπτυχθεί για την ποσοτικοποίηση αυτών των παρατηρήσεων.

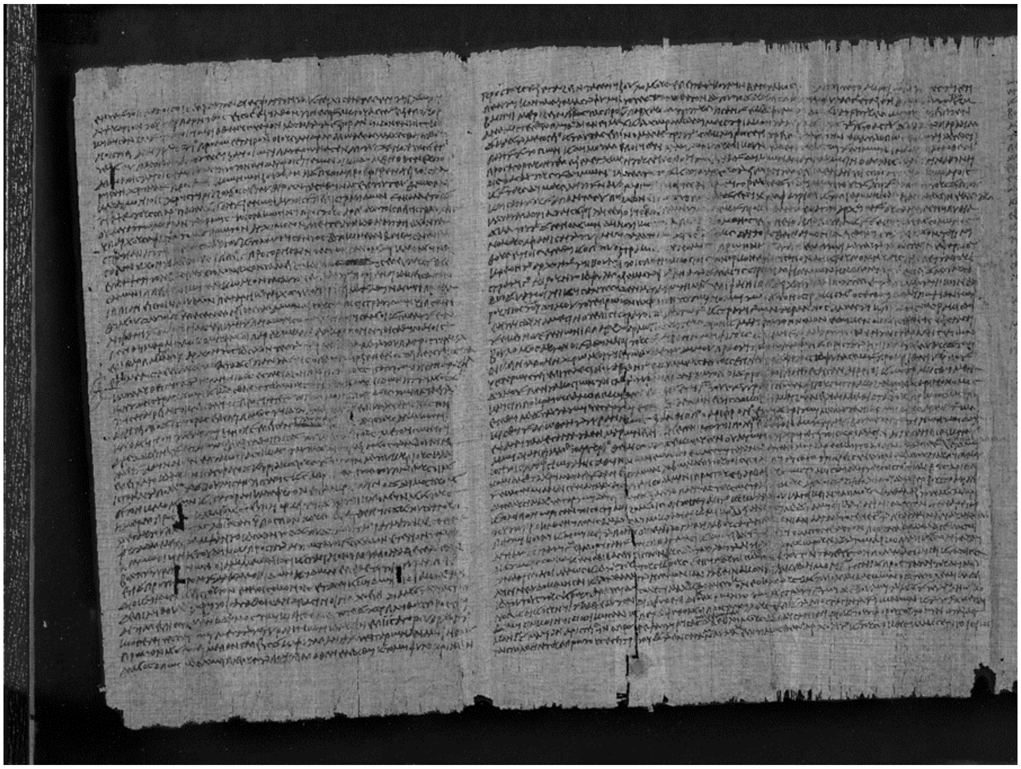
6. Υποθέτουμε επιπλέον ότι όλες οι υλοποιήσεις των γραμμάτων ενός κειμένου ταιριάζουν βέλτιστα με το ιδανικό πρότυπου του ίδιου του γραφέα πρώτα και ενός άλλου γραφέα ακολούθως. Θα περιμέναμε ότι οι δύο περιπτώσεις των ταιριασμένων περιγραμμάτων θα διαφέρουν σημαντικά για έναν τουλάχιστον αριθμό γραμμάτων. Αν τα γράμματα, αλλά και η ιδανική υλοποίηση προέρχονται από τον ίδιο γραφέα, θα περιμέναμε ότι τα ταιριασμένα περιγράμματα θα ταιριάζουν περισσότερο στο πρότυπο. Ένα σύνολο κριτηρίων έχουν διατυπωθεί για να ποσοτικοποιηθούν αυτές οι παρατηρήσεις, ενώ κανένα δοκιμαστικό σύνολο κειμένων δεν έχει χρησιμοποιηθεί για την αναγνώριση του γραφέα.

3.3 Πρώτο στάδιο επεξεργασίας των γραμμάτων

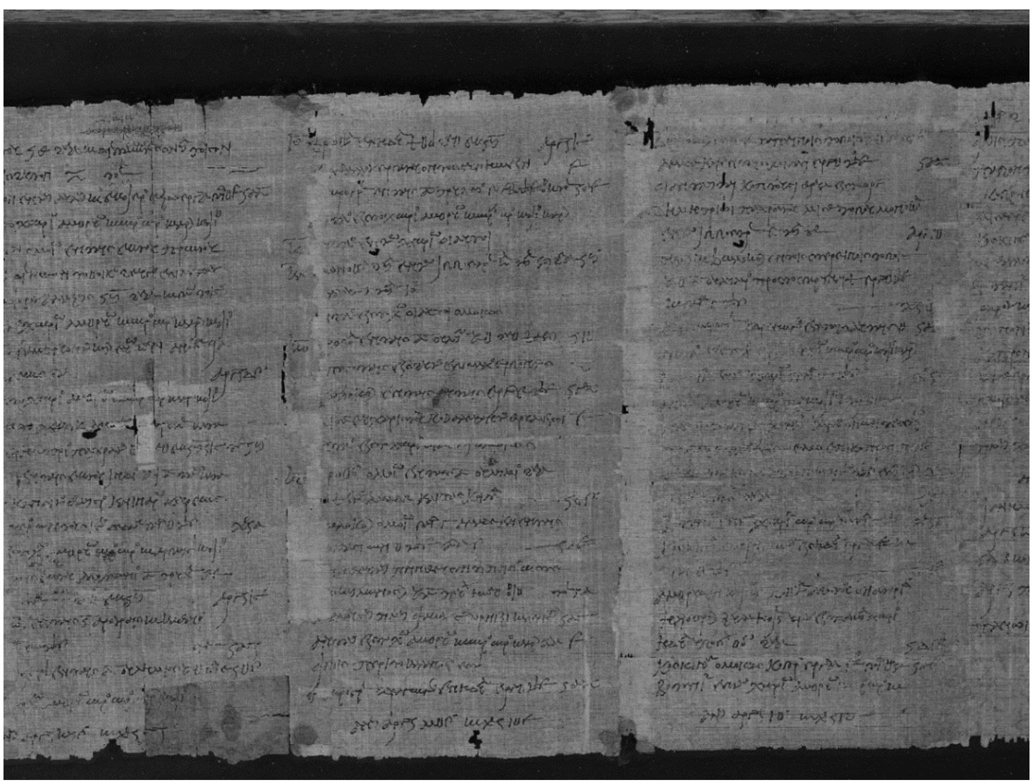
3.3.1 Αυτοματοποιημένη κατάτμηση της εικόνας για την εξαγωγή των γραμμάτων

Για να αποκτήσουμε ευδιάκριτα και ακριβή περιγράμματα του κάθε γράμματος εφαρμόζουμε διάφορες μεθόδους κατάτμησης εικόνας. Στη διάθεσή μας είχαμε ψηφιακές λήψεις από όλα τα προς μελέτη κείμενα, τα οποία προέρχονται από συγκεκριμένα εδάφια της «Αθηναίων Πολιτείας» του Αριστοτέλη. Για κάθε παπυρολογικό κείμενο πραγματοποιήσαμε διαφορετικές λήψεις. Κατ’ αρχήν γίνεται μία πανοραμική λήψη και εν συνεχεία πραγματοποιούνται επί μέρους φωτογραφήσεις πολύ υψηλής ανάλυσης, οι οποίες μάλιστα είναι και οι καταλληλότερες για την επεξεργασία των γραμμάτων που θα ακολουθήσει.

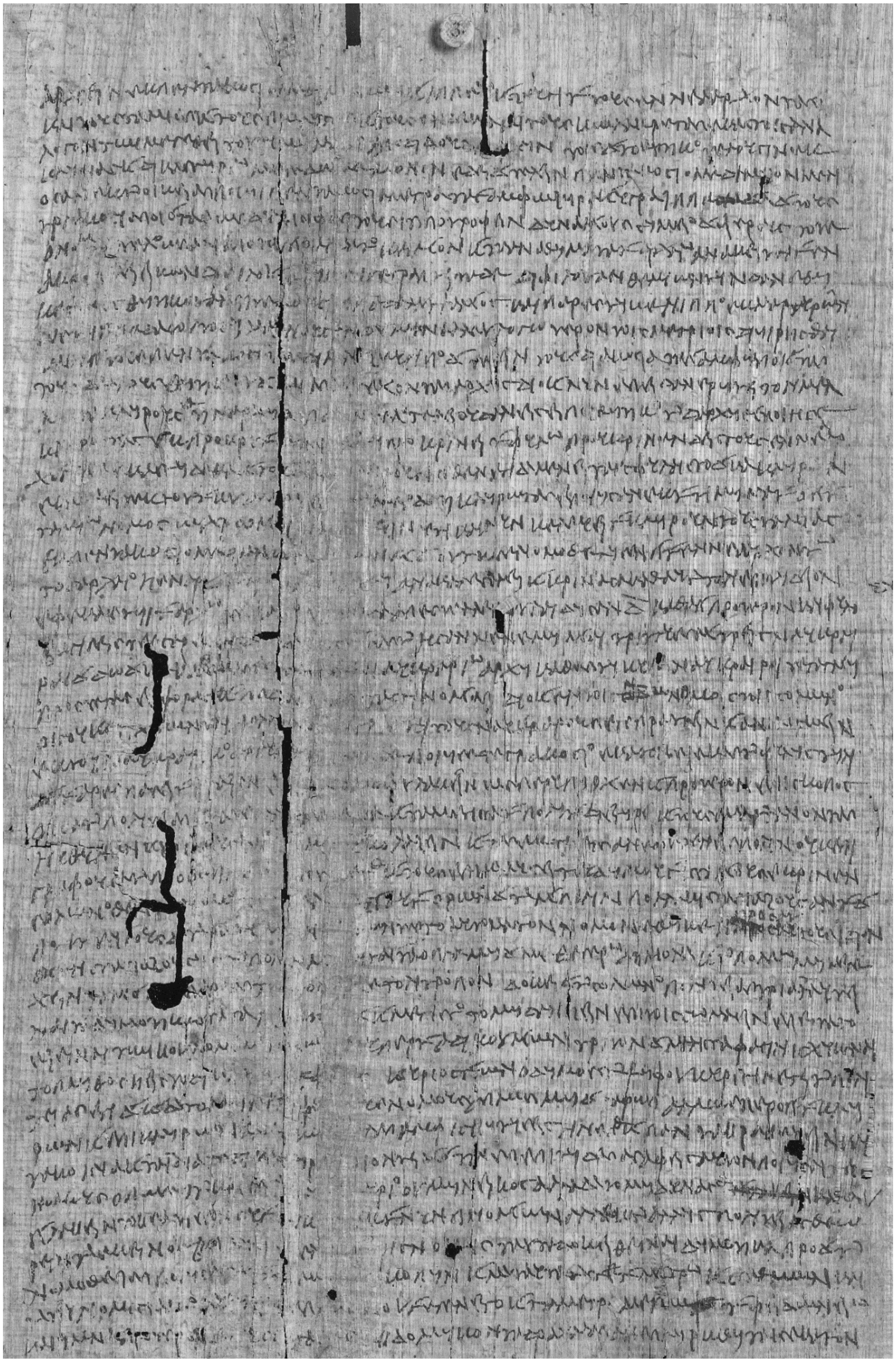
Παρακάτω ακολουθούν ορισμένες χαρακτηριστικές φωτογραφίες από τα προς μελέτη παπυρολογικά κείμενα. Στην συνέχεια, για την επεξεργασία των γραμμάτων γίνεται εξαγωγή τους από το υπόβαθρο του παπύρου. Η εφαρμογή των αλγορίθμων κατάτμησης έγινε αρχικά σε έγχρωμες φωτογραφίες, αλλά στο τέλος παρήχθησαν φωτογραφίες διαβάθμισης του γκρι.



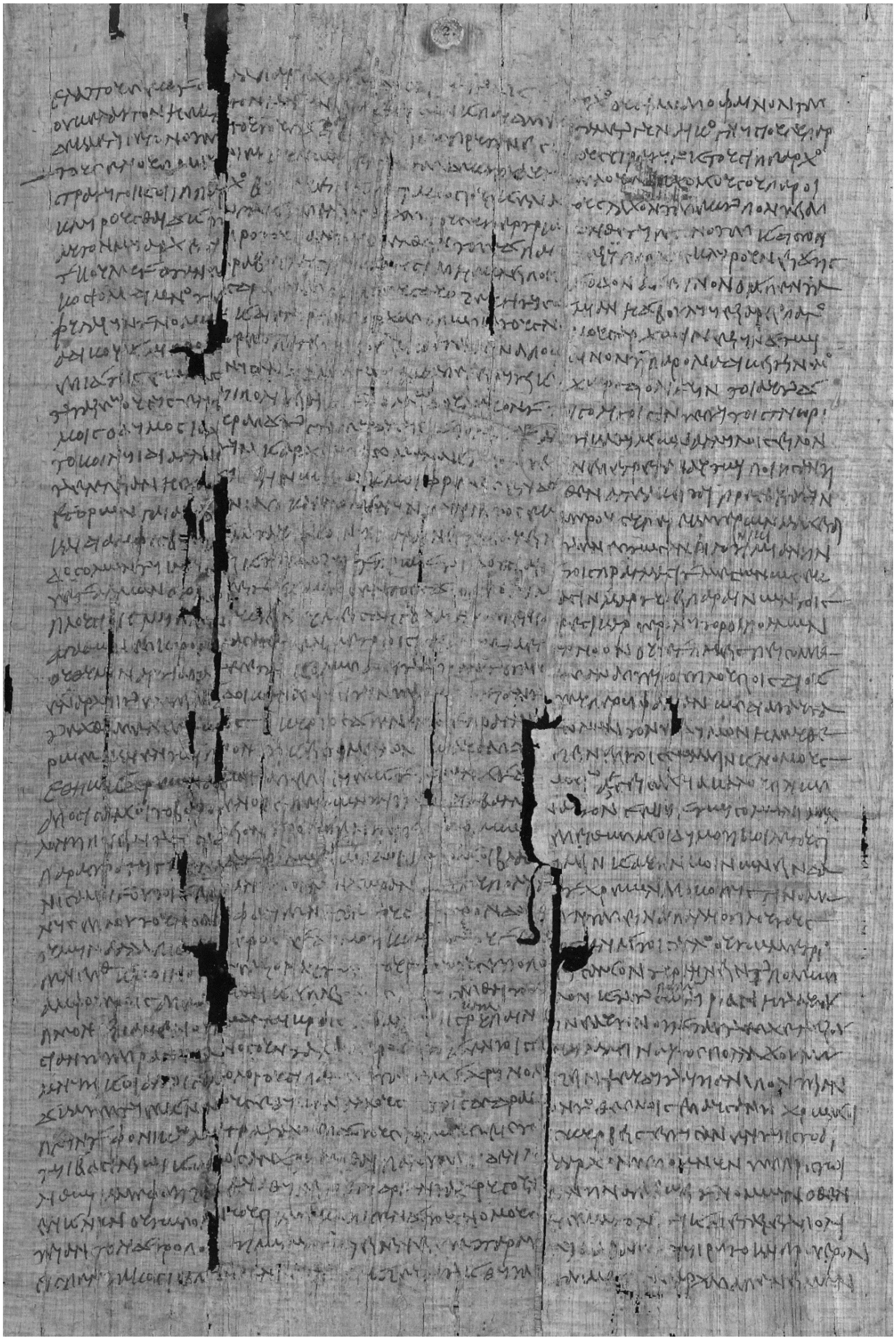
Εικόνα 1: Ψηφιακή φωτογραφία μέρους της «Αθηναίων Πολιτείας»



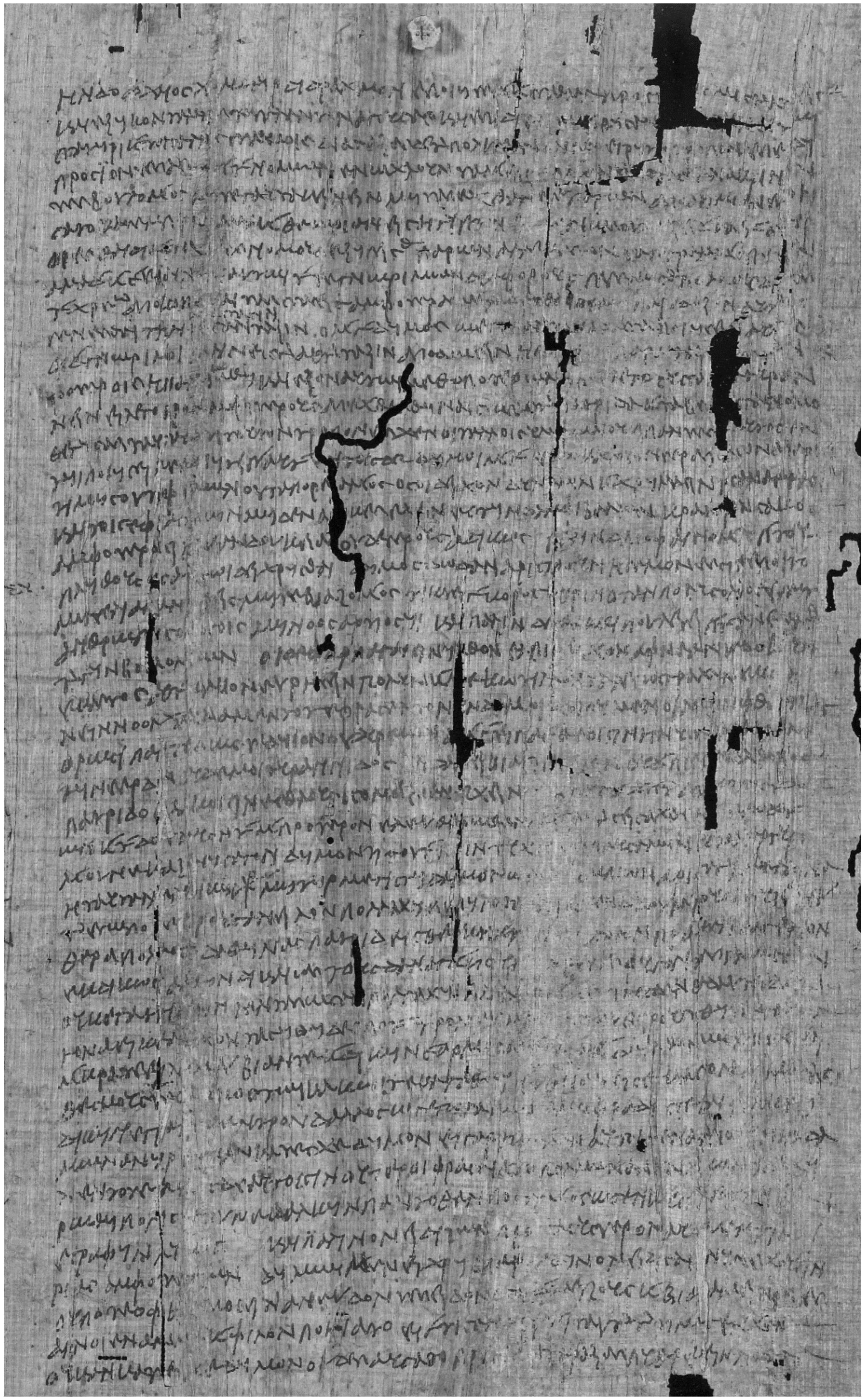
Εικόνα 2: Ψηφιακή φωτογραφία μέρους της «Αθηναίων Πολιτείας»



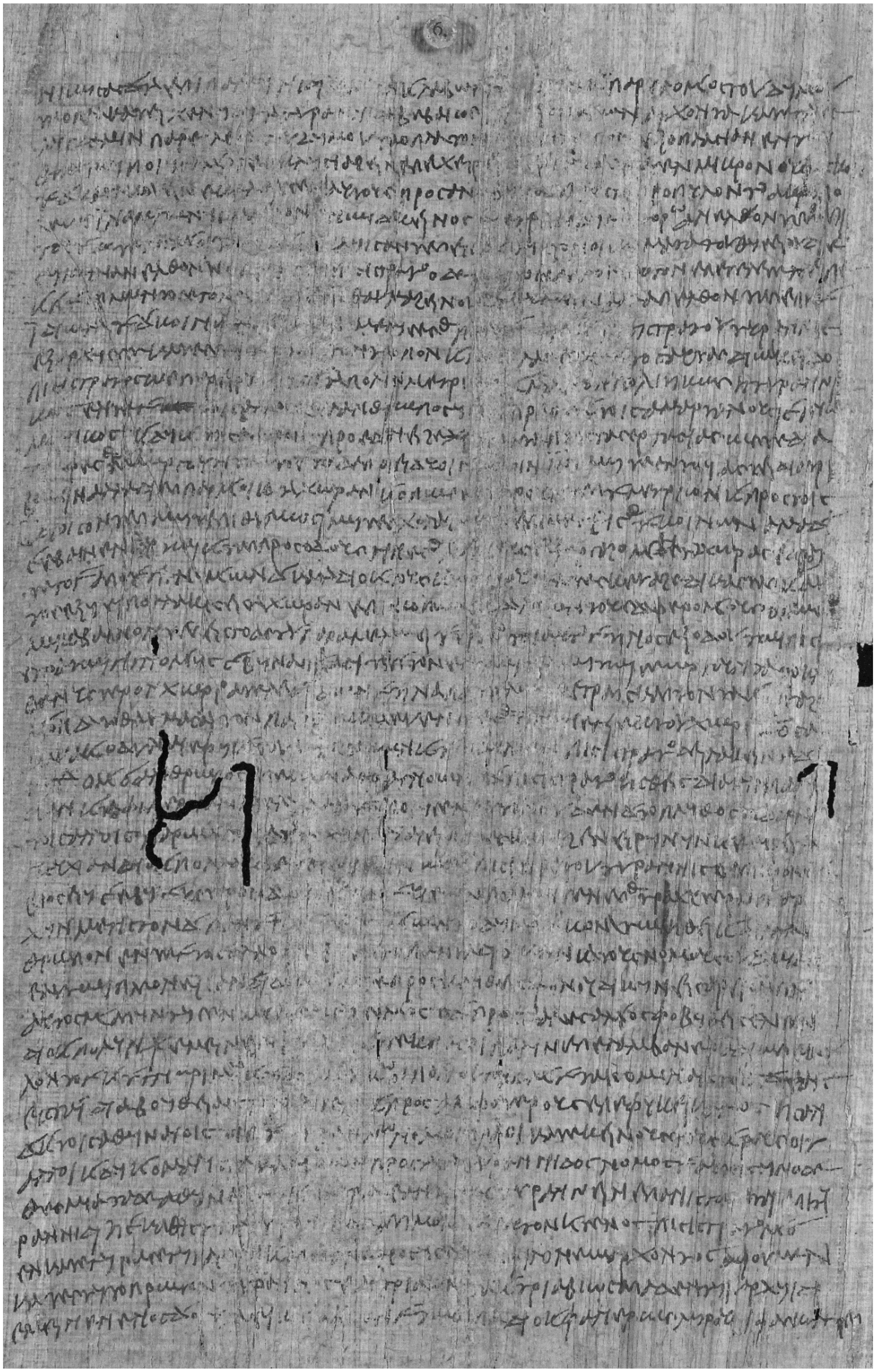
Εικόνα 3: Ψηφιακή φωτογραφία μέρους της «Αθηναίων Πολιτείας»



Εικόνα 4: Ψηφιακή φωτογραφία μέρους της «Αθηναίων Πολιτείας»



Εικόνα 5: Ψηφιακή φωτογραφία μέρους της «Αθηναίων Πολιτείας»



Εικόνα 6: Ψηφιακή φωτογραφία μέρους της «Αθηναίων Πολιτείας»

3.3.2 Αυτοματοποιημένη εξαγωγή περιγραμμάτων

Αρχικά γίνεται κατάτμηση της αρχικής ψηφιακής φωτογραφίας ώστε να απομονώσουμε την υλοποίηση του γράμματος, του οποίου θέλουμε να γίνει εξαγωγή του περιγράμματος, όπως φαίνεται στις εικόνες 7α, 8α.

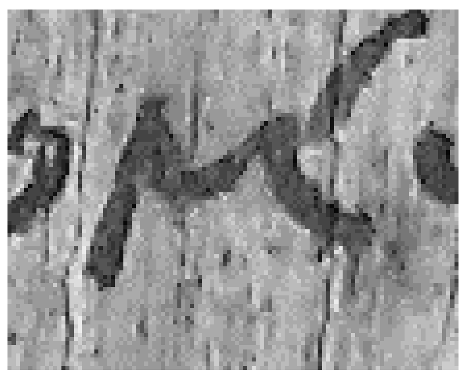
Κάθε περίγραμμα λαμβάνεται με τη βοήθεια της παρακάτω μεθοδολογίας :

- Η διαβάθμιση του χρώματος κάθε εικόνας κάθε γράμματος μετατρέπεται σε μαύρο και άσπρο.
- Το περίγραμμα του γράμματος εξάγεται και εξασφαλίζουμε ότι θα έχει τις παρακάτω ιδιότητες :
  (α) κάθε pixel του περιγράμματος πρέπει να έχει ακριβώς δύο γειτονικά pixels, (β) δεν επιτρέπονται απομονωμένα pixels και (γ) τρία pixels δεν πρέπει να σχηματίζουν ορθή γωνία.

Έχει αναπτυχθεί κατάλληλο λογισμικό που εξασφαλίζει την εξαγωγή περιγραμμάτων της παραπάνω μορφής. Το αποτέλεσμα της παραπάνω μεθοδολογίας φαίνεται στις εικόνες 7β, 8β.

Εικόνα 7α: υλοποίηση γράμματος «Α»

Εικόνα 7β: υλοποίηση γράμματος «Α» σε διαβάθμιση μαύρου και άσπρου



Εικόνα 8α: υλοποίηση γράμματος «Μ»

Εικόνα 8β: υλοποίηση γράμματος «Μ» σε διαβάθμιση μαύρου και άσπρου

3.3.3 Προσδιορισμός κρίσιμων σημείων στην υλοποίηση κάθε γράμματος

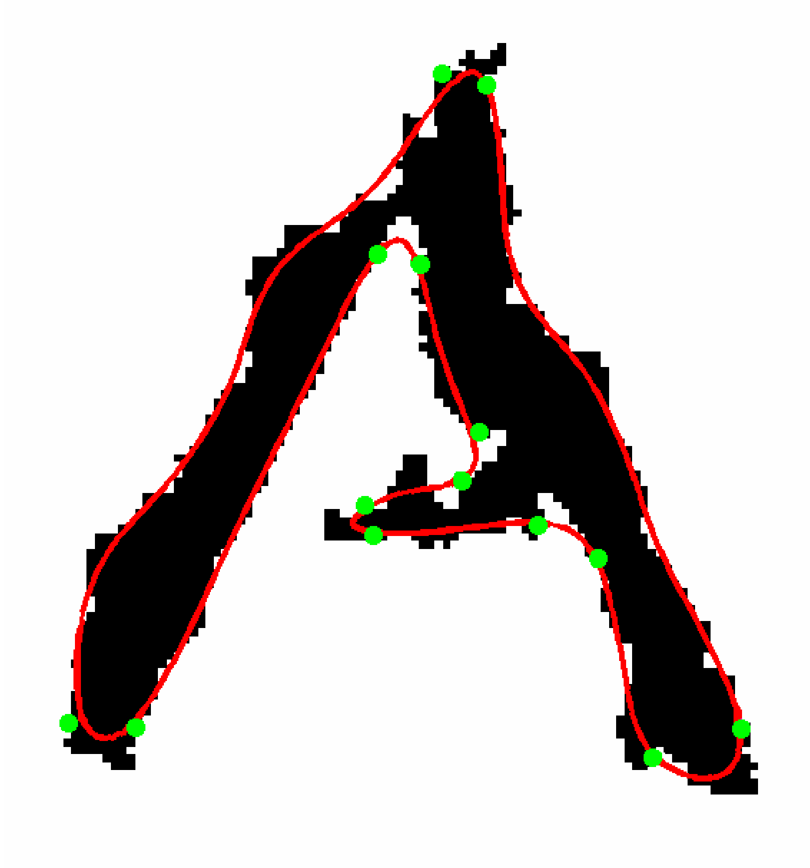
Προσεγγίζουμε το περίγραμμα του γράμματος με ελαφρώς επικαλυπτόμενα πολυώνυμα. Θεωρούμε ένα γράμμα ως ένα σύνολο pixels στο επίπεδο της εικόνας που αποτελείται από συνδεδεμένα υποσύνολα C_i. Λειαίνουμε το περίγραμμα κάθε συνδεδεμένου συνόλου ως εξής :

Για κάθε n-οστό pixel του C_i περιγράμματος δημιουργούμε μία αλυσίδα από διαδοχικά pixels P_n μήκους L_s ξεκινώντας από το n-οστό pixel. Για κάθε pixel της αλυσίδας P_n υπολογίζουμε κάθε απόσταση ξ_j από το αρχικό pixel της αλυσίδας. Στη συνέχεια υπολογίζουμε τα πολυώνυμα τρίτου βαθμού της απόστασης ξ_j που προσεγγίζουν καλύτερα τις συντεταγμένες (x_j, y_j) του P_n. Δηλαδή, υπολογίζουμε τα $\alpha_1, \dots, \alpha_4, b_1, \dots, b_4$, τέτοια ώστε η ποσότητα

$$\sum_{j=1}^{L_s} \left\{ \left(x_j - a_1\xi_j^3 - a_2\xi_j^2 - a_1\xi_j - a_4\right)^2 + \left(y_j - b_1\xi_j^3 - b_2\xi_j^2 - b_1\xi_j - b_4\right)^2 \right\}$$

να είναι ελάχιστη.

Επαναλαμβάνουμε αυτή τη διαδικασία ξεκινώντας από όλα τα σημεία της αλυσίδας P_n. Συνεπώς, για κάθε pixel του C_i περιγράμματος εξασφαλίζουμε L_s προσεγγιστικά πολυώνυμα του συγκεκριμένου pixel του περιγράμματος. Υπολογίζοντας τον μέσο όρο αυτών των πολυωνύμων, αποκτούμε ένα τελικό αποτέλεσμα το οποίο μας εξασφαλίζει μια ομαλή προσέγγιση s_i, αυτού του περιγράμματος.



Εικόνα 9: Απεικόνιση της ομαλής προσέγγισης του περιγράμματος (με κόκκινο χρώμα) και των κρίσιμων σημείων (με πράσινο χρώμα)

Ακολούθως, προσεγγίζουμε το s_i με ελαφρώς επικαλυπτόμενα πολυώνυμα του ξ_j μεγάλου βαθμού όπως και πριν, για παράδειγμα ενδέκατου βαθμού, και υπολογίζουμε την καμπυλότητα αυτών των πολυωνύμων σε κάθε pixel του s_i. Όταν αυτή η καμπυλότητα ξεπερνάει ένα προκαθορισμένο όριο, για παράδειγμα την τριπλάσια τιμή του μέσου όρου της καμπυλότητας ολόκληρου του s_i, τότε συμπεραίνουμε ότι αυτό το σημείο είναι ένα κρίσιμο σημείο καμπής CT_i, του περιγράμματος του γράμματος. Τα πλησιέστερα στο CT_i σημεία στο περίγραμμα του γράμματος ορίζονται ως τα κρίσιμα σημεία του.

3.4 Καθορισμός και προσδιορισμός του «ιδεατού» προτύπου για κάθε γράμμα

Υποθέτουμε ότι ο γραφέας έχει στο μυαλό του ένα ιδεατό πρότυπο, ένα είδος «πλατωνικού» προτύπου που αναπαριστάται από την καμπύλη $x(t)\vec{\imath} + y(t)\vec{\jmath}$. Αντί για αυτό, ο γραφέας χαράσσει μία διαστρεβλωμένη εκδοχή του γράμματος του οποίου το περίγραμμα είναι η καμπύλη $(x(t) + n_x(t))\vec{\imath} + (y(t) + n_y(t))\vec{\jmath}$, όπου $n_x(t), n_y(t)$ αντιπροσωπεύουν το θόρυβο που οφείλεται στην αστάθεια στο χέρι του γραφέα, στην τυχαία αλληλεπίδραση μεταξύ πάπυρου και πένας, των διαφορετικών στάσεων του γραφέα και των εναλλαγών της διάθεσής του, της ποικιλίας των χρησιμοποιούμενων υλικών γραφής, κλπ. Μία πρώτη προσέγγιση προς την αναγνώριση γραφέα είναι η καταστολή του θορύβου $n_x(t), n_y(t)$ από διαφορετικές υλοποιήσεις του ίδιου γράμματος του κειμένου. Υποθέτουμε ότι αφού καταστείλουμε τον θόρυβο, τότε εξασφαλίζουμε μία εκτίμηση του ιδεατού προτύπου του γραφέα, το οποίο επίσης αποκαλούμε πραγματικό πρότυπο (πραγματικό πρότυπο ή ιδεατό αντιπρόσωπο). Για τον σκοπό αυτό, για κάθε γράμμα ορίζουμε τις πλευρές του γράμματος από διαδοχικά κρίσιμα σημεία. Για παράδειγμα, στο γράμμα Α τα δύο αριστερά πιο κρίσιμα σημεία ορίζουν την αριστερή πλευρά του αριστερού ποδιού, ενώ τα δύο δεξιότερα ορίζουν τη δεξιά πλευρά του δεξιού ποδιού, κοκ. Επομένως, το περίγραμμα κάθε γράμματος αποτελείται από έναν αριθμό πλευρών των οποίων η ένωση δεν καλύπτει το περίγραμμα απαραίτητα. Ωστόσο, η χρήση αυτών των πλευρών των γραμμάτων μας παρέχει μία αξιόπιστη αναγνώριση γραφέα στην προβλεπόμενη εφαρμογή μας.

3.4.1 Μία πρώτη προσέγγιση του «ιδεατού» προτύπου

Για να εξασφαλίσουμε ένα βαθμό καταστολής του θορύβου για να δημιουργήσουμε ένα αρχικό ιδεατό πρότυπο για κάθε γράμμα του κειμένου, έχουμε αναπτύξει έναν αλγόριθμο που αποτελείται από τα παρακάτω πέντε βήματα:

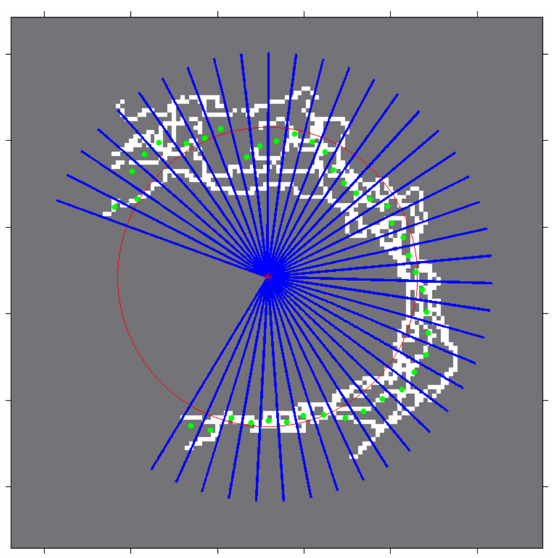
(1) Θεωρούμε ένα κείμενο E_1 και ένα γράμμα, για παράδειγμα το Μ, υλοποιήσεις του οποίου εμφανίζονται στο E_1. Το περίγραμμα Π_1 μίας υλοποίησης του γράμματος επιλέγεται τυχαία και τοποθετείται στο κέντρο ενός επιλεγμένου πλαισίου.

(2) Το περίγραμμα Π_2 μίας άλλης υλοποίησης του γράμματος Μ επιλέγεται τυχαία και του εφαρμόζουμε: (α) περιστροφή γύρω από το κέντρο του, (β) προσαρμογή μεγέθους, (γ) παράλληλη μετατόπιση, ώστε το Π_2 να ταιριάζει βέλτιστα με το Π_1, μέσω της ελαχιστοποίησης της ποσότητας

$$\varepsilon = \frac{\sum_{k=1}^{N_s} \sum_{i=1}^{N_{k,1}} d^2\left(P_i, \Pi_{2,k}\right) + \sum_{k=1}^{N_s} \sum_{j=1}^{N_{k,2}} d^2\left(Q_j, \Pi_{1,k}\right)}{2}$$

όπου N_s είναι ο αριθμός των πλευρών του γράμματος, $\Pi_{1,k}$ είναι η k-οστή πλευρά του Π_1 και $\Pi_{2,k}$ του Π_2, P_i, Q_j αυθαίρετα σημεία της k-οστής πλευράς των Π_1 και Π_2 αντίστοιχα, $N_{k,1}$ ο αριθμός των σημείων του $\Pi_{1,k}$ και $N_{k,2}$ του $\Pi_{2,k}$, $d^2\left(P_i, \Pi_{2,k}\right)$ είναι η τετραγωνική απόσταση του αυθαίρετου σημείου P_i από την καμπύλη Π_2 και $d^2\left(Q_j, \Pi_{1,k}\right)$ του Q_j από την Π_1. Αυτός ο τύπος έχει αποδειχθεί πιο αποτελεσματικός μεταξύ διαφόρων που έχουν δοκιμαστεί. Η συγκεκριμένη μορφή του αντίστοιχου κριτηρίου είναι απαραίτητη. Η άθροιση πρέπει να είναι αμοιβαία αφού αν λάβουμε την απόσταση μόνο του ενός περιγράμματος από το άλλο, τότε ο αλγόριθμος κάνει το άλλο περίγραμμα να συρρικνωθεί. Επίσης, το σφάλμα πρέπει να υπολογιστεί για κάθε πλευρά χωριστά, καθώς το σφάλμα υπολογίζεται μόνο μεταξύ αντίστοιχων πλευρών του κάθε γράμματος της αλφαβήτου.

(3) Μία «μέση καμπύλη» των Π_1 και Π_2, M_1, ορίζεται ως εξής: θεωρούμε τα Π_1 και Π_2 στις καλύτερα ταιριασμένες θέσεις $s_{i,1}$ και $s_{i,2}$ που έχουν βρεθεί στο βήμα (2). Υπολογίζουμε την μέση καμπυλότητα του συνδυασμού των $s_{i,1}$ και $s_{i,2}$ των δύο γραμμάτων μέσω του κύκλου που ταιριάζει καλύτερα σε αυτό το σύνολο. Στην συνέχεια, θεωρούμε το κέντρο αυτού του κύκλου ως σημείο παρατήρησης του συνόλου: κατασκευάζουμε την ελάχιστη γωνία Θ_i με κορυφή στο O, που περιλαμβάνει όλα τα pixels των $s_{i,1}$ και $s_{i,2}$. Εν συνεχεία, αν $\Theta_i \geq \frac{\pi}{6}$ διαιρούμε τη γωνία Θ_i σε N_i μικρότερες γωνίες $\delta\Theta_j = \kappa\sqrt{2}$ όπου κ η καμπυλότητα του κύκλου. Σε κάθε γωνία $\delta\Theta_j$ υπολογίζουμε τη μέση τιμή των συντεταγμένων των pixel που ανήκουν σε αυτή, βρίσκοντας ένα μοναδικό σημείο $A_{j,i}$ όπου έχει γίνει η προσέγγιση για έναν αυθαίρετο αριθμό πλευρών.



Εικόνα 10: Προσδιορισμός του σημείου O και των μικρότερων γωνιών $\delta\Theta_j$

Αν $\Theta_j \leq \frac{\pi}{6}$ υπολογίζουμε το ευθύγραμμο τμήμα ε_i που ταιριάζει καλύτερα στα $s_{i,1}$ και $s_{i,2}$ και το διαιρούμε σε N_i τμήματα $\delta\Pi_j$ μήκους $\sqrt{2}$. Για κάθε τμήμα $\delta\Pi_j$ σημειώνουμε τα pixels των $s_{i,1}$ και $s_{i,2}$ των οποίων οι προβολές στο ε_i βρίσκονται πάνω στο $\delta\Pi_j$ και υπολογίζουμε τη μέση τιμή των συντεταγμένων τους. Έτσι, βρίσκουμε ένα σημείο $A_{j,i}$ για κάθε $\delta\Pi_j$. Η ένωση των $A_{j,i}$ για όλες τις πλευρές και τις γωνίες $\delta\Theta_j$ διαμορφώνουν την καμπύλη M_1.

(4) Το περίγραμμα (Π_3) μίας άλλης υλοποίησης του γράμματος Μ επιλέγεται τυχαία και περιστρέφεται, προσαρμόζεται και μετατοπίζεται, όπως στο βήμα 2, για να ταιριάξει βέλτιστα με το M_1 στη βάση του ίδιου κριτηρίου.

(5) Συνεχίζοντας όπως στο βήμα 3 υπολογίζουμε την μέση καμπύλη των Π_1, Π_2, Π_3 με τον ακριβώς ανάλογο τρόπο, μέσω της μέσης καμπύλης των $s_{i,1}, s_{i,2}$ και $s_{i,3}$

Συνεχίζουμε να εφαρμόζουμε τα βήματα 4 και 5 επιλέγοντας κάθε φορά το περίγραμμα μίας άλλης υλοποίησης του γράμματος, ταιριάζοντάς τα στο προηγούμενο μοντέλο M_i και υπολογίζοντας την καμπύλη M_{i+1} ως τη μέση τιμή των ήδη ταιριασμένων $\Pi_1, \ldots, \Pi_{i+2}$, μέχρι να τελειώσουν όλες οι άλλες υλοποιήσεις. Το αποτέλεσμα της μέσης καμπύλης όλων των διαθέσιμων Π_i θεωρείται ως η πρώτη προσέγγιση I_1του ιδεατού προτύπου του συγκεκριμένου γράμματος, που θα αποκαλούμε επίσης «πραγματικό πρότυπο».

3.4.2 Μία δεύτερη προσέγγιση του «ιδεατού» προτύπου

Όλα τα περιγράμματα $\Pi_1, \dots, \Pi_N$ όλων των υλοποιήσεων του συγκεκριμένου γράμματος του θεωρούμενου κειμένο ταιριάζουν βέλτιστα στο I_1, πάλι μέσω περιστροφής, προσαρμογής μεγέθους και μετατόπισης όπως στο βήμα 2. Μετά την ολοκλήρωση της διαδικασίας ταιριάσματος, γίνεται η εκτίμηση της μέσης καμπύλης των βέλτιστα ταιριασμένων περιγραμμάτων όπως στο βήμα 5 και θεωρείται ως μία καλύτερη προσέγγιση I_2 του ιδεατού προτύπου του συγκεκριμένου γράμματος για το κείμενο που έχουμε επιλέξει.

3.4.3 Τελική εκτίμηση του «ιδεατού» προτύπου

Οι εκτιμήσεις I_1 και I_2 ενός πλατωνικού προτύπου εξαρτώνται, σε έναν βαθμό, από την επιλογή του αρχικού περιγράμματος Π_1 και την ακολουθία των ταιριασμένων περιγραμμάτων Π_i. Για να ελαχιστοποιήσουμε αυτήν την εξάρτηση ορίζουμε όλα τα άλλα περιγράμματα των υλοποιήσεων που έχουμε στη διάθεσή μας για το συγκεκριμένο κείμενο να παίξουν τον ρόλο του Π_1. Επιπλέον, για κάθε διαφορετικό Π_1 αλλάζουμε την σειρά των ταιριασμένων περιγραμμάτων τυχαία, αποκτώντας N διαφορετικές προσεγγίσεις $I_{2,1}, \dots, I_{2,n}$ του ιδεατού προτύπου. Ταιριάζουμε μεταξύ τους τις καμπύλες $I_{2,1}, \dots, I_{2,n}$ όπως στο βήμα 2. Με αυτόν τον τρόπο βρίσκουμε ένα «σύννεφο» από προσεγγίσεις ταιριασμένων ιδεατών καμπυλών. Στη συνέχεια, υπολογίζουμε τη μέση καμπύλη αυτού του σύννεφου όπως στο βήμα 5 με μία μόνο διαφορά: αποδίδουμε ένα σημείο στη γωνία $\delta\Theta_j$ ή ένα ευθύγραμμο τμήμα δp_j μόνο αν τουλάχιστον τα μισά από τα pixels του συνόλου $s_{i,1}, \dots, s_{i,N}$ αντιστοιχούν σε αυτά. Θεωρούμε ότι η καμπύλη I που προκύπτει ως την τελική εκτίμηση του ιδεατού ή «πλατωνικού» προτύπου του γράμματος και του κειμένου που έχουμε επιλέξει.

4. ΣΤΑΤΙΣΤΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΤΩΝ ΜΕΘΟΔΩΝ ΤΗΣ ΠΑΡΑΓΡΑΦΟΥ 3

4.1 Χρήση του πραγματικού προτύπου για τη δημιουργία δύο κριτηρίων για την αναγνώριση γραφέα

Για να αποφανθούμε αν δύο διαφορετικά κείμενα, έστω E_1 και E_2, προέρχονται από το ίδιο χέρι ή όχι, ακολουθούμε ως εξής: Αρχικά, θεωρούμε ένα συγκεκριμένο γράμμα, έστω το Α, και όλες του τις N_1 υλοποιήσεις στο κείμενο E_1 και όλες του τις N_2 υλοποιήσεις στο E_2. Προσδιορίζουμε τα πραγματικά πρότυπα του Α, I_1^A και I_2^A για τα E_1 και E_2 αντίστοιχα. Στη συνέχεια ταιριάζουμε τα δύο πρότυπα I_1^A και I_2^A ορίζοντας το I_1^A να παίξει τον ρόλο του Π_1 και περιστρέφοντας, προσαρμόζοντας και μετατοπίζοντας το I_2^A ώστε το σφάλμα ε να ελαχιστοποιηθεί. Θα μπορούσαμε να περιμένουμε ότι αν το ίδιο χέρι έγραψε τα κείμενα E_1 και E_2 τότε τα I_1^A και I_2^A πρέπει να ταιριάζουν καλά όταν εφαρμόζεται η προαναφερθείσα διαδικασία. Από την άλλη πλευρά, είναι λογικό να θεωρούμε ότι όταν τα δύο κείμενα E_1 και E_2 είναι γραμμένα από διαφορετικά χέρια, τότε τα δύο βέλτιστα ταιριασμένα πραγματικά πρότυπα μπορεί να διαφέρουν σημαντικά, τουλάχιστον για έναν αριθμό γραμμάτων.

Μία άλλη προσέγγιση για να αποφασίσουμε αν δύο κείμενα, έστω E_λ και E_u, είναι γραμμένα από τον ίδιο γραφέα είναι η εξής: ορίζουμε το I_λ ως το πραγματικό πρότυπο ενός γράμματος στο E_λ. Περιστρέφουμε, προσαρμόζουμε και μετατοπίζουμε παράλληλα όλα τα περιγράμματα των υλοποιήσεων του συγκεκριμένου γράμματος στο E_u ώστε να ταιριάζουν στο I_λ. Έτσι αποκτούμε μία σύνθεση όπου τα περιγράμματα του γράμματος στο E_u διαμορφώνουν μία δέσμη που προσπαθεί να ακολουθήσει το πρότυπο I_λ. Όταν τα E_λ και E_u προέρχονται από το ίδιο χέρι, θα περιμέναμε τα περιγράμματα του γράμματος να είναι αρκετά κοντά στο I_λ και συμπαγή γύρω από αυτό. Επιπλέον, είναι λογικό να θεωρούμε ότι το I_λ θα είναι κοντά στη μέση καμπύλη των ταιριασμένων περιγραμμάτων. Θα περιμέναμε το αντίθετο αν τα E_λ και E_u ήταν γραμμένα από διαφορετικούς γραφείς, τουλάχιστον για έναν αριθμό γραμμάτων. Αυτές οι παρατηρήσεις μπορούν να οδηγήσουν στην διατύπωση δύο ποσοτικών κριτηρίων για την αναγνώριση γραφέα.

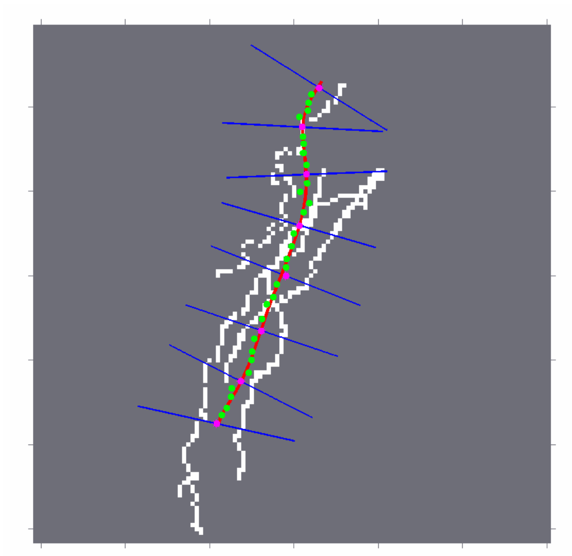
4.2 Εκτίμηση της κατανομής του λάθους προσαρμογής μίας υλοποίησης στο πραγματικό πρότυπο

Για να ποσοτικοποιήσουμε τις προηγούμενες παρατηρήσεις είναι απαραίτητο να εκτιμήσουμε την κατανομή των διαφορετικών υλοποιήσεων του θεωρούμενου γράμματος ενός κειμένου E_μ, που είναι ταιριασμένα γύρω από το πραγματικό πρότυπο του ίδιου γράμματος I_λ ενός άλλου κειμένου E_λ. Για το E_μ, θεωρούμε όλες τις ταιριασμένες υλοποιήσεις του συγκεκριμένου γράμματος στο πραγματικό πρότυπο, και τις συμβολίζουμε με Π_i^{MR}, όπου ο δείκτης MR σημαίνει Matched Realizations, δηλαδή ταιριασμένες υλοποιήσεις.

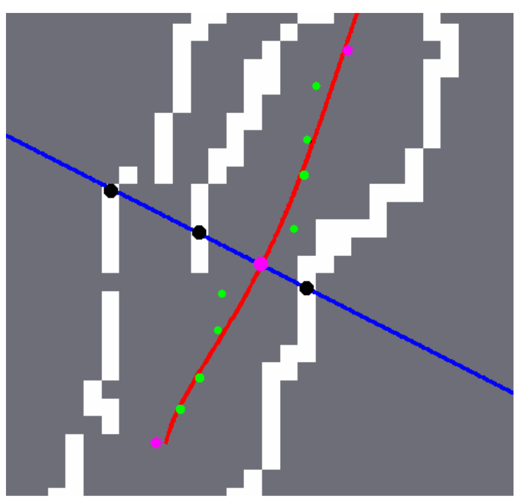
Κάνουμε τις ακόλουθες στατιστικές υποθέσεις:

(α) Έστω I το πραγματικό πρότυπο του θεωρούμενου γράμματος που αποτελείται από D πλευρές $J_1, \ldots, J_D$. Θεωρούμε μία πλευρά J_k και την προσεγγίζουμε βέλτιστα με ένα πολυώνυμο όπως έχουμε περιγράψει στην παράγραφο 3.3.3. Προφανώς, ο βαθμός του πολυωνύμου εξαρτάται από την ανάλυση της εικόνας, από το μέγεθος του γράμματος, κλπ. Σε όλες τις περιπτώσεις που έχουμε εξετάσει μέχρι στιγμής ένα πολυώνυμο εβδόμου βαθμού ήταν σαφώς επαρκές για μία ποσοτικά καλή αναπαράσταση. Σε κάθε σημείο της πλευράς J_k, έστω το J_j, θεωρούμε την κατακόρυφη γραμμή στο πολυώνυμο προσέγγισης και τα σημεία τομής της $T_{i,j}$ με όλες τις υλοποιήσεις $\Pi_{i,k}^{MR}$, όπου ο δείκτης k υποδηλώνει την k-οστή πλευρά του Π_i^{MR} (Εικόνες 1,2). Δεχόμαστε τη στατιστική υπόθεση ότι αν το $J_{k,j}$ είναι το j-οστό pixel του J_k, τότε η σχετική απόσταση των $T_{i,j}$ και $J_{k,j}$ ακολουθεί κανονική κατανομή με μηδενική μέση τιμή και τυπική απόκλιση σ, ανεξάρτητη του j (Εικόνα 2). Παρόλο που αυτή είναι μία εμπειρική υπόθεση, έχουμε εφαρμόσει το test

Kolmogorov και σε πολλές περιπτώσεις τη δοκιμή χι-τετράγωνο για να βρούμε αν υπάρχουν αρκετά στοιχεία που έρχονται σε αντίθεση με αυτή. Οι στατιστικές αυτές δοκιμές δεν οδήγησαν σε απόρριψη της αρχικής υπόθεσης.



Εικόνα 1



Εικόνα 2

(β) Θεωρούμε ένα υποσύνολο J_k^n του J_k, όπου το πρώτο του pixel $J_{k,1}^n$ είναι επίσης το πρώτο pixel του J_k, το δεύτερο pixel $J_{k,2}^n$ είναι n pixel μετά το $J_{k,1}^n$ και ούτω καθεξής (Εικόνα 1). Υποθέτουμε ότι οι κατανομές που περιγράψαμε στο (α) παραπάνω σχετικά με τα pixels του J_k^n είναι στατιστικά ανεξάρτητες όταν το n είναι μεγαλύτερο από μία συγκεκριμένη τιμή κατωφλίου. Σε όλες τις περιπτώσεις που έχουμε εξετάσει μέχρι στιγμής η τιμή $n = 5$ είναι μία ικανή τιμή κατωφλίου. Και πάλι αυτή είναι μία εμπειρική υπόθεση, ωστόσο οι στατιστικές δοκιμές που εφαρμόστηκαν στα διαθέσιμα δεδομένα δεν έρχονται σε αντίθεση με αυτή την υπόθεση.

(γ) Θεωρούμε την πλευρά J_k με μήκος L_k και τα τελικά σημεία A_1, A_2 μίας τυχαίας υλοποίησης Π_i^{MR} με μήκος $L_{i,k}^{MR}$. Κάνουμε την υπόθεση ότι το $L_{i,k}^{MR}\iota$ ακολουθεί κανονική κατανομή με μέση τιμή L_k και τυπική απόκλιση S_L.

4.3 Δύο σχετικά κριτήρια αναγνώρισης γραφέα

Το ένα κριτήριο για την αναγνώριση γραφέα βασίζεται στο πόσο κοντά είναι οι ταιριασμένες υλοποιήσεις Π_i^{MR} στο πραγματικό πρότυπο I_λ. Θεωρούμε το πραγματικό πρότυπο I_λ^A των N_λ^A υλοποιήσεων του γράμματος Α, Π_i^A, του κειμένου E_λ, και μία αυθαίρετη πλευρά J_k αυτού του προτύπου, το υποσύνολό του J_k^n και τις αντίστοιχες κάθετες γραμμές. Ταιριάζουμε όλες τις N_u^A υλοποιήσεις του γράμματος Α του κειμένου E_u με το πραγματικό πρότυπο I_λ^A, δημιουργώντας μία δέσμη $\Pi_i^{A,MR}$. Έστω $Z_1, \dots, Z_\nu$ τα σημεία αυτά του υποσυνόλου J_k^n για τα οποία οι κάθετες γραμμές στην πλευρά J_k τέμνουν όλες τις υλοποιήσεις $\Pi_i^{A,MR}$. Έστω $\Lambda_{i,j}$ το σημείο τομής της $\Pi_i^{A,MR}$ με την κάθετη γραμμή στο σημείο Z_j. Θεωρούμε τις αποστάσεις $Z_j\Lambda_{i,j}$, βρίσκουμε μία κανονική κατανομή για κάθε Z_j και τελικά n στατιστικά ανεξάρτητες κατανομές. Έστω $\overline{Z_j\Lambda_j}$ η μέση τιμή και S_u η τυπική απόκλιση των αποστάσεων $Z_j\Lambda_{i,j}$. Επιπλέον, σε σχέση με την κάθετη γραμμή στο J_k^n, υπάρχει μία άλλη κατανομή σχετική με το πραγματικό πρότυπο. Στην πραγματικότητα, όταν όλα τα περιγράμματα του συγκεκριμένου γράμματος του κειμένου E_λ ταιριάζουν στο I_λ, τότε η τομή αυτών των ταιριασμένων περιγραμμάτων με κάθε κάθετη γραμμή ακολουθεί μία κανονική κατανομή με μηδενική μέση τιμή και εκτιμώμενη διακύμανση S_λ^2. Αν υποθέσουμε ότι τα E_λ και E_u είναι γραμμένα από τον ίδιο γραφέα, τότε η ποσότητα

$$t_{k,j} = \frac{\overline{Z_J \Lambda_J}}{\sqrt{\frac{S_\lambda^2}{N_\lambda^A} + \frac{S_u^2}{N_u^A}}}$$

ακολουθεί την κατανομή Student με d βαθμούς ελευθερίας, όπου:

$$d = \frac{(S_\lambda^2/N_\lambda^A + S_u^2/N_u^A)^2}{\frac{(S_\lambda^2/N_\lambda^A)^2}{N_\lambda^A - 1} + \frac{(S_u^2/N_u^A)^2}{N_u^A - 1}}$$

Χρησιμοποιώντας την τιμή t μπορούμε έχουμε μία πρώτη προσέγγιση ώστε να αποφασίσουμε αν τα δύο κείμενα E_λ και E_u ανήκουν στον ίδιο γραφέα όσον αφορά το συγκεκριμένο γράμμα. Συγκεκριμένα, έστω $f(t_{k,j})$ η τιμή της συνάρτησης πυκνότητας πιθανότητας της κατανομής Student με d βαθμούς ελευθερίας στο σημείο $t_{k,j}$. Μπορούμε να εξετάσουμε τις στατιστικές υποθέσεις σχετικά με την πιθανότητα ότι δύο κείμενα ανήκουν στον ίδιο γραφέα με ένα συγκεκριμένο βαθμό σιγουριάς χρησιμοποιώντας διανύσματα: για μία μόνο πλευρά αρχικά, το διάνυσμα:

$$\alpha_k^A = [t_{k,1} \dots t_{k,\nu}]$$

και αν θεωρήσουμε όλες τις πλευρές του γράμματος, τότε βρίσκουμε το διάνυσμα:

$$\beta^A = [\alpha_1^A \dots \alpha_D^A].$$

Επιπλέον ορίζουμε τις ποσότητες:

$$\delta\alpha_k^A = \sqrt[\nu]{\prod_{j=1}^{\nu} (f(t_{k,j}))}$$

και

$$\delta\alpha^A = \prod_{\substack{k\ over \\ all\ sides}} \delta\alpha_k^A \text{ για να τις}$$

για να τις χρησιμοποιήσουμε σε προσεγγίσεις μέγιστης πιθανότητας. Στην πραγματικότητα, ακόμα και αν τα κείμενα E_λ και E_u ανήκουν στον ίδιο γραφέα, η τιμή $f(t_{k,j})$ μπορεί να είναι αυθαίρετα μικρή. Ωστόσο, έχουμε έναν σημαμτικό αριθμό στατιστικά ανεξάρτητων δειγμάτων του ίδιου πληθυσμού, που αντιστοιχούν στις πολλές $Z_j, j = 1, \dots \nu$ και διαφορετικές πλευρές του θεωρούμενου γράμματος. Επομένως, όταν το κείμενο E_u συγκρίνεται με άλλα για αναγνώριση γραφέα, τότε θα περιμέναμε ότι το E_u έχει τη μεγαλύτερη πιθανότητα να ανήκει στον ίδιο γραφέα με το κείμενο E_μ για το οποίο η αντίστοιχη ποσότητα $\delta\alpha^A$ είναι μέγιστη, όσον αφορά το γράμμα Α. Αν περισσότερα από ένα γράμματα είναι διαθέσιμα, τότε εφαρμόζουμε την ίδια μεθοδολογία χρησιμοποιώντας την ποσότητα

$$\delta\beta = \prod_{\substack{l\ over \\ all\ letters}} \delta\alpha^l$$

όπου έχουμε αντικαταστήσει το Α στο $\delta\alpha^A$ με l ώστε να λάβουμε υπόψη όλα τα διαθέσιμα γράμματα.

Σχετικά με τις εκτιμώμενες διακυμάνσεις S_λ^2 και S_u^2 των κατανομών $Z_j \Lambda_{i,j}$ για τα δύο κείμενα E_λ και E_u, σημειώνουμε ότι η ποσότητα

$$F_{k,j} = \frac{S_\lambda^2/\sigma_\lambda^2}{S_u^2/\sigma_u^2}$$

η οποία ακολουθεί την κατανομή F με $\left(N_\lambda^A - 1, N_u^A - 1\right)$ βαθμούς ελευθερίας και συνάρτηση πυκνότητας πιθανότητας $f_{Sn}(F)$. Αν καταλήξουμε στο συμπέρασμα ότι τα δύο κείμενα E_λ και E_u ανήκουν στον ίδιο γραφέα, τότε $\sigma_\lambda = \sigma_u$ και στην περίπτωση αυτή η ποσότητα $F_{k,j}$ παίρνει μία συγκεκριμένη τιμή. Επομένως, μπορούμε να διατυπώσουμε στατιστικές υποθέσεις και να κάνουμε εκτιμήσεις μέγιστης πιθανότητας χρησιμοποιώντας τα διανύσματα και τις ποσότητες όπως και παραπάνω

$$\zeta_k^A = \left[F_{k,1} \dots F_{k,\nu}\right],$$

$$\gamma^A = [\zeta_1^A \dots \zeta_D^A],$$

$$\delta\zeta_k^A = \sqrt[\nu]{\prod_{j=1}^{\nu}(f_{Sn}(F_{k,j}))},$$

$$\delta\zeta^A = \prod_{\substack{k\ over \\ all\ sides}} \delta\zeta_k^A,$$

$$\delta\gamma = \prod_{\substack{l\ over \\ all\ letters}} \delta\zeta^l.$$

Έτσι, διατυπώνουμε τα παρακάτω δύο κριτήρια για την αναγνώριση γραφέα.

<u>Κριτήριο 1</u>

Υποθέτουμε ότι έχουμε Ν κείμενα $E_1, \dots, E_N$ που προέρχονται από Ν διαφορετικά χέρια $H_i, i = 1, \dots, N$ και ένα κείμενο άγνωστου γραφέα E_u. Για να αποφασίσουμε αν το E_u είναι γραμμένο από ένα από τα χέρια H_i ακολουθούμε τα εξής βήματα: (α) Υπολογίζουμε τα πλατωνικά πρότυπα $I_i^l, l = M, P, \dots etc$ για τα αντίστοιχα γράμματα και όλα τα κείμενα E_i. (β) Στη συνέχεια ταιριάζουμε τις υλοποιήσεις ενός γράμματος l στο κείμενο E_u με το πραγματικό πρότυπο I_1^l, του E_1, αποκτώντας μία δέσμη περιγραμμάτων $\Pi_i^{l,MR}$. (γ) Στη συνέχεια κάνουμε την υπόθεση ότι τα κείμενα E_uκαι E_1 είναι του ίδιου γραφέα και την ελέγχουμε για κάθε γράμμα χωριστά χρησιμοποιώντας τα διανύσματα β_1^A και τις συναρτήσεις πυκνότητας πιθανότητάς τους. Η ίδια υπόθεση γίνεται για όλα τα Ν κείμενα και ποσότητες β_i^A. (δ) Επαναλαμβάνουμε το προηγούμενο βήμα για όλα τα θεωρούμενα γράμματα. Αν για ένα συγκεκριμένο γράμμα η υπόθεση που δοκιμάζεται στο βήμα (γ) απορριφθεί, τότε το συμπέρασμα ότι το αντίστοιχο χέρι H_1 έχει γράψει και το κείμενο E_u απορρίπτεται επίσης. (ε) Επαναλαμβάνουμε τα προηγούμενα βήματα για όλα τα υπόλοιπα κείμενα $E_2, \dots, E_N$, αποκτώντας ένα περιορισμένο σύνολο κειμένων $E_1^{NR}, \dots E_M^{NR}$ για τα οποία η υπόθεση ότι κάθε ένα από αυτά ανήκει στον ίδιο γραφέα με το E_u δεν έχει απορριφθεί, όπου NR σημαίνει Not Rejected. (στ) Τελικά, η απόφαση για το ποιά χέρια ανάμεσα σε αυτά που δεν απορρίφθηκαν στο βήμα (ε), αν υπάρχουν κάποια, έγραψαν το κείμενο E_u παίρνεται επιλέγοντας τον γραφέα που έγραψε το κείμενο που έχει τη μέγιστη τιμή $\delta\beta_i$ που υπολογίζεται για κάθε κείμενο E_i^{NR}.

Κριτήριο 2

Υποθέτουμε ότι έχουμε N κείμενα $E_1, \dots, E_N$ που προέρχονται από N διαφορετικά χέρια $H_i, i = 1, \dots, N$ και ένα κείμενο άγνωστου γραφέα E_u. Για να αποφασίσουμε αν το E_u είναι γραμμένο από ένα από τα χέρια H_i ακολουθούμε τη διαδικασία όπως στο *κριτήριο 1*: (α) κάνουμε την υπόθεση ότι τα κείμενα E_u και E_i ανήκουν στον ίδιο γραφέα για κάθε κείμενο E_i χωριστά και την ελέγχουμε μέσω των διανυσμάτων γ_i για κάθε διαθέσιμο γράμμα. (β) Αν η υπόθεση απορρίπτεται για ακόμα και ένα γράμμα, τότε το συμπέρασμα ότι το αντίστοιχο χέρι H_i έγραψε και το κείμενο E_u, απορρίπτεται επίσης. (γ) Όπως και στο *κριτήριο 1* αποκτούμε ένα περιορισμένο σύνολο κειμένων $E_1^{NR}, \dots, E_M^{NR}$ για τα οποία η υπόθεση ότι κάθε ένα από αυτά έχει γραφτεί από το ίδιο χέρι με το E_u δεν έχει απορριφθεί. (δ) Η απόφαση για το ποιά χέρια ανάμεσα σε αυτά που δεν απορρίφθηκαν στο βήμα (γ), αν υπάρχουν κάποια, έγραψαν και το κείμενο E_u παίρνεται επιλέγοντας τον γραφέα που έγραψε το κείμενο που έχει τη μέγιστη τιμή $\delta\beta_i\delta\gamma_i$ που υπολογίζεται για κάθε κείμενο E_i^{NR}.

5. ΠΡΑΓΜΑΤΟΠΟΙΗΘΕΝΤΑ ΠΕΙΡΑΜΑΤΑ

Αρχικά αναφέρουμε ότι για όλες τις περιπτώσεις γραμμάτων που αντιμετωπίστηκαν σε αυτή τη διπλωματική το περίγραμμα πρέπει να έχει μία συγκεκριμένη μορφή, όπως προαναφέρθηκε συνοπτικά στο κεφάλαιο 3. Επομένως πρέπει να ισχύουν οι εξής περιορισμοί:

(α) Κάθε εικονοστοιχείο πρέπει να έχει ακριβώς δύο γειτονικά. Με τον όρο γειτονικά εννοούμε ότι αν θεωρήσουμε το τυχόν εικονοστοιχείο του περιγράμματος σαν κέντρο μίας περιοχής pixels 3x3, τότε ακριβώς δύο ακόμα pixels του περιγράμματος βρίσκονται εντός της περιοχής αυτής.

(β) Μόνο συγκεκριμένοι σχηματισμοί των τριών εικονοστοιχείων του περιγράμματος επιτρέπονται να βρίσκονται εντός αυτής της περιοχής.

(γ) Όλα τα pixels του περιγράμματος είναι αριθμημένα ωρολογιακά στη φυσική τους σειρά: το υπ’ αριθμόν ένα pixel είναι αυτό που βρίσκεται ψηλότερα και αριστερότερα στην εικόνα. Συνεπώς, το υπ’ αριθμόν n pixel έχει δύο ακριβώς γειτονικά, το υπ’ αριθμόν n-1 και το υπ’ αριθμόν n+1, στη σωστή διάταξη.

(δ) Απαγορεύεται να υπάρχουν συμπλέγματα μεμονομένων pixels συνολικού αριθμού μικρότερου ενός κατωφλίου. Αυτό το κατώφλι είναι συνήθως το 5% του συνολικού αριθμού εικονοστοιχείων του περιγράμματος.

Για να εξασφαλιστεί η τήρηση των παραπάνω περιορισμών αναπτύχθηκε ειδικός πρότυπος κώδικας, ο οποίος στη συνέχεια εφαρμόστηκε σε όλες τις φωτογραφίες του έργου «Αθηναίων Πολιτεία» που είχαμε στη διάθεσή μας και μάλιστα για όλα τα γράμματα της αλφαβήτου.

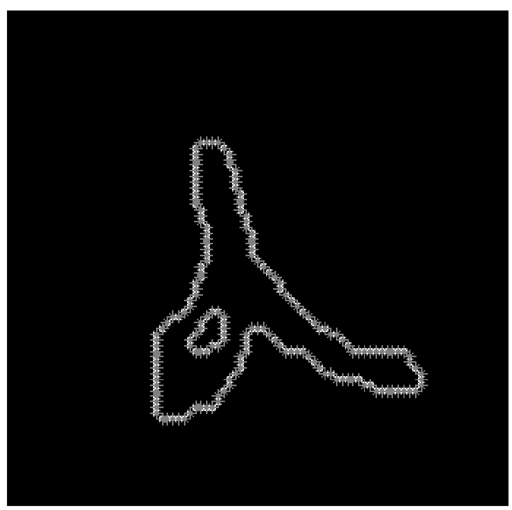
5.1 Παραδείγματα εξαγωγής περιγραμμάτων

Παρακάτω παραθέτουμε επτά από τα πιο χαρακτηριστικά παραδείγματα από τα πραγματοποιηθέντα πειράματα στα οποία εφαρμόστηκε η ίδια ακριβώς διαδικασία και η οποία περιλαμβάνει τα εξής βήματα: (1) εξαγωγή της φωτογραφίας μίας υλοποιήσεως του συμβολογράμματος από την ψηφιακή φωτογραφία ευρύτερου τμήματος του κειμένου, (2) μετατροπή της αρχικά έγχρωμης φωτογραφίας της υλοποιήσεως σε φωτογραφία διαβάθμισης του γκρι και (3) αυτόματη κατάτμηση εικόνας και μετατροπή του αποτελέσματος σε δυαδική (binary) εικόνα, όπου το σώμα του γράμματος παριστάνεται με σύμβαση με μαύρο χρώμα (0), ενώ το υπόβαθρο με άσπρο χρώμα (1) και (4) εξαγωγή του περιγράμματος της συγκεκριμένης υλοποίησης.

I. Το γράμμα «Α»

Εικόνα 1α: Υλοποίηση του γράμματος «Α»

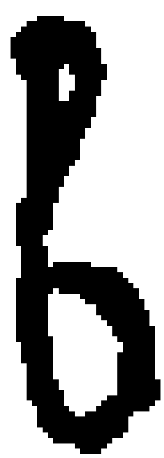
Εικόνα 1β: Υλοποίηση του γράμματος «Α» σε δυαδική εικόνα



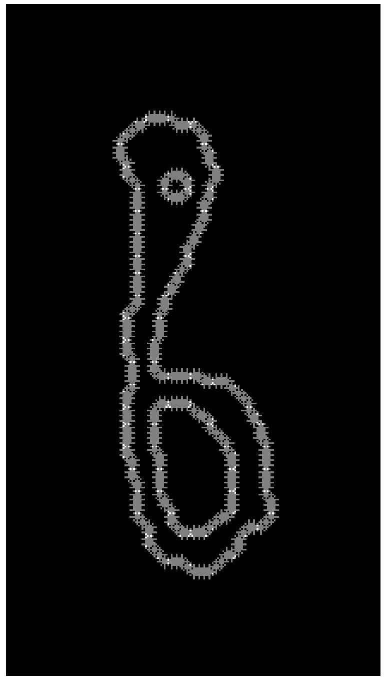
Εικόνα 1γ: Περίγραμμα του γράμματος «Α»

II. Το γράμμα «Β»

Εικόνα 2α: Υλοποίηση του γράμματος «Β»

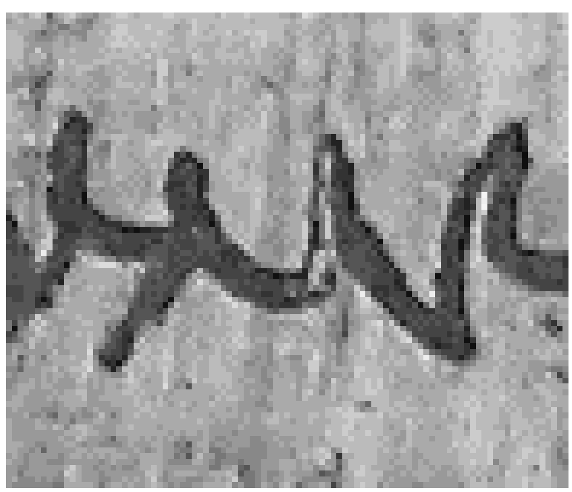


Εικόνα 2β: Υλοποίηση του γράμματος «Β»



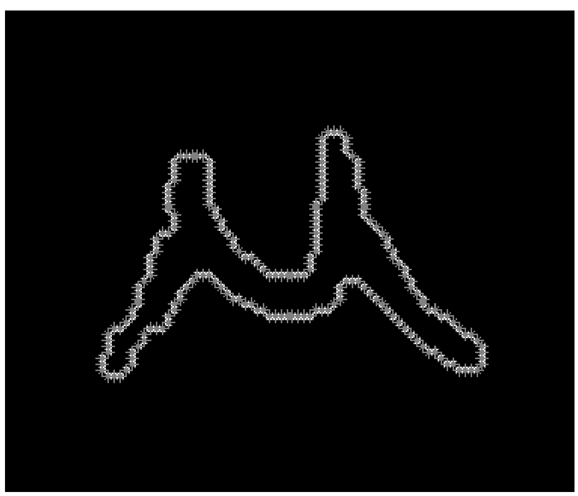
Εικόνα 2γ: Υλοποίηση του γράμματος «Β»

III. Το γράμμα «Μ»



Εικόνα 3α: Υλοποίηση του γράμματος «Μ»

Εικόνα 3β: Υλοποίηση του γράμματος «Μ»

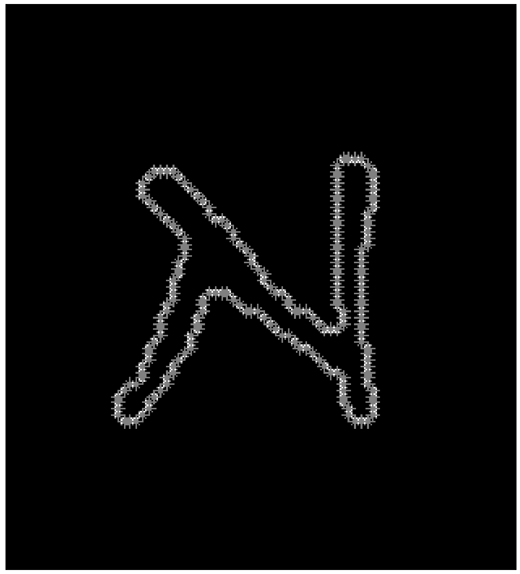


Εικόνα 3γ: Υλοποίηση του γράμματος «Μ»

IV. Το γράμμα «Ν»

Εικόνα 4α: Υλοποίηση του γράμματος «Ν»

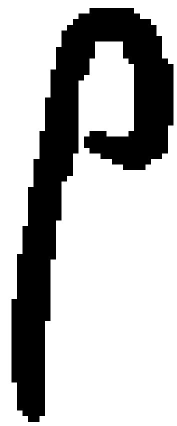
Εικόνα 4β: Υλοποίηση του γράμματος «Ν»



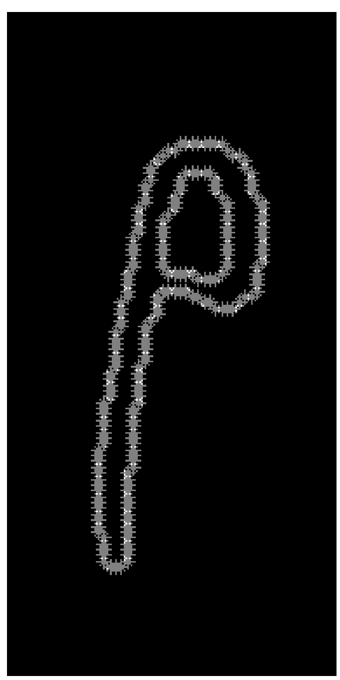
Εικόνα 4γ: Υλοποίηση του γράμματος «Ν»

V. Το γράμμα «Ρ»

Εικόνα 5α: Υλοποίηση του γράμματος «Ρ»



Εικόνα 5β: Υλοποίηση του γράμματος «Ρ»

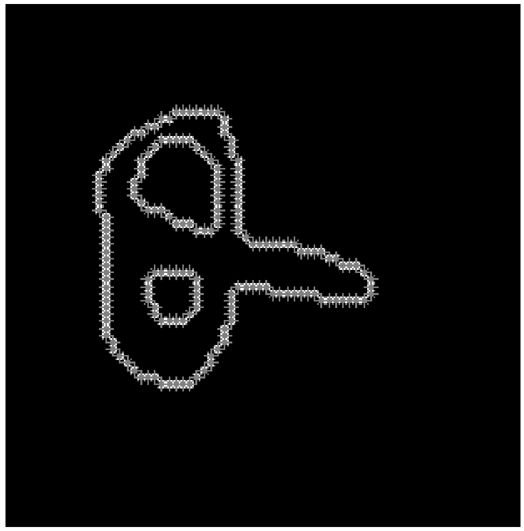


Εικόνα 5γ: Υλοποίηση του γράμματος «Ρ»

VI. Το γράμμα «Θ»

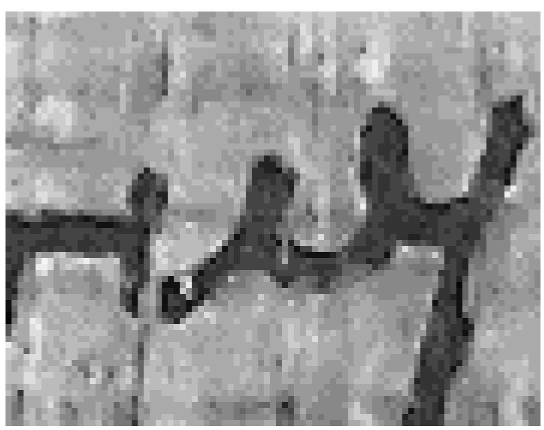
Εικόνα 6α: Υλοποίηση του γράμματος «Θ»

Εικόνα 6β: Υλοποίηση του γράμματος «Θ»



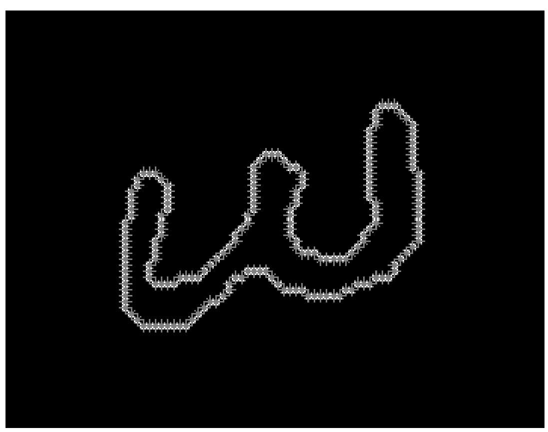
Εικόνα 6γ: Υλοποίηση του γράμματος «Θ»

VII. Το γράμμα «Ω»



Εικόνα 7α: Υλοποίηση του γράμματος «Ω»

Εικόνα 7β: Υλοποίηση του γράμματος «Ω»

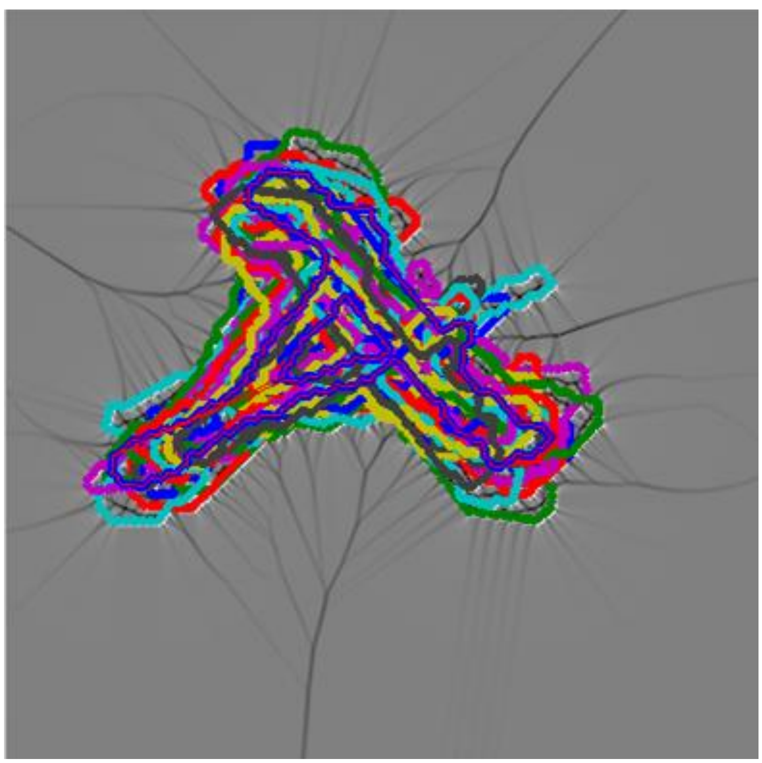


Εικόνα 7γ: Υλοποίηση του γράμματος «Ω»

5.2 Προσδιορισμός πραγματικού προτύπου/ιδεατού αντιπροσώπου

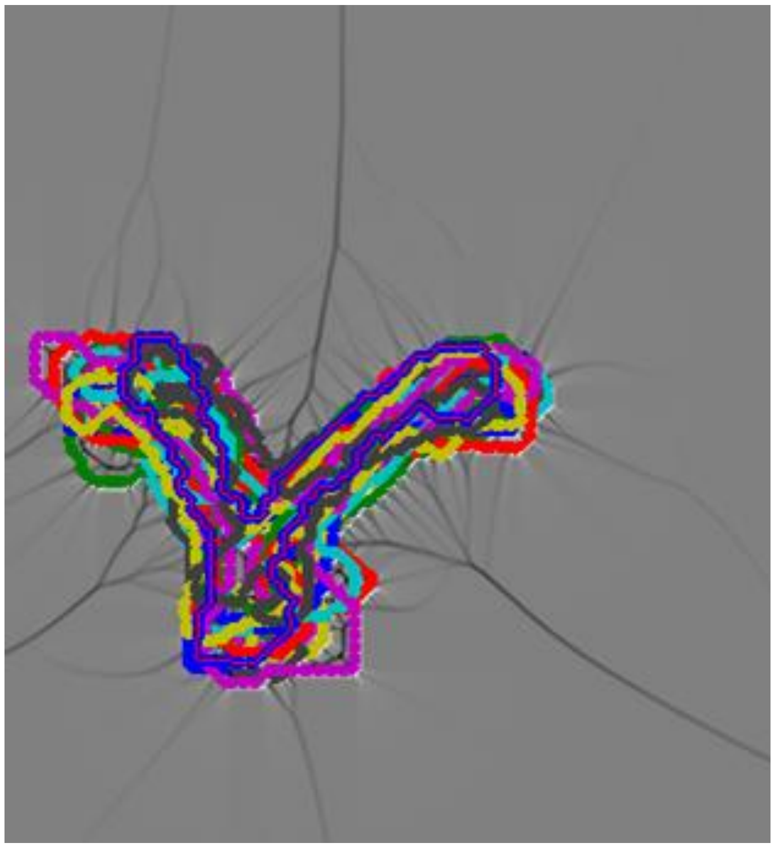
Σε αυτή την παράγραφο παραθέτουμε δύο από τα πιό αντιπροσωπευτικά παραδείγματα προσέγγισης του πραγματικού προτύπου/ιδεατού αντιπροσώπου, τα οποία προέκυψαν με τη μέθοδο που περιγράψαμε στην παράγραφο 3.4.

I. Γράμμα «Α»



Εικόνα 1: Προσδιορισμός πραγματικού αντιπροσώπου για το γράμμα «Α»

II. Γράμμα «Γ»



Εικόνα 2: Προσδιορισμός πραγματικού αντιπροσώπου για το γράμμα «Γ»

6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΤΩΝ ΠΕΙΡΑΜΑΤΩΝ

Τα αποτελέσματα των πειραμάτων που παρουσιάστηκαν στην προηγούμενη παράγραφο μας οδήγησαν στα παρακάτω συμπεράσματα για την αυτόματη αναγνώριση γραφέα:

(1) Φαίνεται ότι η μέθοδος κατάτμησης εικόνας που παρουσιάστηκε και περιγράφεται στο κεφάλαιο 3 της παρούσης εργασίας, λειτουργεί πολύ ικανοποιητικά για την εξαγωγή του σώματος ενός γράμματος από το υπόβαθρο της εικόνας.

(2) Ο αλγόριθμος εξαγωγής περιγραμμάτων λειτουργεί πρακτικά αλάνθαστα, προσφέροντας αυτόματα περιγράμματα στη μορφή που περιγράφηκε αναλυτικά στην παράγραφο 5.

(3) Ο ιδεατός αντιπρόσωπος κάθε συμβόλου της αλφαβήτου που υπολογίζεται από ένα κείμενο φαίνεται να έχει πολύ ομαλή μορφή και να περιγράφει ιδιαιτέρως ικανοποιητικά τις βέλτιστα προσαρμοσμένες υλοποιήσεις του ίδιου γράμματος του υπό μελέτη κειμένου.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] "Αρχαία Ελληνικά Φιλοσοφικός Λόγος", Υπουργείο Εθνικής Παιδείας και Θρησκευμάτων, Παιδαγωγικό Ινστιτούτο, Οργανισμός εκδόσεως διδακτικών βιβλίων, Αθήνα 2010

[2] Αριστοτέλης, Αθηναίων Πολιτεία, Βιβλιοθήκη αρχαίων συγγραφέων, Ι. Ζαχαρόπουλος

[3] "Εισαγωγικά Μαθήματα Στατιστικής Ανάλυσης", Γ. Τζιαφέτας, Αθήνα 1981

[4] www.wikipedia.com

[5] "Στατιστική", Κ. Κωστάκη, Γ. Κωστάκη, Χ. Κωστάκη, Αθήνα 2002

[6] "Εισαγωγή στις Πιθανότητες και τη Στατιστική", Χ. Δαμιανού, Ν. Παπαδάτος, Χ. Α. Χαραλαμπίδης, Τμήμα Μαθηματικών Πανεπιστημίου Αθηνών, Αθήνα 2003

[7] "Automatic writer identification of ancient Greek inscriptions", M. Panagopoulos, C. Papaodysseus, P. Rousopoulos, D. Dafi, S. Tracy, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009

[8] "Ανάπτυξη προηγμένων μεθόδων αναγνώρισης προτύπων και μαθηματική θεμελίωση για τον εντοπισμό του τρόπου δημιουργίας και της ταυτότητας του δημιουργού σημαντικών αρχαιολογικών ευρημάτων", Διδακτορική διατριβή, Ρουσόπουλος Παναγιώτης, 2010