Εθνικό Μετσόβιο Πολυτεχνείο

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Εξόρυξη Δεδομένων και Καταγραφή του Περιεχομένου του Διαδικτύου και των Κοινωνικών Δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΙΑΣΟΝΑ ΚΑΤΣΑΜΕΝΗ

Επιβλέπων: Πέτρος Στεφανέας
Λέκτορας Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2014

Στους γονείς μου

Περιεχόμενα

Πρόλογος ix
Κεφάλαιο 1. Εισαγωγή 1
1.1 Τί είναι τα μέσα κοινωνικής δικτύωσης; 1
1.2 Τα οφέλη της αξιοποίησης των μέσων κοινωνικής δικτύωσης 2
Κεφάλαιο 2. Big Data 5
2.1 Η εκτίναξη του όγκου πληροφοριών στις μέρες μας 5
2.2 Τί είναι τα Big Data; 7
2.3 Ιστορική αναδρομή των Big Data 10
2.4 Η εκτόξευση του ενδιαφέροντος γύρω από τα Big Data 18
2.5 Η αξιοποίηση των Big Data 20
Κεφάλαιο 3. Μεθοδολογίες 23
3.1 Εξόρυξη δεδομένων (Data mining) 23
3.2 Εξόρυξη κειμένου (Text mining) 24
3.3 Μεθοδολογία 1 25
3.4 Μεθοδολογία 2 29
3.4.1 Εκτίμηση του μοντέλου 29
3.4.2 Έλεγχος της στατιστικής σημαντικότητας των επιμέρους παραμέτρων 31
3.4.3 Έλεγχος της ερμηνευτικής ικανότητας του μοντέλου 31
3.4.4 Προϋποθέσεις του μοντέλου 33
3.4.5 Εξέταση υπολοίπων 34
Κεφάλαιο 4. Η αποτύπωση της Ελληνικής οικονομικής κρίσης στο Twitter 37
4.1 Εισαγωγή 37
4.2 Η εμφάνιση και η ανάλυση της “ευτυχίας” στο Twitter 40
4.3 Η εμφάνιση και η ανάλυση της “απογοήτευσης” στο Twitter 44
4.4 Η εμφάνιση και η ανάλυση της “αγανάκτησης” στο Twitter 49
4.5 Συσχέτιση της ανεργίας με την απογοήτευση - αποστροφή 53

Κεφάλαιο 5. Αποτύπωση της οικονομικής δραστηριότητας στο Google 57
5.1 Εισαγωγή 57
5.2 Συλλογή δεδομένων της αυτοκινητοβιομηχανίας Α 59
5.3 Συσχέτιση των δύο μεταβλητών 62
Κεφάλαιο 6. Πρόβλεψη της οικονομικής δραστηριότητας μέσω του Google ..65
6.1 Εισαγωγή 65
6.2 Συλλογή δεδομένων της αυτοκινητοβιομηχανίας Η 67
6.3 Συσχέτιση των δύο μεταβλητών 70
6.4 Απλό γραμμικό μοντέλο.....72
6.4.1 Έλεγχος της στατιστικής σημαντικότητας των επιμέρους παραμέτρων 73
6.4.2 Έλεγχος της ερμηνευτικής ικανότητας του μοντέλου.....74
6.4.3 Έλεγχος για ύπαρξη ετεροσκεδαστικότητας.....75
6.4.4 Έλεγχος για ύπαρξη αυτοσυσχέτισης 77
6.4.5 Γραφική εξέταση της Κανονικής κατανομής.....81
6.5 Γενικό γραμμικό μοντέλο 83
6.5.1 Μέθοδος της διαδοχικής αφαίρεσης και προσδιορισμός παραμέτρων 83
6.5.2 Έλεγχος της ερμηνευτικής ικανότητας του μοντέλου.....85
6.5.3 Έλεγχος για ύπαρξη ετεροσκεδαστικότητας.....86
6.5.4 Έλεγχος για ύπαρξη αυτοσυσχέτισης 87
6.5.5 Γραφική εξέταση της Κανονικής κατανομής.....89
6.5.6 Έλεγχος για συσχέτιση του σφάλματος με τις ανεξάρτητες μεταβλητές 90
6.5.7 Έλεγχος για ύπαρξη πολυσυγγραμμικότητας.....91
6.5.8 Έλεγχος για ύπαρξη σημείων επιρροής.....92
6.6 Πραγματοποίηση πρόβλεψης και συμπεράσματα.....94

Παραρτήματα 97
Παράρτημα Α. Ο απόηχος του κλεισίματος της ΕΡΤ στο διαδίκτυο 97
Α.1 Εισαγωγή 97
Α.2 Η εμφάνιση της ΕΡΤ στο Twitter 98
Α.3 Η ΕΡΤ στο Twitter συναρτήσει του χρόνου 99
Α.4 Η ΕΡΤ στην αναζήτηση του Google 101
Παράρτημα Β. Κώδικας για την αποτύπωση της οικονομικής κρίσης 105
Β.1 Κώδικας σε γλώσσα προγραμματισμού Java 105
Β.2 Δημιουργία διαγραμμάτων χρονοσειρών στην R 108
Β.3 Συσχέτιση στην R 110
Παράρτημα Γ. Κώδικας απλού γραμμικού μοντέλου 111
Γ.1 Συσχέτιση στην R (για την Α) 111
Γ.2 Συσχέτιση στην R (για την Η) 111
Γ.3 Εκτίμηση της συνάρτησης στην R 112
Γ.4 Δημιουργία διαγραμμάτων στην R 112
Παράρτημα Δ. Κώδικας γενικού γραμμικού μοντέλου 113
Δ.1 Μέθοδος διαδοχικής αφαίρεσης στην R 113
Δ.2 Εκτίμηση της συνάρτησης στην R 113
Δ.3 Δημιουργία διαγραμμάτων στην R 114
Δ.4 Έλεγχος για ύπαρξη πολυσυγγραμμικότητας 115
Δ.5 Πρόβλεψη πωλήσεων στην R 115
Βιβλιογραφία 117

Πρόλογος

Η παρούσα διπλωματική εργασία αφορά στη μελέτη των πληροφοριών που παράγονται μέσα από το διαδίκτυο και κυρίως από τα μέσα κοινωνικής δικτύωσης.

Αρχικά γίνεται μία ιστορική αναδρομή γύρω από τα Big Data, ενώ παράλληλα αναφέρονται διάφορα παραδείγματα αυτών, καθώς και τα βασικά θεωρητικά γνωρίσματα που τα διακρίνουν.

Στη συνέχεια μελετήθηκε ο βαθμός στον οποίο αποτυπώνεται στο διαδίκτυο ένα κοινωνικοπολιτικό φαινόμενο που, κατά τη διάρκεια μίας χρονικής περιόδου, μονοπωλεί το ενδιαφέρον των ΜΜΕ και της κοινωνίας γενικότερα. Έτσι λοιπόν εξετάστηκε αν εμφανίζεται ή όχι στα social media και συγκεκριμένα στο Twitter, η αρνητική ψυχολογία που διακρίνει τους Έλληνες κατά τη διάρκεια των τελευταίων ετών, λόγω της έντονης οικονομικής κρίσης. Προκειμένου να πραγματοποιηθεί η συγκεκριμένη ανάλυση, συλλέχθηκαν περίπου τέσσερα εκατομμύρια Tweets, που αναρτήθηκαν στη δημοφιλή σελίδα κοινωνικής δικτύωσης, μέσα σε διάστημα τεσσεράμισι ετών. Για την υλοποίηση της συγκεκριμένης εφαρμογής γράφτηκε κώδικας σε γλώσσα προγραμματισμού Java.

Τέλος αποδείχθηκε ότι οι αναζητήσεις μιας αυτοκινητοβιομηχανίας στο Google συμμεταβάλλονται με τις πωλήσεις της εταιρείας. Έτσι λοιπόν, χρησιμοποιώντας ανάλυση παλινδρόμησης, επιχειρήθηκε η κατασκευή ενός μοντέλου, το οποίο είναι ικανό να πραγματοποιεί μελλοντικές προβλέψεις των πωλήσεων της εταιρείας, χρησιμοποιώντας τον αριθμό των αναζητήσεών της, στη δημοφιλή μηχανή αναζήτησης. Για την στατιστική ανάλυση και την εξαγωγή συμπερασμάτων χρησιμοποιήθηκε το στατιστικό πακέτο R.

Κεφάλαιο 1. Εισαγωγή

1.1 Τί είναι τα μέσα κοινωνικής δικτύωσης;

Ο όρος μέσα κοινωνικής δικτύωσης (*social media*) αναφέρεται στη δημιουργία και ανταλλαγή πληροφοριών μέσω των διαφόρων διαδικτυακών κοινοτήτων (Ahlqvist, Asta, Halonen, & Heinonen, 2008). Ουσιαστικά μέσω των social media πραγματοποιείται η ένωση της διαδικτυακής κοινωνικοποίησης με τις διάφορες μορφές της πληροφορίας, όπως το κείμενο, οι φωτογραφίες και το video. Επομένως, από τον ορισμό τους και μόνο, αντιλαμβανόμαστε ότι μέσω της επικοινωνίας των χρηστών τους, προκύπτει μία πληθώρα δεδομένων, η οποία είναι ικανή να αξιοποιηθεί από διάφορους φορείς, όπως για παράδειγμα κοινωνιολογικούς, πολιτικούς, εμπορικούς και οικονομικούς.

Προκειμένου κάτι τέτοιο να γίνει αντιληπτό, αρκεί να αναλογιστούμε για παράδειγμα ότι στο Twitter (Stross, 2010) κοινοποιούνται ημερησίως εκατομμύρια μηνύματα, γεγονός που δεν αφήνει ασυγκίνητους πολλούς ιστορικούς, σύμφωνα με τους οποίους αν μελετήσουμε συλλογικά τον μεγάλο όγκο πληροφοριών που μας παρέχεται μπορούμε να καταγράψουμε επαρκώς και με κάθε λεπτομέρεια την ιστορία των τελευταίων χρόνων. Παράλληλα η αναζήτηση ενός γεγονότος καθίσταται μια εξαιρετικά εύκολη διαδικασία, λόγω της ψηφιακής μορφής των πληροφοριών, σε αντίθεση με το παρελθόν όπου η αναζήτηση σε βιβλία και εφημερίδες ήταν αρκετά περίπλοκη και χρονοβόρα. Τέλος αξίζει να σημειωθεί ότι λόγω του τεράστιου όγκου των χρηστών του Twitter, διαμορφώνεται ένα μεγάλο φάσμα απόψεων που πηγάζουν από διάφορα κοινωνικά και οικονομικά στρώματα, πολιτικές πεποιθήσεις και τρόπους σκέψης. Το γεγονός αυτό αποτελεί ισχυρό εργαλείο στα χέρια των ιστορικών, μέσω του οποίου είναι δυνατή η σκιαγράφηση ενός γεγονότος μέσα από διαφορετικές αντιλήψεις και από διαφορετικές οπτικές γωνίες. Κατά το παρελθόν κάτι τέτοιο ήταν σαφώς αδύνατο, καθώς η κύρια πηγή άντλησης πληροφοριών για τους ιστορικούς, κατά τους προηγούμενους αιώνες, δεν ήταν άλλη από τις εφημερίδες οι οποίες προφανώς φέρουν μονάχα τη γνώμη συντακτών και δημοσιογράφων. Το συμπέρασμα που βγάζουμε από το παραπάνω παράδειγμα είναι ότι, εκτός από το γεγονός ότι τα social media επηρεάζουν σημαντικά την καθημερινότητά μας, είναι επίσης ικανά να προσθέσουν ένα ισχυρό εργαλείο για την πραγματοποίηση μελετών, λόγω του μεγάλου όγκου δεδομένων που αυτά διαθέτουν.

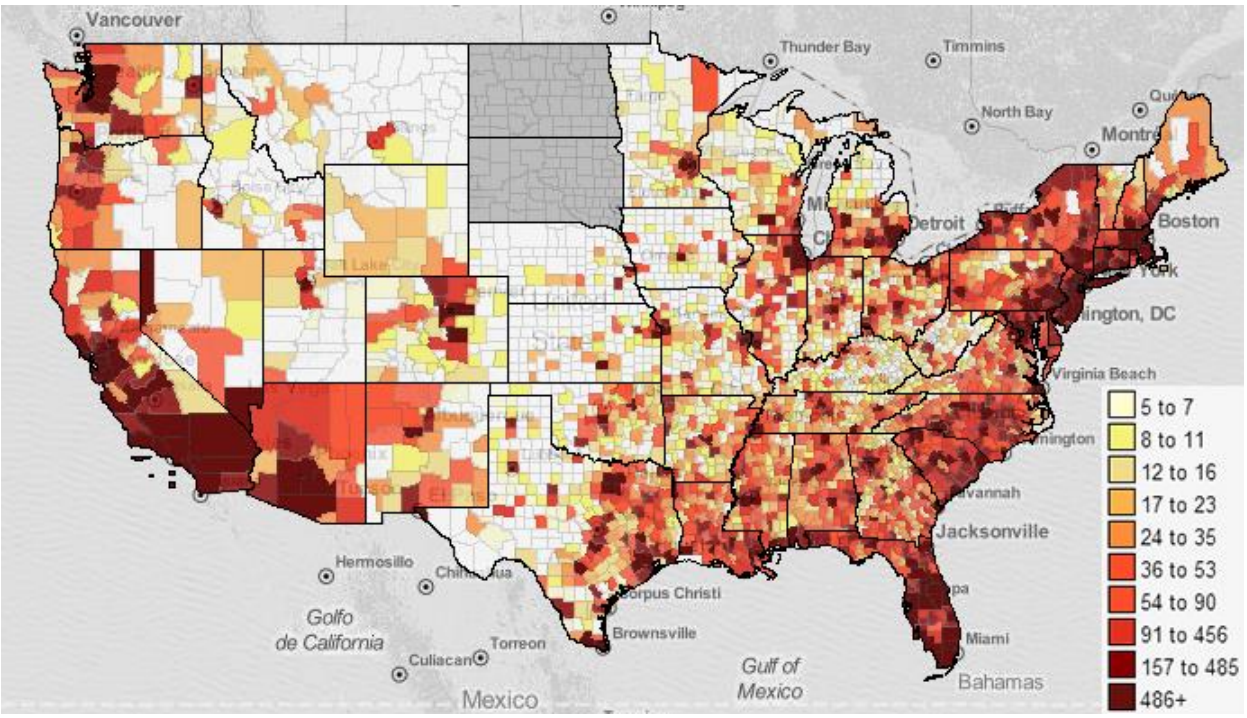
1.2 Τα οφέλη της αξιοποίησης των μέσων κοινωνικής δικτύωσης

Δεν είναι δύσκολο να αντιληφθεί κάποιος ότι η χρήση των μέσων κοινωνικής δικτύωσης, στη σύγχρονη κοινωνία, τείνει πολλές φορές να ξεφεύγει από τον ορισμό που δόθηκε στην προηγούμενη παράγραφο. Έτσι λοιπόν ενώ αρχικά τα social media στόχευαν μόνο στην αλληλεπίδραση μεταξύ των χρηστών τους, παρατηρούμε πλέον ότι το εύρος της χρήσης τους έχει επεκταθεί σε πολύ μεγάλο βαθμό. Χαρακτηριστικό είναι το παράδειγμα των εταιρειών που χρησιμοποιούν όλο και πιο ενεργά τα social media για διάφορους σκοπούς, όπως η προώθηση των προϊόντων και των υπηρεσιών τους, κυρίως μέσω του Facebook και του Youtube, καθώς και η αναζήτηση ανθρώπινου δυναμικού για την κάλυψη θέσεων εργασίας, κυρίως μέσω του LinkedIn.

Παράλληλα τονίζεται ότι τα μέσα κοινωνικής δικτύωσης μπορούν να συνδράμουν και σε πολυπλοκότερα πεδία, όπως αυτό της επιστήμης. Για παράδειγμα οι επιστήμονες δηλώνουν (Rivero, 2014) ότι το Twitter μπορεί να αποτελέσει χρησιμότατο εργαλείο για τον εντοπισμό των κρουσμάτων του HIV και μελλοντικά θα μπορεί ακόμη και να προλαμβάνει την εξάπλωση του ιού. Πιο συγκεκριμένα μία ομάδα επιστημόνων από το UCLA (*University of California, Los Angeles*) πραγματοποίησε μία μελέτη, χρησιμοποιώντας το Twitter, στην οποία συλλέχθηκαν 550 εκατομμύρια Tweets τα οποία δημοσιοποιήθηκαν από τις 26 Μαΐου έως τις 9 Δεκεμβρίου του 2012 στις ΗΠΑ. Στη συνέχεια κατασκεύασαν έναν αλγόριθμο, μέσω του οποίου είναι δυνατή η πρόβλεψη επικίνδυνων σεξουαλικών συμπεριφορών. Οι Αμερικανοί ερευνητές εστίασαν την προσοχή τους και ανέλυσαν περίπου 10.000 Tweets. Έτσι λοιπόν τοποθετώντας την προέλευσή τους στο χάρτη (Βλ. Εικόνα 1.1), τα συνέκριναν με τη γεωγραφική κατανομή των κρουσμάτων λοίμωξης από τον HIV. Τα γεωγραφικά δεδομένα, τα οποία χρησιμοποιήθηκαν από τους ερευνητές και αφορούν τον αριθμό κρουσμάτων HIV στις ΗΠΑ, προήλθαν (AIDSVu, 2010) από έναν διαδραστικό χάρτη[a], ο οποίος απεικονίζει την εξάπλωση του ιού (Βλ. Εικόνα 1.2). Η έρευνα εστίασε στη σύγκριση των δεδομένων των δύο χαρτών, έτσι ώστε να αποφανθεί αν είναι δυνατή η παρακολούθηση της εξάπλωσης των κρουσμάτων του ιού, μέσω του των μέσων κοινωνικής δικτύωσης και συγκεκριμένα του Twitter.

[a] Τα δεδομένα του χάρτη που χρησιμοποίησαν οι ερευνητές προέρχονται από την ιστοσελίδα *AIDSVu.org* και αφορούν το έτος 2009. Στην Εικόνα 1.2 φαίνεται ο αντίστοιχος ανανεωμένος χάρτης του 2010. Σημειώνεται ότι το τμήμα υγείας των πολιτειών που είναι χρωματισμένες με γκρι χρώμα (Βόρεια και Νότια Ντακότα), δεν δημοσιοποίησαν τα δεδομένα τους στη συγκεκριμένη ιστοσελίδα.

Εικόνα 1.1 - Προέλευση των επικίνδυνων Tweets, που σχετίζονται με τον HIV



Εικόνα 1.2 – Κρούσματα HIV στις ΗΠΑ (2010)

Χρησιμοποιώντας στατιστικά μοντέλα, το συμπέρασμα των ερευνητών, ήταν ότι υπάρχει σημαντική σχέση μεταξύ των δεδομένων που προέκυψαν από τη μελέτη των Tweets, και των πολιτειών που αναφέρονται τα περισσότερα κρούσματα λοιμώξεων από τον ιό HIV.

Ειδικότερα, όπως φαίνεται και στην Εικόνα 1.1, παρατηρήθηκε ότι μεγαλύτερα ποσοστά των επικίνδυνων Tweets, σε σχέση με τον συνολικό τους αριθμό, εμφανίστηκαν στις εξής πολιτείες: Καλιφόρνια (9.4%), Τέξας (9,0%), Νέα Υόρκη (5,7%) και Φλόριντα (5,4%). Παράλληλα, στην Εικόνα 1.2 παρατηρούμε ότι οι συγκεκριμένες πολιτείες έχουν ιδιαίτερα υψηλό αριθμό κρουσμάτων HIV. Έτσι λοιπόν εμφανίζεται μία συμμεταβολή του αριθμού των σημείων και στους δύο χάρτες, γεγονός που επαληθεύει το συμπέρασμα της έρευνας. Σε αυτό το σημείο τονίζεται ότι σύμφωνα με άλλες έρευνες (HIV Surveillance Report, 2011) σε δύο από τις παραπάνω πολιτείες, και συγκεκριμένα στο Τέξας και στη Φλόριντα, δεν είναι υποχρεωτική η παράδοση μαθημάτων σεξουαλικής διαπαιδαγώγησης από τα δημόσια σχολεία, καθώς και η παροχή πληροφοριών, που σχετίζονται με τον HIV, στους μαθητές.

Τα δεδομένα του παραπάνω παραδείγματος αποτέλεσαν την πρώτη μελέτη που αποδεικνύει ότι το Twitter αποτελεί μια αποτελεσματική μέθοδο για την παρακολούθηση των κρουσμάτων και την πρόβλεψη του κινδύνου εξάπλωσης και ουσιαστικά αποδεικνύουν ότι τα κοινωνικά δίκτυα μπορούν να δώσουν σημαντικές πληροφορίες για την εξάπλωση σοβαρών λοιμώξεων. Γενικεύοντας λοιπόν, τα social media αποτελούν μία αστείρευτη πηγή δεδομένων, τα οποία αν αναλυθούν κατάλληλα είναι ικανά να συμβάλλουν στην ανάπτυξη ενός μεγάλου φάσματος της κοινωνίας, το οποίο περιλαμβάνει και εξαιρετικά πολύπλοκα πεδία, όπως αυτό της επιστημονικής προόδου.

Κεφάλαιο 2. Big Data

2.1 Η εκτίναξη του όγκου πληροφοριών στις μέρες μας

Πίσω στο 2000, κατά την έναρξη του προγράμματος SDSS (*Sloan Digital Sky Survey*), το τηλεσκόπιο που ήταν τοποθετημένο στην αμερικανική πολιτεία Νέο Μεξικό, κατάφερε να συγκεντρώσει (Cukier, 2010) μέσα στις πρώτες βδομάδες λειτουργίας του περισσότερα δεδομένα από αυτά που είχαν συλλεχθεί κατά τη διάρκεια ολόκληρης την ιστορίας της αστρονομίας. Ύστερα από μία δεκαετία λειτουργίας του η τελική απεικόνιση κάλυψε 14.555 τετραγωνικές μοίρες, που αναλογεί στο 35% του ουρανού, ενώ συγκεντρώθηκαν 140 terabytes πληροφοριών σε μορφή εικόνας. Ο διάδοχος του SDSS αναμένεται να τεθεί σε λειτουργία το 2016 στη Χιλή και υπολογίζεται ότι θα αποκτά το συγκεκριμένο όγκο πληροφοριών σε μόλις πέντε ημέρες.

Ένα δεύτερο παράδειγμα είναι η ημερήσια αύξηση του όγκου των φωτογραφιών του Facebook, καθώς σύμφωνα (Smith, 2013) με τα στατιστικά του δημοφιλούς social media, που αφορούν τους πρώτους μήνες του 2013, οι χρήστες ανεβάζουν καθημερινά 350 εκατομμύρια φωτογραφίες. Αυτές προστίθενται στις ήδη υπάρχουσες, με αποτέλεσμα το Facebook να αριθμεί 240 δισεκατομμύρια φωτογραφίες μέσα σε μόλις εννέα χρόνια ζωής.

Το συμπέρασμα από τα παραπάνω παραδείγματα είναι ότι βρισκόμαστε σε μία εποχή κατά την οποία παράγουμε μία υπερβολικά μεγάλη ποσότητα ψηφιακών πληροφοριών. Η κύρια μάζα των πληροφοριών αυτών, παράγεται από τους χρήστες του διαδικτύου και διαδίδεται μάλιστα με υπερβολικά μεγάλη ταχύτητα κατά κύριο λόγω μέσω των social media. Επομένως, η συγκεκριμένη διαδικασία της ολοένα αυξανόμενης πληροφορίας, καθώς και η αξιοποίηση αυτής, μπορεί να αποτελέσει κομβικό παράγοντα σε τομείς όπως η πρόληψη ασθενειών, η καταπολέμηση του εγκλήματος, η πρόοδος της επιστήμης και της έρευνας, η ανάπτυξη των επιχειρήσεων και ούτω καθεξής.

Είναι εύκολο λοιπόν να αντιληφθούμε ότι στις μέρες μας ο όγκος των πληροφοριών που παράγουμε μεγαλώνει συνεχώς και μάλιστα εκθετικά, γεγονός που μας αναγκάζει να τον διαχειριζόμαστε με διαφορετικό τρόπο σε σχέση με το παρελθόν. Οι πληροφορίες αυτές εκτός από τη μεγάλη τους έκταση χαρακτηρίζονται και από έντονη ποικιλομορφία καθώς καλύπτουν

κάθε πτυχή της κοινωνίας και της καθημερινότητας μας. Τα δεδομένα, τα οποία δεν είναι πια περιορισμένα σε όγκο και ποικιλία, είναι γνωστά ως Big Data.

Μερικά παραδείγματα των Big Data είναι τα εξής:

- 340 εκατομμύρια Tweets δημιουργούνται σε ημερήσια βάση, που αναλογούν σε περίπου 4.000 Tweets το δευτερόλεπτο.
- Πάνω από ένα δισεκατομμύριο ενεργοί χρήστες του Facebook παράγουν πληροφορίες, αλληλεπιδρώντας καθημερινά μεταξύ τους.
- 70 ώρες βίντεο προστίθενται στο Youtube κάθε λεπτό.
- Περισσότεροι από πέντε δισεκατομμύρια άνθρωποι συνομιλούν μέσω των κινητών τηλεφώνων τους μέσω φωνητικών κλήσεων, γραπτών μηνυμάτων και βιντεοκλήσεων.
- Κάθε δευτερόλεπτο πραγματοποιούνται 10.000 συναλλαγές μέσω πιστωτικών καρτών.
- Έχει υπολογιστεί (W. P. Carey School of Business, 2012) ότι ο όγκος δεδομένων των επιχειρήσεων παγκοσμίως διπλασιάζεται κάθε περίπου 1,2 χρόνια.
- Η Αμερικάνικη πολυεθνική εταιρεία Wal-Mart πραγματοποιεί κάθε ώρα περισσότερες από ένα εκατομμύριο συναλλαγές, τα δεδομένα των οποίων αποθηκεύονται σε βάσεις δεδομένων, οι οποίες περιέχουν 2560 terabytes δεδομένων, τα οποία προσεγγιστικά είναι 167 φορές περισσότερα από τη βιβλιοθήκη του Κογκρέσου των Ηνωμένων Πολιτειών, που θεωρείται η μεγαλύτερη και πλουσιότερη βιβλιοθήκη του κόσμου.

2.2 Τί είναι τα Big Data;

Τα Big Data και η αξιοποίησή τους αποτελούν μια από τις σημαντικότερες, αλλά και παράλληλα περίπλοκες προκλήσεις για τη σημερινή κοινωνία και τις διάφορες πτυχές της, όπως για παράδειγμα τη βιομηχανία, τις επιχειρήσεις και τους μη κερδοσκοπικούς οργανισμούς. Τα δεδομένα αυτά είναι αδύνατο να αναλυθούν και να επεξεργαστούν με τα συστήματα των συμβατικών βάσεων δεδομένων, καθώς είναι πολύ μεγάλα σε όγκο, ή διακινούνται εξαιρετικά γρήγορα, ή απλά δεν ταιριάζουν με την αρχιτεκτονική των παραδοσιακών βάσεων δεδομένων.

Επομένως, προκειμένου να κατανοήσουμε τι ακριβώς είναι τα Big Data, θα μπορούσαμε να πούμε ότι είναι τα δεδομένα και οι πληροφορίες που εκτείνονται στις παρακάτω τρεις διαστάσεις:

- **Μεγάλος όγκος δεδομένων**
  Για παράδειγμα, το τμήμα μάρκετινγκ μιας επιχείρησης καλείται να αναλύσει δεκάδες terabytes από Tweets, που δημιουργούνται από τους χρήστες, μέσα σε μία μόλις μέρα, προκειμένου να προσεγγίσουν τις αντιδράσεις τους γύρω από τα παραγόμενα προϊόντα.
- **Μεγάλη ταχύτητα διάδοσης δεδομένων**
  Για παράδειγμα, μερικά μόλις δευτερόλεπτα πιθανότατα αποτελούν εξαιρετικά μεγάλο χρονικό διάστημα, όταν πρόκειται για ευαίσθητες χρονικά διεργασίες, όπως η ανάλυση των πληροφοριών για την εξακρίβωση μίας οικονομικής απάτης. Επομένως γίνεται αντιληπτό ότι καθημερινά απαιτείται διερεύνηση εκατομμυρίων λογαριασμών και συναλλαγών σε πραγματικό χρόνο, προκειμένου να εντοπιστούν πιθανές απάτες.
- **Μεγάλη ποικιλία τύπων δεδομένων**
  Τα Big Data δεν είναι απαραίτητο να βρίσκονται σε μορφή κειμένου. Πληροφορίες μπορούμε να αντλήσουμε και από άλλους τύπους δεδομένων όπως από εικόνες, βίντεο, ήχους, αποτελέσματα αισθητήρων (π.χ. ένα διάγραμμα θερμοκρασίας συναρτήσει του χρόνου, μέσω ενός θερμομέτρου), clickstreams (Moe & Fader, 2004) (η καταγραφή των κλικ ενός χρήστη κατά τη διάρκεια της περιήγησης του στο διαδίκτυο), καθώς και πολλούς άλλους. Για παράδειγμα η παρατήρηση σε ζωντανή μετάδοση εκατοντάδων βίντεο από κάμερες επιτήρησης.

Τα παραπάνω χαρακτηριστικά, τα οποία είναι γνωστά (O'Reilly, 2012) και ως *"τα τρία V των Big Data"* από τις λέξεις *Volume*, *Velocity* και *Variety*, αποτελούν τα θεμελιώδη στοιχεία των Big Data και ουσιαστικά προσδιορίζουν τον χαρακτήρα τους. Παράλληλα μέσω αυτών των τριών εργαλείων

καθίσταται δυνατός ο προσδιορισμός της φύσης της κάθε πληροφορίας, άρα και κατ' επέκταση η εκμετάλλευσή της.

Πίσω από τα Big Data όμως υποβόσκει ένας σοβαρός κίνδυνος, καθώς η εξόρυξη και η ανάλυσή τους έχει αποδειχθεί εξαιρετικά περίπλοκη διαδικασία. Αυτό συμβαίνει κατά κύριο λόγο διότι τα δεδομένα που υπάρχουν, λόγω του τεράστιου όγκου τους και της ποικιλομορφίας τους, είναι σε μεγάλο βαθμό αδόμητα, με αποτέλεσμα να απαιτείται εξαιρετικά πολύς χρόνος προκειμένου να αναλυθούν. Ένα χαρακτηριστικό παράδειγμα που θα μπορούσαμε να σκεφτούμε είναι ότι αν κάποιος χρήστης του Facebook γράψει το μήνυμα "*Είμαι νευριασμένος*" στη προσωπική του σελίδα, δεν θα μπορέσουμε άμεσα να ερμηνεύσουμε τους λόγους για τους οποίους ισχύει αυτό, τη σοβαρότητα της κατάστασης ή ακόμη και αν το μήνυμα είναι σοβαρό ή χιουμοριστικό και άρα ψευδές. Επομένως αν αναλογιστούμε τα χιλιάδες μηνύματα που δημιουργούνται κάθε δευτερόλεπτο, και πόσες πιθανές αποκλίσεις και ιδιαιτερότητες μπορεί να έχει το καθένα, καταλαβαίνουμε ότι είναι εξαιρετικά πιθανό να οδηγηθούμε σε λάθος συμπεράσματα όταν κληθούμε να τα αναλύσουμε.

Ένας δεύτερος παράγοντας, που καθιστά τα Big Data δυσπρόσιτες πηγές πληροφοριών, είναι τα μέσα που διαθέτουν οι οργανισμοί προκείμενου να τα αναλύσουν. Πιο συγκεκριμένα η εταιρεία Hewlett Packard πραγματοποίησε τον Απρίλιο του 2013 μία παγκόσμια έρευνα (Έρευνα της HP, 2013), σύμφωνα με την οποία περισσότεροι από ένας στους δύο εκ των εργαζομένων σε επιχειρήσεις, που διεξάγουν ανάλυση Big Data, πιστεύουν ότι οι οργανισμοί τους δεν διαθέτουν τον κατάλληλο εξοπλισμό, αλλά και τις σωστές ιδέες και γνώσεις, προκειμένου να αναλύσουν έναν τόσο μεγάλο και ποικιλόμορφο όγκο δεδομένων. Μάλιστα νέα έρευνα, της ίδιας εταιρείας τον Μάιο του 2013 με τίτλο "*Big Data and Cloud*", αποκαλύπτει ότι ένας στου τρεις οργανισμούς που προσπάθησε να πραγματοποιήσει ανάλυση των συγκεκριμένων δεδομένων απέτυχε.

Ωστόσο, παρά το γεγονός ότι η αξιοποίηση των Big Data κρύβει σημαντική πιθανότητα αποτυχίας, σύμφωνα με έρευνες το 60% των οργανισμών που συμμετείχαν διατίθενται, να ξοδέψει το 10% του προϋπολογισμού τους πάνω στη ανάλυση αυτών των πληροφοριών.

Ο επικεφαλής του τμήματος software στην εταιρεία Hewlett Packard αναφέρει χαρακτηριστικά: "*Τα big data επιτρέπουν στους οργανισμούς να επωφεληθούν από το σύνολο των πληροφοριών τους, τόσο των εσωτερικών όσο και των εξωτερικών, σε πραγματικό χρόνο. Παράγουν εξαιρετικά γρήγορη διαδικασία λήψης αποφάσεων και, ως αποτέλεσμα, μοναδικούς και*

καινοτόμους τρόπους για την προσφορά υπηρεσιών στους πελάτες και την κοινωνία". Το άμεσο ερώτημα επομένως που προκύπτει είναι ποιά είναι τα οφέλη από την αξιοποίηση του ατέρμονου αυτού όγκου δεδομένων. Στη συνέχεια θα αναφερθούν ορισμένα παραδείγματα εκμετάλλευσης των Big Data, καθώς και η συνεισφορά της σε κάθε πτυχή της σύγχρονης κοινωνίας, όπως για παράδειγμα η επιστήμη, η υγεία και η οικονομία.

2.3 Ιστορική αναδρομή των Big Data

Η εμφάνιση των Big Data χρονολογείται πολλά χρόνια πριν την σημερινή εκτίναξη του όγκου των πληροφοριών. Η απαρχή τους πραγματοποιείται κατά την δεκαετία του 1940, και πιο συγκεκριμένα το 1941 όταν για πρώτη φορά εμφανίζεται ο όρος[a] "*information explosion*". Τα παρακάτω αποτελούν τα βασικά ιστορικά ορόσημα της εξέλιξης της ιδέας των Big Data, καθώς και των πληροφοριών γενικότερα:

- Το 1944 ο συγγραφέας και βιβλιοθηκάριος του Wesleyan University, Fremont Rider αναφέρει στο βιβλίο του "*The Scholar and the Future of the Research Library. A problem and Its Solution.*" (Rider, 1944) ότι οι βιβλιοθήκες των πανεπιστημίων της Αμερικής διπλασιάζουν τον όγκο των βιβλίων τους κάθε δεκαέξι χρόνια. Με βάση τη συγκεκριμένη παρατήρηση ο Rider εκτίμησε ότι το 2040 η βιβλιοθήκη του πανεπιστημίου του Yale θα κατέχει περίπου 200 εκατομμύρια βιβλία, τα οποία θα καταλαμβάνουν ράφια συνολικού μήκους 6000 μιλίων.
- Το 1949 ο μαθηματικός και ηλεκτρολόγος μηχανικός Claude Shannon (James, 2009), ο οποίος είναι γνωστός ως "*ο πατέρας της θεωρίας της πληροφορίας*" και της σύγχρονης κρυπτογραφίας, μελέτησε τη χωρητικότητα διαφόρων μέσων αποθήκευσης σε bits. Για παράδειγμα εστίασε τη προσοχή του στα δεδομένα που περιέχονται στη διάτρητη κάρτα[b] (*punch card*), στη φωτογραφία, ενώ παράλληλα αξιοσημείωτο είναι το γεγονός ότι είναι ο πρώτος που ανέφερε ότι το γονιδίωμα στη μοριακή βιολογία είναι ουσιαστικά ένα μέσο αποθήκευσης πληροφοριών, που μετριέται σε bits. Στην κορυφή της λίστας των αντικειμένων που μελέτησε ο Shannon βρέθηκε η βιβλιοθήκη του Κογκρέσου των Ηνωμένων Πολιτειών, η οποία σύμφωνα με τη μελέτη του ισοδυναμούσε με 100 τρισεκατομμύρια bits.
- Το 1961 ο Derek Price εκδίδει το βιβλίο "*Science Since Babylon*" στο οποίο μελετά τη πρόοδο της επιστημονικής γνώσης, μετρώντας τον αριθμό των

[a] Σύμφωνα με το *Oxford English Dictionary* ο όρος "information explosion" εμφανίζεται πρώτη φορά το 1941.

[b] Η διάτρητη κάρτα (Cortada, 1993) αποτελεί την παλαιότερη μορφή αποθηκευτικού μέσου και είναι εφεύρεση του Herman Hollerith, ο οποίος χρησιμοποιούσε διάτρητο χαρτί προκειμένου να αποθηκεύει επαναλαμβανόμενα μοτίβα για τη σχεδίαση ρούχων. Ο Herman Hollerith στις 23 Σεπτεμβρίου 1884 χρησιμοποίησε για πρώτη φορά τον συγκεκριμένο τρόπο, προκειμένου να αποθηκεύσει δεδομένα. Οι κάρτες αυτές χρησιμοποιήθηκαν μέχρι τα τέλη της δεκαετίας του 1970, για την αποθήκευση εντολών και ρυθμίσεων διαφόρων αυτοματοποιημένων μηχανών.

επιστημονικών (Price, 1961) άρθρων και δημοσιεύσεων, συναρτήσει του χρόνου. Μέσα από τα στοιχεία που συλλέγει, βγάζει το συμπέρασμα ότι ο αριθμός των νέων δημοσιεύσεων δεν αυξάνει γραμμικά, αλλά εκθετικά και συγκεκριμένα παρατηρεί ότι τα επιστημονικά άρθρα διπλασιάζονται κάθε δεκαπέντε έτη, ή ισοδύναμα δεκαπλασιάζονται κάθε πενήντα. Σύμφωνα με τον Price, αυτό συμβαίνει διότι κάθε επιστημονική πρόοδος δημιουργεί πολλά νέα πεδία, τα οποία προσφέρουν εύφορο έδαφος για περαιτέρω έρευνα. Το άμεσο συμπέρασμα είναι επομένως ότι το εύρος της επιστημονικής έρευνας μεγαλώνει ολοένα και περισσότερο και μάλιστα με συνεχώς αυξανόμενους ρυθμούς, με αποτέλεσμα να αυξάνουν με ανάλογο τρόπο και οι επιστημονικές ανακαλύψεις, ανά μονάδα χρόνου.

- Το 1975 ξεκινά στην Ιαπωνία, υπό την επίβλεψη του υπουργείου που είναι αρμόδιο για τις τηλεπικοινωνίες και τις ταχυδρομικές συναλλαγές της χώρας, το πρόγραμμα "*Information Flow Census*" το οποίο καταγράφει την ετήσια ροή των πληροφοριών σε ολόκληρη την Ιαπωνία. Το IFC καθιέρωσε το "*word"* σαν μονάδα μέτρησης της ροής μίας πληροφορίας, το οποίο όπως γίνεται εύκολα αντιληπτό από το όνομά του, μετρά το μέγεθος της πληροφορίας με βάση τον αριθμό των λέξεων που αυτή περιέχει. Για παράδειγμα, σύμφωνα με τα στοιχεία του IFC, ένα λεπτό προφορικού διαλόγου ισοδυναμεί με 71 words, το άκουσμα ενός λεπτού μουσικής με 120 words, ενώ η παρακολούθηση ενός λεπτού έγχρωμης τηλεόρασης με 672 words. Παράλληλα το πρόγραμμα κατηγοριοποιεί (Akiyama, 2003) τις πληροφορίες, που μετρώνται σε words, στα εξής πέντε ήδη: Αρχική πληροφορία (*Original information*), Μεταδιδόμενη πληροφορία (*Transmitted information*), Διαθέσιμη πληροφορία (*Available information*), Αναλώσιμη πληροφορία (*Consumable information*), Καταναλωμένη πληροφορία (*Consumed information*). Από τις τρεις τελευταίες, εκ των πέντε παραπάνω κατηγοριών, είναι δυνατός ο προσδιορισμός του ρυθμού με τον οποίο καταναλώνονται οι πληροφορίες. Προκειμένου αυτό να γίνει αντιληπτό αρκεί να αναλογιστούμε ότι στην περίπτωση της τηλεόρασης η "*διαθέσιμη πληροφορία*" είναι το σύνολο όλων των εκπομπών του τηλεοπτικού προγράμματος, η "*αναλώσιμη πληροφορία*" είναι οι εκπομπές ενός μόνο τηλεοπτικού καναλιού (καθώς ο τηλεθεατής προφανώς δεν μπορεί να παρακολουθεί περισσότερα από ένα πρόγραμμα ταυτοχρόνως) και τέλος η "*καταναλωμένη πληροφορία*" είναι οι εκπομπές που τελικά ο τηλεθεατής παρακολούθησε. Σύμφωνα λοιπόν με τα στοιχεία που συλλεχθήκαν το 1975

από το IFC, οι διαθέσιμες πληροφορίες αυξάνονται με πολύ μεγαλύτερο ρυθμό, συγκριτικά με αυτόν που καταναλώνονται. Παράλληλα, σύμφωνα με τις μετρήσεις από το 1978 και ύστερα, παρατηρήθηκε μία έντονη στασιμότητα όσον αφορά τη ζήτηση των πληροφοριών που προέρχονται από τα μέσα μαζικής ενημέρωσης, τα οποία αποτελούν ένα είδος *μονόδρομης*[a] επικοινωνίας. Αντίστροφα παρατηρήθηκε θεαματική αύξηση στη ζήτηση πληροφοριών, που εκφράζονται μέσω *αμφίδρομων*[b] τρόπων επικοινωνίας, όπως για παράδειγμα η τηλεφωνική επικοινωνία, η αλληλογραφία και αργότερα η επικοινωνία μέσω των κινητών τηλεφώνων και του διαδικτύου.

- Το 1981 η Στατιστική Υπηρεσία της Ουγγαρίας ξεκινά ένα ερευνητικό πρόγραμμα (Hilbert, 2012), μέσω το οποίου παρακολουθείται και μετριέται ο συνολικός όγκος των πληροφοριών, που παράγουν οι υπηρεσίες της χώρας. Στις υπηρεσίες αυτές συγκαταλέγονται βιβλιοθήκες, σχολεία, πανεπιστήμια, μέσα μαζικής ενημέρωσης, νοσοκομεία, βιομηχανίες, κυβερνητικοί οργανισμοί και ούτω καθεξής. Τα δεδομένα αυτά μετρώνται σε bits και η συγκεκριμένη έρευνα συνεχίζει να πραγματοποιείται μέχρι και σήμερα.
- Τον Αύγουστο του 1983 ο κοινωνιολόγος Ithiel de Sola Pool δημοσιεύει στο περιοδικό Science το άρθρο "*Tracking Flow of Information*" (Pool, 1983) στο οποίο εξετάζει τους ρυθμούς ανάπτυξης, από το 1960 έως το 1977, των 17 μεγαλύτερων μέσων επικοινωνίας της Αμερικής, όπως για παράδειγμα του ραδιοφώνου, της τηλεόρασης, των εφημερίδων, των περιοδικών, και ούτω καθεξής. Πιο συγκεκριμένα παρατηρεί ότι οι συνολικές ετήσιες λέξεις που διατίθενται στους Αμερικάνους (άνω των 10 ετών), από τα μέσα αυτά, αυξάνει εκθετικά με ρυθμό 8,9% κάθε χρόνο. Παράλληλα όμως αναφέρει ότι οι συνολικές λέξεις, οι οποίες τελικά απορροφήθηκαν από τους Αμερικάνους (βλέποντας τηλεόραση, ακούγοντας ραδιόφωνο, διαβάζοντας εφημερίδες και περιοδικά, κλπ), αυξάνονταν κατά μόλις 2,9% ανά έτος. Επομένως, όπως γίνεται εύκολα αντιληπτό από τις παρατηρήσεις του Ithiel de Sola Pool, κατά το χρονικό διάστημα 1960-1977 δημιουργήθηκε μία έκρηξη μετάδοσης πληροφοριών από τα μέσα ενημέρωσης των ΗΠΑ, αλλά και παράλληλα ένα

[a] Κατά την μονόδρομη επικοινωνία παρατηρείται μεταφορά της πληροφορίας από τον πομπό στον δέκτη.

[b] Κατά την αμφίδρομη επικοινωνία παρατηρείται μετάδοση πληροφορίας και από τα δύο πρόσωπα που αλληλεπιδρούν, με αποτέλεσμα ο πομπός και ο δέκτης να μην είναι σταθεροί, αλλά να εναλλάσσονται συνεχώς.

συνεχώς αυξανόμενο χάσμα ανάμεσα στις πληροφορίες που εκπέμπονταν και στις πληροφορίες που τελικά έφταναν για απορρόφηση στους δέκτες τους.

- Τον Ιούλιο του 1986 ο συγγραφέας και κοινωνιολόγος Hal Becker αναφέρει στο άρθρο του (Becker, 1986) "*Can users really absorb data at today's rates? Tomorrow's?*" ότι οι πήλινες επιγραφές, που χρησιμοποιούσαν οι Σουμέριοι το 4000 π.Χ., εμπεριείχαν ένα σύμβολο ανά κυβική ίντσα. Στη συνέχεια τονίζει ότι ο εφευρέτης Ιωάννης Γουτεμβέργιος (Johaness Gutenberg), ο οποίος γύρω στο 1439 εισήγαγε την τυπογραφία στην Ευρώπη (Childress, 2008), κατάφερε με την εφεύρεσή του να αυξήσει δραματικά τη πυκνότητα των συμβόλων (γραμμάτων), η οποία πλέον εκτοξεύτηκε στους 500 χαρακτήρες ανά κυβική ίντσα. Τέλος αξίζει να σημειωθεί ότι μία μνήμη RAM που κατασκευάστηκε το 2000 αποθηκεύει $1{,}25 \cdot 10^{11}$ bytes[a] ανά κυβική ίντσα, γεγονός που μας υποδεικνύει ότι η πυκνότητα των δεδομένων που αποθηκεύονται ανά μονάδα όγκου, όχι απλώς αυξάνει, αλλά αυξάνει εκθετικά και μάλιστα με ραγδαίους ρυθμούς.
- Από το 1996 και μετά το κόστος της αποθήκευσης αρχείων και δεδομένων σε ψηφιακή μορφή γίνεται πλέον αποδοτικότερο (Morris & Truskowski, 2003), συγκριτικά με το αντίστοιχο κόστος αποθήκευσής τους σε χαρτί.
- Το 1997 ο Michael Lesk, στη δημοσίευσή του "*How much information is there in the world?*" (Lesk, 1997), υπογραμμίζει ότι οι συνολικές πληροφορίες, μέχρι εκείνη την εποχή, ανέρχονται σε μερικά petabytes (το ένα petabyte είναι ίσο με 10^{15} bytes, ή ισοδύναμα ένα εκατομμύριο gigabytes). Παράλληλα παρατηρεί ότι η παραγωγή των μέσων αποθήκευσης εκείνων των ημερών (κασέτες, σκληροί δίσκοι, CD, δισκέτες, κ.ο.κ.) θα καλύψει και θα προσπεράσει τα συγκεκριμένα μεγέθη μέχρι το έτος 2000. Το αποτέλεσμα θα είναι όλα τα παραγόμενα αλλά και ήδη υπάρχοντα δεδομένα να μπορούν να αποθηκευτούν και έτσι να μην χάνεται το παραμικρό ίχνος πληροφορίας.
- Τον Οκτώβριο του 1997 οι ερευνητές της NASA, Michael Cox και David Ellsworth, αναφέρουν στη δημοσίευσή τους "*Application controlled demand paging for out of core visualization*" (Cox & Ellsworth, 1997) για πρώτη φορά στην ιστορία τον όρο Big Data. Πιο συγκεκριμένα τονίζουν ότι η δραματική

---

[a] Σημείωση προκειμένου να γίνει αντιληπτή η σχέση μεταξύ των bytes και των χαρακτήρων: Ένας χαρακτήρας του κώδικα ASCII καταλαμβάνει στη μνήμη 1 byte. Αυτό ωστόσο δεν αποτελεί γενική περίπτωση, καθώς στο πρότυπο Ubicode για παράδειγμα ένας χαρακτήρας μπορεί να αντιστοιχεί και σε περισσότερα από 1 bytes (π.χ. στο UTF-16 και στο UTF-32 αντιστοιχεί σε 2 και 4 bytes αντίστοιχα).

αύξηση του όγκου των δεδομένων αποτελεί μείζον πρόβλημα, αλλά παράλληλα και πρόκληση, για τα συστήματα ηλεκτρονικών υπολογιστών της εποχής εκείνης. Παράλληλα υπογραμμίζουν ότι οι δυνατότητες της μνήμης και των σκληρών δίσκων των υπολογιστών εκείνων των ημερών, αδυνατούσε να ανταπεξέλθει στην ολοένα αυξανόμενη ποσότητα πληροφοριών, με αποτέλεσμα να απαιτούνται περισσότεροι υπολογιστικοί πόροι από αυτούς που κατείχαν. Έτσι λοιπόν ορίζουν το συγκεκριμένο εμπόδιο στο χώρο των υπολογιστών ως "*Το πρόβλημα των Big Data*".

- Τον Οκτώβριο του 1998 ο K.G. Coffman και ο Andrew Odlyzko παρατηρούν, στο άρθρο τους "*The Size and Growth Rate of the Internet*" (Coffman & Odlyzko, 1998), ότι ο ετήσιος ρυθμός αύξησης της κίνησης του διαδικτύου (*Internet traffic*) ανέρχεται στο 100%. Έτσι προσθέτουν ότι, αν η συγκεκριμένη τάση παραμείνει στα ίδια επίπεδα, η διακίνηση δεδομένων μέσω του διαδικτύου θα ξεπεράσει, μέχρι το 2002, τη διακίνηση πληροφοριών μέσω φωνητικών συνδιαλέξεων (π.χ. επικοινωνία μέσω τηλεφώνου), με αποτέλεσμα το Internet να κυριαρχήσει πλήρως στο χώρο της επικοινωνίας.
- Τον Οκτώβριο του 2000 οι καθηγητές του Berkeley, Peter Lyman και Hal R. Varian, δημοσιεύουν το άρθρο "*How much information?*" (Lyman & Varian, 2000), στο εμφανίζεται η πρώτη ολοκληρωμένη μελέτη που περιλαμβάνει ακριβή υπολογισμό και ποσοτικοποίηση σε bytes των νέων και πρωτότυπων πληροφοριών (χωρίς δηλαδή να συνυπολογίζεται η αντιγραφή και η αναπαραγωγή των συγκεκριμένων δεδομένων από τρίτους) που παράγονται σε ολόκληρο τον κόσμο, μέσα σε ένα χρόνο. Πιο συγκεκριμένα στην έρευνα υπολογίζεται ότι το 1999 παράχθηκε παγκοσμίως συνολικός όγκος πληροφοριών που ανέρχεται στα 1.5 exabytes (το ένα exabyte είναι ίσο με 10^{18} bytes, ή ισοδύναμα ένα δισεκατομμύριο gigabytes) νέων πληροφοριών, ο οποίος ισοδυναμεί με 250 megabytes ανά άτομο. Παράλληλα σύμφωνα με τα στοιχεία της μελέτης η μεγαλύτερη πλειοψηφία των δεδομένων δημιουργήθηκε από μεμονωμένα άτομα (και όχι για παράδειγμα από οργανισμούς, εταιρείες, σχολεία, πανεπιστήμια κ.ο.κ.), παρατήρηση η οποία καθιερώθηκε ως ο "*Εκδημοκρατισμός της Πληροφορίας*". Επίσης φάνηκε ότι τα ψηφιακά δεδομένα όχι απλά κυριαρχούν σε όγκο, συγκριτικά με τις υπόλοιπες μορφές πληροφοριών, αλλά παράλληλα αναπτύσσονται και με τους υψηλότερους ρυθμούς, παρατήρηση η οποία καθιερώθηκε ως η "*Κυριαρχία των Ψηφιακών*". Η συγκεκριμένη μελέτη επαναλήφθηκε τρία

χρόνια αργότερα και στην έκβαση της παρατηρήθηκε ότι το 2002 ο συνολικός όγκος νέων δεδομένων σκαρφάλωσε στα ύψη των 5 exabytes, γεγονός που αποδεικνύει ότι παρουσιάστηκε μια αλματώδης αύξηση συγκριτικά με το 1999. Αξιοσημείωτο είναι επίσης το γεγονός ότι το 92% εκ των παραγόμενων νέων πληροφοριών του 2002 αποτελούν ψηφιακό υλικό αποθηκευμένο σε σκληρούς δίσκους, γεγονός που αποδεικνύει τη κυριαρχία των ψηφιακών στις μέρες μας.

- Τον Φεβρουάριο του 2001 συνδέονται με τα Big Data, για πρώτη φορά στην ιστορία, οι λέξεις Volume, Velocity και Variety, στη δημοσίευση "*3D Data Management: Controlling Data Volume, Velocity and Variety*" του αναλυτή Doug Laney (Laney, 2001). Οι συγκεκριμένες λέξεις οι οποίες όπως έχει προαναφερθεί αναφέρονται αντίστοιχα στον μεγάλο όγκο, στη μεγάλη ταχύτητα διάδοσης και στη μεγάλη ποικιλία τύπων δεδομένων, θα γίνουν μία δεκαετία αργότερα γνωστές ως τα "*τρία V των Big Data*" και θα αποτελέσουν τις συνιστώσες, πάνω στις οποίες αυτά βασίζονται.
- Τον Μάρτιο του 2007 οι John F. Gantz και David Reinsel σε συνεργασία και με άλλους ερευνητές του IDC (*International Data Corporation*) στη δημοσίευση τους "*The Expanding Digital Universe: A Forecast of Worldwide Information through 2010*" (Gantz, et al., 2007) επιχειρούν να κάνουν την πρώτη πρόβλεψη σχετικά με τον όγκο των πληροφοριών (πρωτότυπων καθώς και αναπαραχθέντων) των επόμενων ετών. Πιο συγκεκριμένα στην έρευνα αναφέρεται ότι το 2006 παρήχθησαν ή αναπαρήχθησαν 161 exabytes δεδομένων, με αποτέλεσμα να υπολογίζεται ότι ο αντίστοιχος αριθμός του 2010 θα έχει εξαπλασιαστεί και θα ανέρχεται στα 988 exabytes. Κάτι τέτοιο συνεπάγεται το γεγονός ότι ο ετήσιος όγκος των πληροφοριών, που προστίθεται στο ψηφιακό σύμπαν, θα διπλασιάζεται κάθε 18 μήνες. Τέσσερα χρόνια αργότερα, σε νέα δημοσίευση της συγκριμένης έρευνας, πραγματοποιήθηκε η ανακοίνωση των τελικών αριθμών του 2010 και του 2011, η οποία όχι απλά επιβεβαίωσε τις προβλέψεις, αλλά τις ξεπέρασε παράλληλα, καθώς διαπιστώθηκε ότι το 2010 δημιουργήθηκαν 1200 exabytes δεδομένων, ενώ το 2011 ο αριθμός αυτός αυξήθηκε ακόμα περισσότερο, στα 1800 exabytes.
- Τον Ιανουάριο του 2008 οι οικονομολόγοι Bret Swanson και George Gilder δημοσιεύουν τη μελέτη "*Estimating the Exaflood*" (Swanson & Gilder, 2008), στην οποία προβλέπουν ότι το 2015, μόνο στις ΗΠΑ, η διακίνηση δεδομένων

μέσω του διαδικτύου θα ανέρχεται στο ένα zettabyte (το ένα zettabyte είναι ίσο με 10^{21} bytes, ή ισοδύναμα ένα δισεκατομμύριο terabytes), ένας αριθμός ο οποίος είναι πενήντα φορές μεγαλύτερος από τον αντίστοιχο του 2006.

- Τον Ιούνιο του 2008 η εταιρεία Cisco δημοσιεύει το άρθρο "Cisco Visual Networking Index – Forecast and Methodology, 2007-2012" (Cisco, 2008) προβλέπει ότι το 2012 η διακίνηση δεδομένων μέσω του διαδικτύου θα αγγίξει το μισό zettabyte, κάτι που συνεπάγεται ότι θα διπλασιάζεται κάθε δύο χρόνια. Στις 30 Μαΐου του 2012 η Cisco εξέδωσε αναφορά, σχετική με τη μελέτη της, στην οποία ανέφερε ότι οι προβλέψεις που έκανε επιβεβαιώθηκαν, γεγονός που την οδήγησε στο συμπέρασμα ότι η συναλλαγή πληροφοριών μέσω του διαδικτύου οκταπλασιάστηκε, μέσα σε μόλις μία πενταετία.
- Τον Δεκέμβριο του 2009 ο Roger E. Bohn σε συνεργασία με τον James E. Short δημοσιεύουν τη μελέτη "*How much Information? 2009 Report on American Consumers*" (Bohn & Short, 2009), στην οποία διαπιστώθηκε ότι κατά τη διάρκεια του 2008 συνολικά οι Αμερικανοί πολίτες "κατανάλωσαν" περίπου 3,6 zettabytes ψηφιακών δεδομένων καθώς και επίσης 11 τετράκις εκατομμύρια λέξεις[a]. Οι αριθμοί αυτοί ισοδυναμούν με 100.500 λέξεις και 34 gigabytes δεδομένων ανά μέσο άτομο, σε μία μέση ημέρα της καθημερινότητας του.
- Σύμφωνα με το άρθρο του Kenneth Cukier, "*Data, data everywhere*" (Cukier, 2010), στο βρετανικό περιοδικό *Economist* τον Φεβρουάριο του 2010, ο κόσμος στις μέρες μας κατέχει έναν αφάνταστα μεγάλο όγκο ψηφιακών πληροφοριών, ο οποίος τείνει ολοένα και περισσότερο στο να πάρει αχανείς διαστάσεις. Η επίδραση των πληροφοριών αυτών γίνεται αισθητή σε κάθε πτυχή της κοινωνίας μας, από την επιστήμη στις επιχειρήσεις και από τις τέχνες στις κυβερνήσεις. Έτσι, σύμφωνα με το άρθρο, οι επιστήμονες έχουν δημιουργήσει έναν νέο όρο για το συγκεκριμένο φαινόμενο, ο οποίος δεν είναι άλλος από τον όρο "*Big Data*".
- Τον Φεβρουάριο του 2011 δημοσιεύεται στο περιοδικό *Science*, από τους Martin Hilbert και Priscila Lopez, το άρθρο "*The World's Technological Capacity to Store, Communicate, and Compute Information*" (Hilbert & Lopez, 2011), στο οποίο υπολογίστηκε ότι το πλήθος των αποθηκευμένων δεδομένων, σε

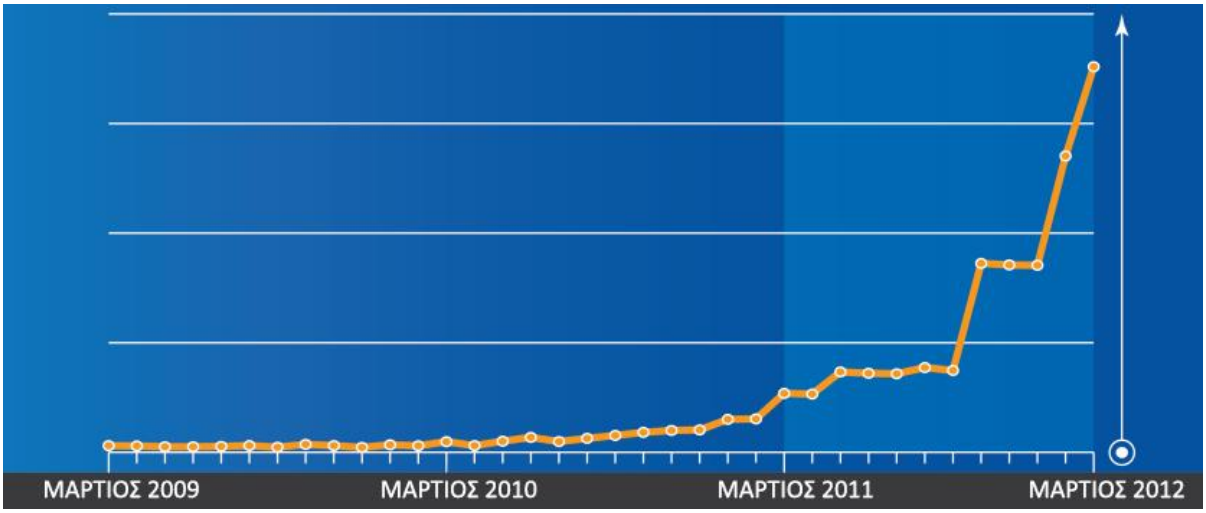
[a] Συγκεκριμένα: $1.0845 \cdot 10^{16}$ λέξεις.

ολόκληρο τον κόσμο, αυξήθηκε με μέσο ετήσιο ρυθμό 25% από το 1986 έως το 2007. Παράλληλα εκτιμήθηκε ότι το 1986, από τον συνολικό όγκο των μέσων αποθήκευσης, το 99,2% βρισκόταν σε αναλογική μορφή (για παράδειγμα: δίσκοι βινυλίου, μαγνητικές ταινίες, βιντεοκασέτες, κ.ο.κ.), ενώ αντίθετα το 2007 το 94% των μέσων αποθήκευσης είχαν ψηφιακή μορφή (CD, DVD, σκληροί δίσκοι, Mini Discs, κ.ο.κ.). Το γεγονός αυτό αποδεικνύει την αντιστροφή των ρόλων, όσον αφορά τα αναλογικά και τα ψηφιακά μέσα αποθήκευσης, ενώ τέλος σημειώνεται ότι για πρώτη φορά στην ιστορία τα ψηφιακά μέσα ξεπέρασαν σε όγκο τα αναλογικά κατά τη διάρκεια του 2002.

- Τον Απρίλιο του 2012, στη δημοσίευση "*Tracking the flow of information into home*" των W. Russell Neuman, Young Jin Park και Elliot Panek, μελετήθηκε (Neuman, Park, & Panek, 2012) η σημαντικότατη αύξηση της ημερήσιας προσφοράς δεδομένων, μέσω των ΜΜΕ, κατά την τελευταία πεντηκονταετία. Πιο συγκεκριμένα υπολογίστηκε ότι το 2005 σε ημερήσια βάση η συνολική προσφορά πληροφοριών, μέσω των ΜΜΕ, σε έναν Αμερικανό πολίτη ανέρχεται στα 900.000 λεπτά, από εκεί που το 1960 ο αντίστοιχος αριθμός ήταν ίσος με 50.000 λεπτά. Αξιοσημείωτο είναι επίσης το γεγονός ότι η συγκεκριμένη μελέτη ακολούθησε όμοια μεθοδολογία με αυτή που χρησιμοποιήθηκε το 1975 από το Ιαπωνικό υπουργείο, καθώς και το 1983 από τον κοινωνιολόγο Ithiel de Sola Pool.
- Μέσα στο 2012 δημοσιεύεται η μελέτη των David Bounie και Laurent Gille, με όνομα "*International Production and Dissemination of Information*" (Bounie & Gille, 2012), στην οποία με τρόπο όμοιο με αυτόν που χρησιμοποίησαν το 2000 οι καθηγητές του Berkeley, Peter Lyman και Hal R. Varian, υπολογίστηκε ότι μέσα στο 2008 παράχθηκε παγκοσμίως συνολικός όγκος νέων και πρωτότυπων δεδομένων που ανήλθε στα 14,7 exabytes. Ο αριθμός αυτός είναι τριπλάσιος σε σχέση με τον αντίστοιχο του έτους 2003.

2.4 Η εκτόξευση του ενδιαφέροντος γύρω από τα Big Data

Όπως εύκολα παρατηρείται από την ιστορική τους αναδρομή, τα Big Data προσελκύουν όλο και περισσότερο το ενδιαφέρον όχι μόνο των επιστημόνων και της έρευνας, αλλά παράλληλα και άλλων φορέων, με χαρακτηριστικά παραδείγματα αυτά της πολιτικής και των επιχειρήσεων. Αυτό συμβαίνει διότι δια μέσω αυτών επιτυγχάνεται η πρόσβαση σε μία αστείρευτη πηγή πληροφοριών, η οποία αν μελετηθεί και αξιοποιηθεί κατάλληλα είναι ικανή να αποδώσει τεράστια οφέλη. Έτσι λοιπόν παρατηρήθηκε ότι στις αρχές του 2011 τα Big Data και η αξιοποίηση αυτών άρχισαν να συγκεντρώνουν καθολικό ενδιαφέρον από κάθε πτυχή της κοινωνίας και όχι κυρίως από επιστήμονες που σχετίζονταν με το χώρο της πληροφορικής. Προκειμένου αυτό να γίνει αντιληπτό παρατίθεται το παρακάτω διάγραμμα (Visible Intelligence, 2012) στο οποίο φαίνεται η μεταβολή του πλήθους των αναφορών, στα μέσα κοινωνικής δικτύωσης, που σχετίζονταν με τα Big Data, κατά τη διάρκεια 36 μηνών (Μάρτιος 2009 έως Μάρτιος 2012).

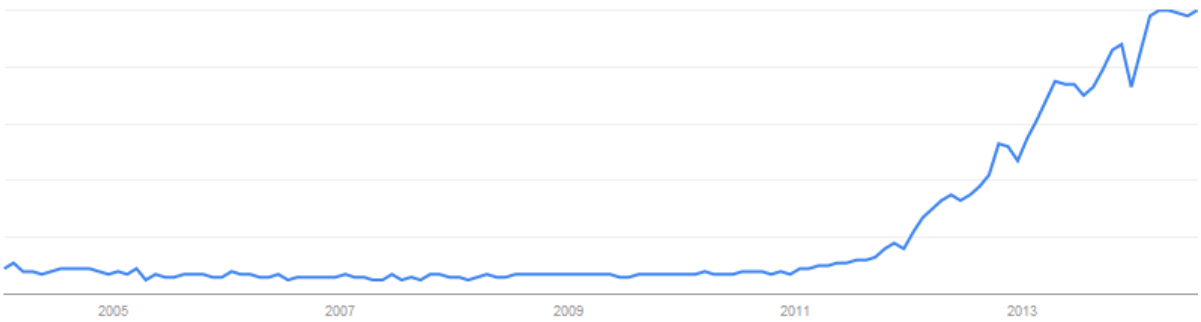


Διάγραμμα 2.1 - Αναφορές στα Big Data (Συλλεγμένες από όλα τα Social Media) Μάρτιος 2009 – Μάρτιος 2012

Όπως εύκολα αντιλαμβάνεται κάποιος οι αναφορές των Big Data στα Social Media αυξάνονται δραματικά από το 2011 και μετά. Πιο συγκεκριμένα η αύξηση των αναφορών σε αυτά, από τον Μάρτιο του 2011 μέχρι τον Μάρτιο του 2012 (σκιαγραφημένη περιοχή στο διάγραμμα), υπολογίστηκε ότι ξεπερνά το 1200% σε σχέση με τα προηγούμενα έτη. Τέλος αξιοσημείωτο είναι το γεγονός ότι οι συγκεκριμένες αναφορές αφορούν μόνο την αγγλική γλώσσα, γεγονός που συνεπάγεται ότι η μεταβολή αυτή θα είναι ακόμα μεγαλύτερη, αν συμπεριληφθούν στη μελέτη και οι υπόλοιπες γλώσσες του κόσμου.

Με απολύτως όμοιο τρόπο αποτυπώνεται η εξάπλωση των Big Data και στα νούμερα που αφορούν τις αναζητήσεις του θέματος στο διαδίκτυο μέσω της δημοφιλέστερους μηχανής αναζήτησης που δεν είναι άλλη από αύτη της Google.

Προκειμένου να διαπιστωθεί η παρατήρηση αυτή χρησιμοποιήθηκε η υπηρεσία Google Trends, η οποία εμφανίζει, συναρτήσει του χρόνου, τη σχετική συχνότητα των αναζητήσεων μίας λέξης, ανά το πλήθος των συνολικών παγκόσμιων αναζητήσεων. Το συγκεκριμένο εργαλείο της Google εμφανίζει αποτελέσματα από το 2004 μέχρι και σήμερα, ανανεώνοντας τα δεδομένα του ημερησίως και ανά μία ώρα στα δημοφιλή θέματα. Όπως γίνεται εύκολα αντιληπτό, το Google Trends αποτελεί ισχυρό όπλο στα χέρια των τμημάτων marketing των επιχειρήσεων, προκειμένου αυτά να παρακολουθήσουν τη δημοτικότητα των προϊόντων τους στο καταναλωτικό κοινό, ενώ παράλληλα αξίζει να τονιστεί ότι έχει χρησιμοποιηθεί και σε λιγότερο προφανείς έρευνες όπως για παράδειγμα η παρακολούθηση της εξάπλωσης της γρίπης σε έναν πληθυσμό. Στο παρακάτω διάγραμμα ο οριζόντιος άξονας αναπαριστά τον χρόνο (από τον Ιανουάριο του 2004 μέχρι τον Ιανουάριο του 2014), ενώ ο κάθετος το πλήθος των αναζητήσεων της φράσης “Big Data”, σε σύγκριση με το συνολικό πλήθος αναζητήσεων που πραγματοποιήθηκαν, σε ολόκληρο τον κόσμο.



Διάγραμμα 2.2 - Παγκόσμιες αναζητήσεις των “Big Data” στο Google (Ιανουάριος 2004 – Ιούλιος 2014)

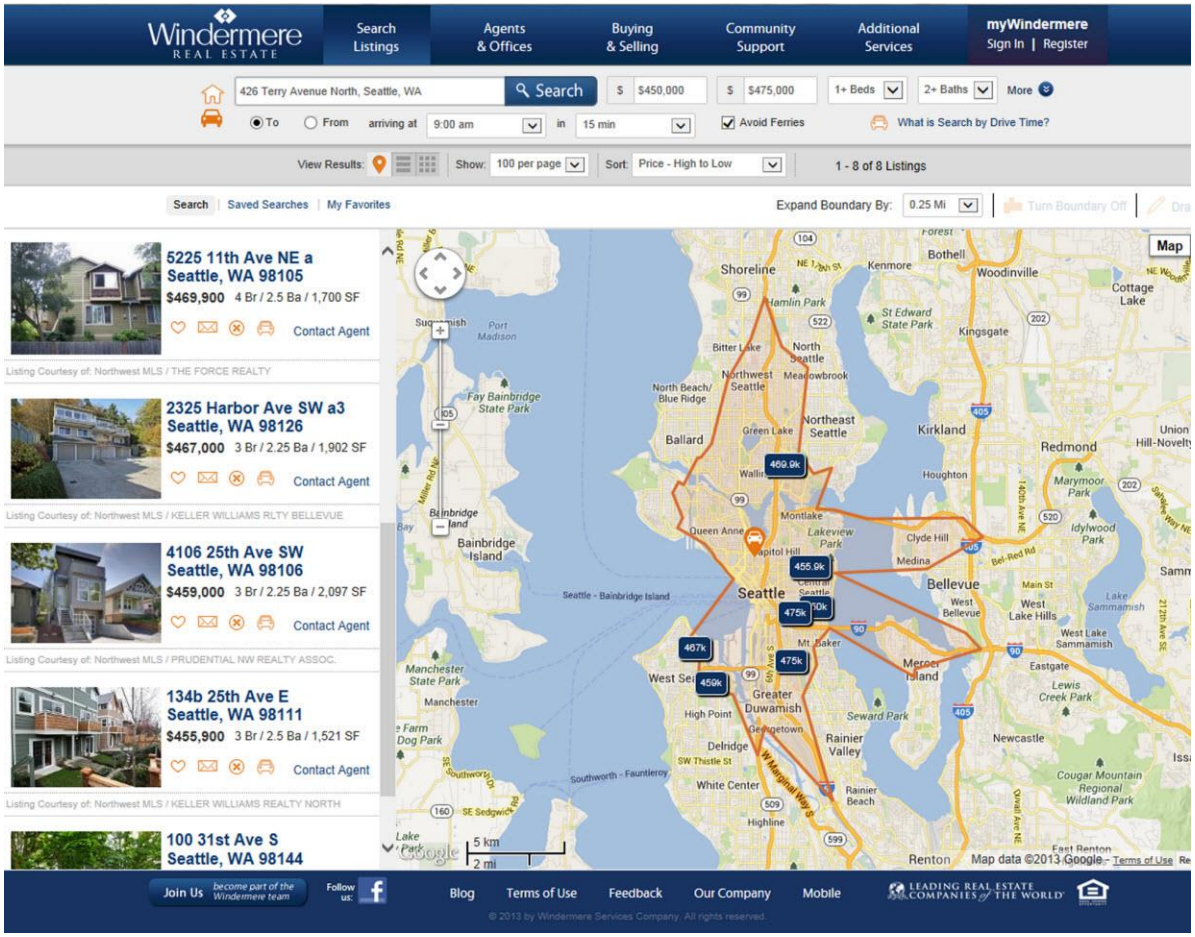
Όπως αναμενόταν η παραπάνω γραφική παράσταση παρουσιάζει την ίδια ακριβώς μορφή, από τον Μάρτιο του 2009 μέχρι τον Μάρτιο του 2012, με αυτή που αφορούσε τις αναφορές των Big Data στα μέσα κοινωνικής δικτύωσης και μάλιστα κατά τα επόμενα δύο σχεδόν χρόνια παρατηρείται εξίσου ραγδαία αύξηση του ενδιαφέροντος γύρω από το συγκεκριμένο θέμα. Πιο συγκεκριμένα αναφέρεται ότι το ποσοστό αύξησης των αναζητήσεων στο Google, από τον Ιανουάριο του 2010 έως τον Ιούλιο του 2014, γύρω από τα Big Data, υπερβαίνει το 1300%.

2.5 Η αξιοποίηση των Big Data

Έχουν ήδη αναφερθεί κάποιες περιπτώσεις αξιοποίησης των Big Data, όπως αυτή της απεικόνισης ενός μέρους του ουρανού μέσω των εικόνων που συλλέγει το τηλεσκόπιο του προγράμματος SDSS. Παρακάτω παρατίθενται μερικά ακόμα παραδείγματα (Acquia, 2014) συλλογής, ανάλυσης και αξιοποίησης των Big Data:

- Το τμήμα marketing των εταιριών που ασχολούνται με το εμπόριο παρακολουθούν μέσω των μηνυμάτων του Facebook και του Twitter τις αντιδράσεις του καταναλωτικού κοινού γύρω από τα προϊόντα τους. Παράλληλα αναλύουν τις τάσεις και τις επιθυμίες των καταναλωτών προκειμένου να τις καλύψουν στο μέλλον.
- Η εταιρεία Ford (Hemsoth, 2013) δείχνει έμπρακτα ότι έχει πίστη στους αισθητήρες αυτοκινήτων και έτσι το νέο υβριδικό της αυτοκίνητο με το όνομα *Fusion* είναι ικανό να συλλέγει μέχρι και 25 GB πληροφοριών την ώρα μέσω των 74 αισθητήρων που διαθέτει, οι οποίοι καταγράφουν ασταμάτητα τις κινήσεις της περιμέτρου γύρω από το αυτοκίνητο, προκείμενου να διευκολυνθεί η οδηγική εμπειρία. Πιο συγκεκριμένα, τα δεδομένα που συλλέγονται από το αυτοκίνητο, χρησιμοποιούνται για λόγους ασφάλειας (καταγραφή και κατανόηση της οδικής συμπεριφοράς του οδηγού προκειμένου να αποφευχθούν πιθανά ατυχήματα, καθώς και αποφυγή εμποδίων και συγκρούσεων) αλλά και οικονομίας (οικονομία καυσίμων, καθώς και αξιολόγηση των μηχανικών μερών του αυτοκινήτου, έτσι ώστε οποιαδήποτε πιθανή φθορά να γίνει άμεσα αντιληπτή από τον ιδιοκτήτη, προκειμένου να ελαχιστοποιηθεί το κόστος συντήρησης και να αποφευχθούν περαιτέρω ζημίες). Τέλος, αξίζει να αναφερθεί ότι η Ford (Πηλίδης, 2012) αυτή τη περίοδο έχει ρίξει μεγάλη βαρύτητα στην έρευνα αισθητήρων πραγματικού χρόνου (*real time*), οι οποίοι μέσω διαφόρων καμερών και ραντάρ θα συλλέγουν πληροφορίες από το περιβάλλον εκτός του αυτοκινήτου και βοηθούν τον οδηγό να μένει συγκεντρωμένος στην οδήγηση, χωρίς να αποσπάται η προσοχή του από εξωτερικούς παράγοντες. Επίσης ιδιαίτερη έμφαση (McCue, 2013) δίνει και στους βιομετρικούς αισθητήρες, οι οποίοι υπολογίζουν τα επίπεδα άγχους του οδηγού. Οι συγκεκριμένοι αισθητήρες τοποθετούνται στο κάθισμα, έτσι ώστε λίγα δευτερόλεπτα πριν από μία πιθανή σύγκρουση, όπου τα επίπεδα του στρες ανεβαίνουν ενώ παράλληλα τα αντανακλαστικά και οι ικανότητες του οδηγού μειώνονται, να ενημερωθεί το σύστημα ασφαλείας του αυτοκινήτου, και να ομαλοποιήσει την οδήγηση με σκοπό την αποφυγή ενός ατυχήματος.

- Η Windermere (Frank, 2013) (μία από τις μεγαλύτερες Real Estate εταιρείες των ΗΠΑ) στις 9 Αυγούστου του 2013 ανακοινώνει ότι οι πελάτες της πλέον θα μπορούν να μαθαίνουν πόση ώρα θα χρειάζονται προκειμένου να πραγματοποιήσουν τα συνηθισμένα τους δρομολόγια, από και προς το σπίτι που πρόκειται να αγοράσουν. Έτσι η εταιρεία θα μπορεί να βοηθά τον υποψήφιο αγοραστή να επιλέξει το κατάλληλο σπίτι, γνωστοποιώντας του για παράδειγμα πόση ώρα θα απαιτείται προκειμένου να πάει το πρωί στη δουλεία του και να γυρίσει το απόγευμα από αυτή. Αυτό επιτυγχάνεται μέσω ενός αλγορίθμου ο οποίος λαμβάνει και αναλύει ανώνυμα τα σήματα GPS από 100 εκατομμύρια οδηγούς κατά τη διάρκεια όλης της ημέρας.



Εικόνα 2.3 - Παράδειγμα χρήσης του αλγορίθμου της Windermere

- Η εταιρεία Union Pacific Railroad (Murphy, 2012) έχει τοποθετήσει κατά μήκος των γραμμών ειδικά θερμόμετρα, μικρόφωνα και δέκτες υπερήχων, που συλλέγουν δεδομένα για κάθε τρένο που μετακινείται. Τα δεδομένα αυτά αναλύονται με σκοπό να προβλεφθεί έγκαιρα οποιαδήποτε βλάβη στα μηχανικά μέρη των τρένων, με σκοπό την ασφάλεια των επιβατών και την

οικονομία στα κόστη συντήρησης λόγω της μειωμένης πιθανότητας εμφάνισης μεγάλων βλαβών.

- Οι μεγάλες ναυτιλιακές εταιρείες χρησιμοποιούν τηλεματικές εφαρμογές (όλες εκείνες οι υπηρεσίες που μας προσφέρει η σύγχρονη τεχνολογία, μέσω των οποίων μπορούμε να αποστείλουμε και να λάβουμε κάθε φύσης πληροφορία), μέσω των οποίων τα πλοία αποστέλλουν πληροφορίες σχετικά με την θέση τους και τη πορεία τους, προκειμένου να ελαχιστοποιηθεί το κόστος της μετακίνησης του στόλου. Έτσι, αναλύοντας αυτά τα δεδομένα για χιλιάδες πλοία και δρομολόγια, καθίσταται δυνατή η εύρεση της αποδοτικότερης λύσης για τη μεταφορά των πλοίων σε χρόνο και καύσιμα. Σύμφωνα με την General Electric Company η μείωση των εξόδων στον τομέα της ναυτιλίας, μέσω της αξιοποίησης των Big Data, θα συνεισφέρει 30 τρισεκατομμύρια δολάρια στον παγκόσμιο ΑΕΠ.

Κεφάλαιο 3. Μεθοδολογίες

3.1 Εξόρυξη δεδομένων (Data mining)

Το 1982 ο συγγραφέας John Naisbitt (Naisbitt, 1982) ανέφερε στο βιβλίο του "*Megatrends*" ότι "Πνιγόμαστε από πληροφορίες, αλλά διψάμε για γνώση" ("*we are drowning in information, but starving for knowledge*"). Περίπου τριάντα χρόνια αργότερα, δεν είναι δύσκολο να αντιληφθεί κάποιος ότι ζούμε στην "Εποχή της Πληροφορίας", καθώς ένα κύριο γνώρισμα των ημερών μας είναι η αλματώδης αύξηση των παραγόμενων δεδομένων. Έτσι λοιπόν έχουμε μία συνεχή εισροή πληροφοριών, στις ήδη ογκώδεις βάσεις δεδομένων, οι οποίες μάλιστα βρίσκονται σε διαφορετικές μορφές, όπως για παράδειγμα κείμενο, φωτογραφία, ήχος και video. Το συμπέρασμα που μπορούμε να βγάλουμε είναι ότι η αναζήτηση και η επεξεργασία των δεδομένων, με σκοπό την εξαγωγή χρήσιμης γνώσης από αυτά, καθίσταται μία περίπλοκη και χρονοβόρα διαδικασία, λόγω της υπερφόρτωσης και της μεταβλητής μορφής των δεδομένων. Η άμεση συνέπεια λοιπόν είναι η αναζήτηση αυτοματοποιημένων τεχνικών, που στοχεύουν στην εξόρυξη πληροφοριών μέσα από τα τεράστια σε όγκο δεδομένα.

Καθώς, κατά τη διάρκεια των ετών, οι συλλογές δεδομένων αυξήθηκαν τόσο σε όγκο όσο και σε πολυπλοκότητα, η αυτόματη επεξεργασία δεδομένων έχει αντικαταστήσει τη χειρωνακτική ανάλυση τους. Πιο συγκεκριμένα, η εξόρυξη δεδομένων (*data mining*) ορίζεται (Larose, 2004) ως η διαδικασία εξαγωγής χρήσιμης γνώσης, υπό τη μορφή συσχετίσεων, προτύπων και τάσεων, μέσω της εξέτασης, της ανάλυσης και της επεξεργασίας βάσεων δεδομένων, με χρήση αλγορίθμων, συνδυάζοντας αρχές της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης[a], των συστημάτων βάσεων δεδομένων και της οπτικοποίησης. Υπογραμμίζεται ότι η εξόρυξη δεδομένων δεν περιορίζεται σε ένα μόνο τύπο δεδομένων, ωστόσο όμως οι αλγόριθμοί της μπορεί να διαφέρουν εφαρμοσμένοι σε διαφορετικά είδη δεδομένων.

Μία από τις τεχνικές, που περιλαμβάνει η εξόρυξη δεδομένων, είναι η ανάλυση παλινδρόμησης. Η συγκεκριμένη τεχνική θα αναλυθεί στην Παράγραφο 3.4 και θα εφαρμοστεί στο 6ο Κεφάλαιο.

---

[a] Ο κλάδος της τεχνητής νοημοσύνης που ασχολείται με τη κατασκευή αλγορίθμων, οι οποίοι επιτρέπουν στους ηλεκτρονικούς υπολογιστές να "μαθαίνουν".

3.2 Εξόρυξη κειμένου (Text mining)

Τα τελευταία χρόνια παρατηρείται μία αλματώδης αύξηση του όγκου των εγγράφων κειμένου που δημοσιεύεται καθημερινά στο διαδίκτυο. Το γεγονός αυτό, σε συνδυασμό με την αναζήτηση των επιθυμητών πληροφοριών στα συγκεκριμένα κείμενα, αποτέλεσε ιδανικό κίνητρο για την έρευνα και την ανάπτυξη τεχνικών εξόρυξης γνώσης από δεδομένα. Έτσι λοιπόν αναπτύχθηκε ο τομέας της εξόρυξης κειμένου (*text mining*), η οποία αποτελεί ένα παρακλάδι της εξόρυξης δεδομένων (*data mining*), εστιάζοντας (Hearst, 2003) στην εξαγωγή χρήσιμων πληροφοριών από κείμενα φυσικής γλώσσας και όχι από δομημένες βάσεις δεδομένων. Η διαφοροποίηση αυτή έγκειται στο γεγονός ότι οι βάσεις δεδομένων σχεδιάζονται από προγράμματα, τα οποία τις επεξεργάζονται αυτόματα. Αντίθετα το κείμενο γράφεται προκειμένου να διαβαστεί από ανθρώπους, γεγονός που καθιστά την εξόρυξη πληροφοριών από αυτό μία περίπλοκη διαδικασία. Η πλήρης κατανόηση του κειμένου φυσικής γλώσσας αποτελεί έναν από τους πρωταρχικούς στόχους της πληροφορικής. Πολλοί ερευνητές θεωρούν (Hearst, 2003) ότι, προκειμένου να επιτευχθεί κάτι τέτοιο και να είμαστε σε θέση να γράψουμε προγράμματα που κατανοούν πλήρως ένα κείμενο, απαιτείται η πλήρης προσομοίωση της λειτουργίας του ανθρώπινου μυαλού. Ωστόσο ο τομέας της υπολογιστικής γλωσσολογίας (*Natural Language Processing*) σημειώνει σημαντική πρόοδο σε μικρές εργασίες που αφορούν την ανάλυση κειμένου. Για παράδειγμα υπάρχουν προγράμματα τα οποία δέχονται σαν είσοδο ένα κείμενο, όπως για παράδειγμα ένα άρθρο ή ένα βιβλίο, και αποσπούν τις λέξεις και φράσεις κλειδιά που το συνοψίζουν με το βέλτιστο τρόπο.

Η εξόρυξη κειμένου δεν πρέπει να συγχέεται με το γνωστό σε όλους μας *web search*. Πιο συγκεκριμένα κατά την αναζήτηση ο χρήστης ψάχνει μία πληροφορία που είναι ήδη γνωστή και έχει γραφεί από κάποιον άλλον. Αντίθετα η εξόρυξη κειμένου ορίζεται (Hearst, 2003) ως η διαδικασία εξαγωγής νέας, έγκυρης και χρήσιμης γνώσης, με αυτόματο τρόπο, από τους διάφορους γραπτούς πόρους.

Στο Παράρτημα Α, καθώς και στα επόμενα Κεφάλαια εφαρμόζονται τα παραπάνω εργαλεία, προκειμένου να αντλήσουμε χρήσιμη γνώση από κείμενο, καθώς και από δεδομένα γενικότερα. Στη συνέχεια θα περιγραφούν οι μεθοδολογίες που θα εφαρμοστούν στα επόμενα κεφάλαια της συγκεκριμένης διπλωματικής εργασίας.

3.3 Μεθοδολογία 1

Η συγκεκριμένη μεθοδολογία θα εφαρμοστεί στο **Κεφάλαιο 4**, προκειμένου να μελετηθεί και να απαντηθεί το ερώτημα, που σχετίζεται με το αν αποτυπώνεται ή όχι στο Twitter η αρνητική ψυχολογία των Ελλήνων, που οφείλεται στην οικονομική κρίση των τελευταίων ετών. Έτσι λοιπόν συλλέχθηκαν και αναλύθηκαν 3.958.648 Tweets, που αναρτήθηκαν μέσα σε διάστημα τεσσεράμισι ετών (από την 1η Ιανουαρίου 2009, έως και την 26η Ιουνίου 2013). Προκειμένου να πραγματοποιηθεί η ανάλυση των Tweets και η εξαγωγή συμπερασμάτων μέσω αυτών, γράφτηκε κώδικας σε γλώσσα προγραμματισμού Java (Βλ. Παράρτημα Β). Τα συγκεκριμένα μηνύματα, τα οποία χωρίζονται σε τέσσερις κατηγορίες ανάλογα με το περιεχόμενο τους, αποθηκεύτηκαν σε βάσεις δεδομένων έτσι ώστε να αναλυθούν. Πιο συγκεκριμένα τα Tweets διαχωρίστηκαν στις εξής κατηγορίες:

- Μηνύματα που εκφράζουν **ευτυχία**.
- Μηνύματα που εκφράζουν **απογοήτευση**.
- Μηνύματα που εκφράζουν **αγανάκτηση**.
- **Ουδέτερα** μηνύματα.

Κάθε μία από τις παραπάνω ομάδες περιλαμβάνει κάποιο αριθμό από οικογένειες λέξεων, οι οποίες με τη σειρά τους περιλαμβάνουν συνώνυμα και παραπλήσιες λέξεις, με την αντίστοιχη κατηγορία στην οποία ανήκουν. Με τη λέξη οικογένεια λέξεων εννοούμε τη συνώνυμη λέξη, μαζί με τους διάφορους λεκτικούς τύπους της. Για παράδειγμα μία οικογένεια λέξεων που αναζητήθηκε και συμπεριλήφθηκε στα μηνύματα που εκφράζουν ευτυχία αποτελείται από τη λέξη χαρά, μαζί με κάποιο αριθμό λέξεων που είναι πηγάζουν από την ίδια ρίζα με αυτή. Έτσι η συγκεκριμένη οικογένεια περιλαμβάνει τα ουσιαστικά *χαρά* και *χαρές*, το ρήμα *χαίρομαι*, μαζί με διάφορους τύπους (πρόσωπα, χρόνους, εγκλίσεις) όπως *χάρηκα*, *χαιρόμαστε*, *χαρήκαμε*, *χαρώ*, *χαρούμε* και τα επίθετα *χαρούμενος*, *χαρούμενη*, *χαρούμενο*, *χαρούμενοι*, *χαρούμενες*, *χαρούμενα*, *χαρούμενου*, *χαρούμενης*, *χαρούμενων* και *χαρούμενους*. Παράλληλα συγκαταλέγονται και οι παραπάνω λέξεις χωρίς να είναι τονισμένες, καθώς όπως παρατηρήθηκε περίπου το 20% των χρηστών δεν χρησιμοποιούν τονισμό στις αναρτήσεις τους στο Twitter. Με αυτόν τον τρόπο αναζητήθηκαν 368 λέξεις, οι οποίες σχημάτισαν 15 οικογένειες λέξεων που παραπέμπουν στο συναίσθημα της ευτυχίας, 289 λέξεις οι οποίες σχημάτισαν 23 οικογένειες λέξεων που παραπέμπουν στο συναίσθημα της απογοήτευσης, 270 λέξεις οι οποίες δημιούργησαν 23 οικογένειες λέξεων που παραπέμπουν στο συναίσθημα της αγανάκτησης,

καθώς και 4 ουδέτερες λέξεις (μία οικογένεια λέξεων). Οι συγκεκριμένες 935 λέξεις αναζητήθηκαν στο Search του Twitter μέσα σε χρονικό διάστημα 1638 ημερών προκειμένου να μελετηθεί η διακύμανση της εμφάνισης τους μέσα στη διάρκεια των τεσσεράμισι χρόνων.

Έτσι λοιπόν οι τρεις πρώτες κατηγορίες μηνυμάτων μετρούν λέξεις που σχετίζονται με τρία διαφορετικά συναισθήματα. Την ευτυχία, την απογοήτευση και την αγανάκτηση. Παράλληλα σχηματίστηκε και μία τέταρτη ομάδα από Tweets, στην οποία μετρήθηκαν τέσσερις ουδέτερες λέξεις. Αυτές είναι οι λέξεις "*είμαι*", "*είναι*" καθώς και οι αντίστοιχες λέξεις χωρίς τονισμό ("*ειμαι*" και "*ειναι*"). Το πλήθος των συγκεκριμένων λέξεων διαιρέθηκε, για κάθε μία από τις 1638 ημέρες, με το συνολικό πλήθος των λέξεων που αφορούσαν κάποιο από τα τρία συναισθήματα. Η συγκεκριμένη διαδικασία πραγματοποιήθηκε για τους εξής λόγους: Αρχικά, το πλήθος των ενεργών χρηστών του Twitter, έχει αυξηθεί σε μεγάλο βαθμό μέσα στο χρονικό διάστημα των τεσσεράμισι ετών. Αν λοιπόν δεν χρησιμοποιηθούν τα ουδέτερα μηνύματα το αποτέλεσμα θα είναι όσο περνούν οι ημέρες να εμφανίζεται μια ραγδαία αύξηση κάποιου συναισθήματος, η οποία προφανώς δεν θα οφείλεται σε κοινωνικά αίτια. Ο δεύτερος λόγος έχει να κάνει με το γεγονός ότι κατά τη διάρκεια ορισμένων περιόδων οι χρήστες τείνουν να δημοσιεύουν λιγότερα, ή περισσότερα μηνύματα απ' ότι συνήθως. Για παράδειγμα κατά τη διάρκεια των καλοκαιρινών διακοπών αρκετοί χρήστες δεν έχουν πρόσβαση σε ηλεκτρονικό υπολογιστή, ή αποφεύγουν να τον χρησιμοποιούν, με αποτέλεσμα να μειώνεται η συχνότητα των μηνυμάτων. Κάτι τέτοιο θα συνεπαγόταν επομένως ότι το συναίσθημα της ευτυχίας κατά τους καλοκαιρινούς μήνες φθίνει, συμπέρασμα που προφανώς δεν ευσταθεί.

Η επιλογή των λέξεων, που συμπεριλήφθηκαν στις αναζητήσεις, πραγματοποιήθηκε με γνώμονα την πληρέστερη δυνατή κάλυψη των τριών συναισθημάτων. Χρησιμοποιήθηκαν διάφορα λεξικά προκειμένου να βρεθούν συνώνυμα των λέξεων, ενώ παράλληλα προστέθηκαν και λέξεις που, αν και δεν είναι συνώνυμες με την "*ευτυχία*", την "*απογοήτευση*", ή την "*αγανάκτηση*", παραπέμπουν σε αυτές. Εν συνεχεία για τις λέξεις αυτές αναζητήθηκαν νέα συνώνυμα, με αποτέλεσμα μέσα από τη συνεχή επανάληψη αυτής της διαδικασίας να δημιουργηθεί ένα πλέγμα λέξεων παραπλήσιου νοήματος. Έτσι δημιουργήθηκε μία επαρκής ομάδα λέξεων, για το κάθε συναίσθημα αντίστοιχα, προκειμένου να επιτευχθεί η χρονική αναπαράστασή και η ανάλυση αυτού. Τέλος προστέθηκαν οι διάφοροι λεκτικοί τύποι για κάθε μία από τις λέξεις, προκειμένου η αναζήτησή μας να ισχυροποιηθεί ακόμα περισσότερο και η κατάληξη σε συμπεράσματα κοινωνικού περιεχομένου να καταστεί ασφαλέστερη και ακριβέστερη.

Προτού προχωρήσουμε στην παρουσίαση των αποτελεσμάτων της μελέτης, οφείλουμε να υπογραμμίσουμε πρώτα τις όποιες αδυναμίες του μοντέλου. Αρχικά όπως αναφέρθηκε αναζητήθηκαν συνολικά 935 λεκτικοί τύποι. Δεν είναι δύσκολο επομένως να αντιληφθούμε ότι μέσα στις εκατομμύρια ελληνικές λέξεις θα υπάρχουν αρκετές που δεν έχουν συμπεριληφθεί στο δείγμα μας για λόγους οικονομίας χρόνου. Παρόλα αυτά αναζητήθηκαν λέξεις που εμφανίζονται με μεγάλη συχνότητα στη καθημερινότητά μας, οι οποίες είναι και αυτές που θα οδηγήσουν σε σημαντικότερες μεταβολές και ασφαλέστερα συμπεράσματα. Στη συνέχεια τονίζεται ότι δεν κρίθηκε σημαντικό να συμπεριληφθεί η ανορθογραφία, καθώς υπάρχουν δεκάδες τρόποι με τους οποίους κάποιος μπορεί να μην γράψει σωστά μια λέξη. Ωστόσο εξαιρέσεις αποτέλεσε για παράδειγμα η έλλειψη τονισμού των λέξεων, η οποία ήταν αρκετά συχνή και επομένως αδύνατη να αγνοηθεί, καθώς και κάποιες μεμονωμένες ειδικές περιπτώσεις όπως για παράδειγμα η λέξη "*στεναχώρια*", η οποία αποτελεί την ελληνιστική κοινή της λέξεως "*στενοχώρια*" και εμφανίζεται διπλάσιες φορές σε σχέση με αυτή. Παράλληλα δεν εξετάστηκαν λέξεις που είναι γραμμένες σε μορφή greeklish (ελληνική γλώσσα γραμμένη σε λατινικό αλφάβητο), καθώς μέσα από μία λέξη προκύπτουν δεκάδες, λόγω του γεγονότος ότι οι χρήστες, μέσω της γραφής σε greeklish, επιθυμούν να υπεραπλουστεύσουν τα κείμενά τους, προκειμένου να κάνουν εξοικονόμηση χρόνου. Έτσι για παράδειγμα η λέξη "*απέχθεια*", ανάλογα με τον τρόπο που έχει συνηθίσει ο καθένας να την γράψει, θα εμφανιστεί ως "*apextheia*", "*apex8eia*", "*apexthia*", "*apex8ia*", και ούτω καθεξής, λέξεις οι οποίες εμφανίζονται με εξαιρετικά χαμηλή συχνότητα και άρα μπορούν να αγνοηθούν. Τέλος δεν εξετάστηκε η πιθανότητα μέσα σε μία πρόταση να εμπεριέχεται άρνηση, με αποτέλεσμα να αλλάζει η σημασία της. Για παράδειγμα η φράση "*δεν είμαι νευριασμένος*" μετρήθηκε σαν φράση με χαρακτήρα απογοήτευσης, ενώ προφανώς κάτι τέτοιο είναι λανθασμένο. Η ελαστική αυτή αντιμετώπιση πραγματοποιήθηκε διότι η ύπαρξη μιας αρνητικής λέξης σε μια πρόταση δεν συνεπάγεται αυτομάτως την αντιστροφή του χαρακτήρα της, καθώς θα μπορούσαμε να πούμε "*είμαι νευριασμένος, διότι δεν έχω λεφτά*", φράση η οποία παρόλη την ύπαρξη της λέξης "*δεν*", διατηρεί τον χαρακτήρα αγανάκτησης, που τη διακρίνει. Προκειμένου επομένως να μην καταφύγουμε σε μία παραδοχή, η οποία μπορεί ενδεχομένως να οδηγήσει σε άλλα χειρότερα σφάλματα, επιλέξαμε η λέξη αύτη καθαυτή να προσδιορίζει το ύφος ενός δεδομένου tweet, με μία μικρή πιθανότητα σφάλματος, το οποίο ωστόσο θα είναι καθολικό για όλα τα συναισθήματα που εξετάζουμε. Προκειμένου να διορθωθεί το πιθανό σφάλμα από την εμφάνιση άρνησης σε κάποιες προτάσεις, απαιτούνται πολύπλοκες

τεχνικές text mining, με τις οποίες δεν θα ασχοληθούμε στη παρούσα διπλωματική εργασία.

Στο 4ο Κεφάλαιο θα παρουσιαστούν τρεις χρονοσειρές, που περιέχουν την καθημερινή αναπαράσταση καθενός εκ των τριών συναισθημάτων (ευτυχία, απογοήτευση, αγανάκτηση). Κάτι τέτοιο επιτυγχάνεται αφού πρώτα υπολογιστεί το ποσοστό εμφάνισης μίας συγκεκριμένης λέξης στο Twitter, συγκριτικά με την ουδέτερη κατηγορία μηνυμάτων, για μία δεδομένη ημέρα. Τα ποσοστά αυτά των λέξεων, που αφορούν κάποιο από τα τρία συναισθήματα, προστίθενται έτσι ώστε να αποκτήσουμε μία συνολική εικόνα της εμφάνισης του κάθε συναισθήματος για την κάθε δεδομένη ημέρα. Επαναλαμβάνοντας αυτή τη διαδικασία, για κάθε μία από τις 1638 ημέρες που μελετάμε, προκύπτουν τρία χρονοδιαγράμματα, που αντιστοιχούν σε κάθε ένα από τα τρία συναισθήματα, τα οποία μας αποκαλύπτουν πώς φαίνονται μέσα από το Twitter οι αλλαγές στη διάθεση της κοινής γνώμης, κατά τη χρονική περίοδο των τεσσεράμισι ετών.

3.4 Μεθοδολογία 2

Η συγκεκριμένη μεθοδολογία θα εφαρμοστεί στο **Κεφάλαιο 6** και περιλαμβάνει στοιχεία της ανάλυσης παλινδρόμησης, τα θεωρητικά στοιχεία της οποίας θα αναφερθούν στη συνέχεια. Ο τρόπος με τον οποίο συμπεριφέρεται μία τυχαία μεταβλητή, συνήθως καθορίζεται από την συμπεριφορά άλλων τυχαίων μεταβλητών. Στη περίπτωση που έχουμε στη διάθεσή μας παρατηρήσεις αυτών των μεταβλητών, τότε μπορούμε να εστιάσουμε στον προσδιορισμό της ποσοτικής σχέσης μεταξύ των εμπλεκόμενων μεταβλητών. Ο λόγος για τον οποίο μπαίνουμε στη διαδικασία να επεξεργαστούμε με αυτόν τον τρόπο τις παρατηρήσεις μίας ή περισσοτέρων μεταβλητών, αποσκοπεί στη διερεύνηση της συμπεριφοράς και στην παραγωγή εκτιμήσεων, που αφορούν τη μελλοντική εξέλιξη, ενός φαινομένου. Έτσι λοιπόν στη συνέχεια θα χρησιμοποιήσουμε ανάλυση παλινδρόμησης, η οποία αποτελεί μία μεθοδολογία που στοχεύει στον προσδιορισμό μίας σχέσης μεταξύ των μεταβλητών που εμπλέκονται στο φαινόμενο. Με αυτόν τον τρόπο καθίσταται εφικτή η δημιουργία προβλέψεων και η κατανόηση της συμπεριφοράς του φαινομένου. Η ανάλυση παλινδρόμησης εφαρμόζεται τόσο σε διαστρωματικά δεδομένα (για παράδειγμα ο αριθμός τηλεφωνημάτων που έγιναν από τα νοικοκυριά της Αθήνας τη χθεσινή ημέρα, ή το ποσό που κατέθεσαν 100 άτομα σε μία τράπεζα) όσο και σε στοιχεία χρονολογικών σειρών (για παράδειγμα οι μηνιαίες πωλήσεις αυτοκινήτων μίας αυτοκινητοβιομηχανίας μέσα σε μία δεκαετία, ή το σύνολο των ημερήσιων καταθέσεων σε μία τράπεζα κατά τη διάρκεια μίας εβδομάδας).

3.4.1 Εκτίμηση του μοντέλου

Η ανάλυση προβλημάτων στα οποία έχουμε ενδείξεις, ή υποπτευόμαστε ότι οι τιμές μίας μεταβλητής (εξαρτημένη μεταβλητή y) επηρεάζονται από τις τιμές μίας ή περισσοτέρων μεταβλητών (ανεξάρτητες μεταβλητές x_1, x_2, … , x_k), μας οδηγούν στο να ορίσουμε το γενικό γραμμικό μοντέλο:

$$y_t = b_0 + b_1 x_{t1} + b_2 x_{t2} + \ldots + b_k x_{tk} + \varepsilon_t \quad , \quad (\textit{Σχέση 1})$$

Το μοντέλο προσαρμόζεται σε: $\hat{y}_t = \hat{b}_0 + \hat{b}_1 x_{t1} + \hat{b}_2 x_{t2} + \ldots + \hat{b}_k x_{tk}$

Το παραπάνω υπόδειγμα λέγεται εκτιμηθέν υπόδειγμα και τα ε_t είναι τα τυχαία σφάλματα, τα οποία ορίζονται (Οικονόμου & Καρώνη, 2010) ως η

κατακόρυφη απόκλιση της τιμής y_t από την ευθεία της συνάρτησης παλινδρόμησης. Σε αυτό το σημείο τονίζεται ότι τα τυχαία σφάλματα διαφέρουν από τις παρακάτω ποσότητες:

$$e_t = y_t - \hat{y}_t = y_t - \hat{b}_0 - \hat{b}_1 x_t - \hat{b}_2 x_{t2} - \ldots - \hat{b}_k x_{tk}$$

οι οποίες ονομάζονται υπόλοιπα ή αλλιώς κατάλοιπα (*residuals*) και ορίζονται (Οικονόμου & Καρώνη, 2010) ως η κατακόρυφη απόκλιση της τιμής y_t από την ευθεία της προσαρμοσμένης συνάρτησης παλινδρόμησης. Η (*Σχέση 1*) αποτελεί ένα στοχαστικό υπόδειγμα, εκφράζει δηλαδή τη θεωρητική πληθυσμιακή γραμμική σχέση μεταξύ της εξαρτημένης και της ανεξάρτητης μεταβλητής, καθώς περιλαμβάνει και τον διαταρακτικό όρο ε_t , οι τιμές του οποίου δεν παρατηρούνται. Αυτό συμβαίνει διότι μπορεί να προέρχονται από άλλες ποσοτικά μετρήσιμες μεταβλητές, οι οποίες δεν περιλαμβάνονται στο υπόδειγμα, από άλλες ποσοτικά μη μετρήσιμες μεταβλητές που επηρεάζουν την εξαρτημένη μεταβλητή, από την γραμμικότητα του υποδείγματος, η οποία σε περίπτωση που δεν τηρείται έχει σαν συνέπεια την αύξηση του ε_t , καθώς και τέλος από όλους τους άλλους τυχαίους και απρόσμενους παράγοντες που επηρεάζουν την εξαρτημένη μεταβλητή.

Ο λόγος που πραγματοποιείται αυτή η παλινδρόμηση είναι για να εκτιμηθούν οι άγνωστες πληθυσμιακές παράμετροι b_0 , b_1 , b_2 , … , b_k , οι οποίες θα προσδιοριστούν με τη βοήθεια των εκτιμητριών[a] που θα προκύψουν από την παλινδρόμηση. Η εκτίμηση των παραμέτρων b_0 , b_1 , b_2 , … , b_k είναι ικανή να πραγματοποιηθεί με αρκετούς τρόπους, ο συνηθέστερος εκ των οποίων είναι η μέθοδος ελαχίστων τετραγώνων (*least squares method*). Η συγκεκριμένη μέθοδος, η οποία στηρίζεται στο Θεώρημα Gauss-Markov, είναι εξαιρετικά απλή και το κριτήριο στο οποίο βασίζεται είναι η ελαχιστοποίηση του αθροίσματος των τετραγώνων των υπολοίπων e_t. Έτσι λοιπόν οι παράμετροι του υποδείγματος b_0 , b_1 , b_2 , … , b_k υπολογίζονται με τέτοιο τρόπο ώστε το άθροισμα των τετραγώνων των υπολοίπων από τη γραμμή παλινδρόμησης να είναι το ελάχιστο δυνατό. Εναλλακτικά επομένως μπορούμε να πούμε ότι δεν μπορούμε να βρούμε άλλη γραμμή παλινδρόμησης της οποίας το άθροισμα των τετραγώνων των υπολοίπων να είναι μικρότερο από αυτό που προκύπτει ύστερα από την εφαρμογή της μεθόδου ελαχίστων τετραγώνων.

Όπως θα δούμε στη συνέχεια ο έλεγχος της ερμηνευτικής ικανότητας του μοντέλου διακρίνεται σε δύο επίπεδα: Στον έλεγχο της στατιστικής

[a] *Δηλαδή των* $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$, … , $\hat{b}_k$

σημαντικότητας των επιμέρους παραμέτρων (b_0 , b_1 , b_2 , … , b_k), καθώς και στον έλεγχο της ερμηνευτικής ικανότητας του μοντέλου συνολικά.

3.4.2 Έλεγχος της στατιστικής σημαντικότητας των επιμέρους παραμέτρων

- Έλεγχος υπόθεσης t για τους συντελεστές b_1 , b_2 , … , b_k:
  Στον έλεγχο της στατιστικής σημαντικότητας των επιμέρους παραμέτρων για i=1 , 2 , … , k οι υποθέσεις για τον συντελεστή b_i είναι:

$$H_0 : b_i = 0 \text{ έναντι της } H_1 : b_i \neq 0$$

- Σταθερός όρος b_0:
  Ο σταθερός όρος, σχεδόν χωρίς καμία εξαίρεση, πρέπει να συμπεριλαμβάνεται στο γραμμικό μοντέλο (Οικονόμου & Καρώνη, 2010). Παρόλο που γενικά η ανάλυση παλινδρόμησης συμπεριλαμβάνει έναν έλεγχο t με υπόθεση b_0=0, στην πλειονότητα των περιπτώσεων αυτός αγνοείται δεν χρησιμοποιείται προκειμένου να μας οδηγήσει σε μία πιθανή αφαίρεση του όρου από το μοντέλο. Τονίζεται επομένως ότι μία ενδεχόμενη αφαίρεση του σταθερού όρου b_0 οφείλεται μόνο στη φύση του φαινομένου που σκοπεύουμε να μοντελοποιήσουμε και όχι σε μία ενδεχόμενη δεκτή μηδενική υπόθεση. Η συμπεριφορά των μοντέλων, στα οποία παραλείπεται το b_0, διαφέρει σημαντικά. Ενδεικτικά αναφέρουμε ότι σε περίπτωση που αφαιρεθεί ο σταθερός όρος, ο συντελεστής προσδιορισμού R^2, ο οποίος θα αναλυθεί στη συνέχεια, δεν θεωρείται καλό μέτρο για την προσαρμογή του μοντέλου.

3.4.3 Έλεγχος της ερμηνευτικής ικανότητας του μοντέλου

- Στατιστικός Έλεγχος F:
  Προκειμένου να ελέγξουμε τη σημαντικότητα της παλινδρόμησης χρησιμοποιούμε τη στατιστική συνάρτηση F. Σε αντίθεση επομένως με τους επιμέρους ελέγχους υπόθεσης t, με αυτόν τον τρόπο ελέγχουμε όλους τους συντελεστές του μοντέλου ταυτόχρονα: συσχέτισης:

$$H_0 : b_1 = b_2 = \ldots = b_k = 0 \text{ έναντι της } H_1 : \text{τουλάχιστον ένα } b_j \neq 0$$

- Συντελεστής προσδιορισμού R^2:
  Μόλις προσδιοριστούν οι τιμές των συντελεστών και το υπόδειγμα είναι έτοιμο να δεχθεί τις τιμές των ανεξάρτητων μεταβλητών, προκειμένου να

βρεθεί η τιμή της εξαρτημένης μεταβλητής, θα πρέπει να εκτιμηθεί η ερμηνευτική ικανότητα του μοντέλου. Εναλλακτικά μπορούμε να πούμε ότι πρέπει να προσδιοριστεί ο βαθμός στον οποίο η δειγματική γραμμή της παλινδρόμησης εφαρμόζεται ικανοποιητικά στις παρατηρήσεις του δείγματος. Η ερμηνευτική ικανότητα του μοντέλου είναι ίση (Οικονόμου & Καρώνη, 2010) με τον συντελεστή προσδιορισμού R^2. Με άλλα λόγια το R^2 δηλώνει το βαθμό στον οποίο η εκτιμηθείσα γραμμή της παλινδρόμησης εφαρμόζεται στα δεδομένα του δείγματος. Ο συντελεστής προσδιορισμού δηλώνει επομένως το ποσοστό της μεταβλητότητας των τιμών της εξαρτημένης μεταβλητής που ερμηνεύεται από την παλινδρόμηση και ισούται με:

$$R^2 = \frac{\sum_{t=1}^{n}\left(\hat{y}_t - \bar{y}\right)^2}{\sum_{t=1}^{n}\left(y_t - \bar{y}\right)^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

όπου: $SST = SSR + SSE$, $SSE = \sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2$, $SST = \sum_{t=1}^{n}\left(y_t - \bar{y}_t\right)^2$

Το R^2 είναι επομένως ο λόγος δύο διακυμάνσεων: της διακύμανσης που προκύπτει από τις εκτιμηθείσες τιμές της εξαρτημένης μεταβλητής και του δειγματικού μέσου, προς τη διακύμανση που προκύπτει από τις παρατηρήσεις του δείγματος και τον δειγματικό μέσο. Όσο μεγαλύτερη είναι η τιμή του R^2, τόσο μεγαλύτερη είναι η ερμηνευτική ικανότητα του μοντέλου. Το R^2 δεν έχει μονάδες μέτρησης και παίρνει τιμές από το μηδέν μέχρι το ένα. Στη περίπτωση που το R^2 ισούται με τη μονάδα υπάρχει τέλεια γραμμική συσχέτιση και όλα τα ζεύγη των παρατηρήσεων βρίσκονται πάνω στην εκτιμηθείσα γραμμή. Αντίθετα όταν το R^2 είναι ίσο με το μηδέν δεν υπάρχει γραμμική σχέση και το μοντέλο δεν διαθέτει ερμηνευτική ικανότητα. Η τιμή 1-R^2 αναπαριστά το κατά πόσο η μεταβλητότητα των τιμών της εξαρτημένης μεταβλητής ερμηνεύεται από άλλους παράγοντες εκτός του υποδείγματος.

Παράλληλα όμως υπογραμμίζεται ότι σε ένα πολλαπλό γραμμικό υπόδειγμα (με περισσότερες από μία ανεξάρτητες μεταβλητές), το R^2 δεν είναι το καταλληλότερο κριτήριο αξιολόγησης της ερμηνευτικής ικανότητας του μοντέλου, καθώς όσο περισσότερες οι ανεξάρτητες μεταβλητές που διαθέτουμε, τόσο ψηλότερη θα είναι η τιμή του R^2 χωρίς αυτό να συνεπάγεται απαραίτητα και ισχυρότερη ερμηνευτική ικανότητα του υποδείγματος. Γι’ αυτό το λόγο αντί του συντελεστή προσδιορισμού R^2, χρησιμοποιείται ο προσαρμοσμένος συντελεστής προσδιορισμού, ο οποίος λαμβάνει υπ’ όψιν το πλήθος των ανεξάρτητων μεταβλητών, και ισούται με:

$$\bar{R}^2 = R^2_{adj} = 1 - \frac{SSE/(n-k)}{SST/(n-1)}$$

Όπου k το πλήθος των ανεξάρτητων μεταβλητών και n ο αριθμός των παρατηρήσεων.

3.4.4 Προϋποθέσεις του μοντέλου

Ωστόσο προκειμένου να είμαστε σε θέση να ελέγξουμε και να αξιολογήσουμε το προσαρμοσμένο μοντέλο, που προκύπτει ύστερα από την εφαρμογή της μεθόδου ελαχίστων τετραγώνων, πρέπει πρώτα να εξετάσουμε αν παραβιάζονται κάποιες προϋποθέσεις. Σε περίπτωση λοιπόν που παραβιάζεται ένας ή περισσότεροι εκ των παρακάτω ισχυρισμών, τότε το μοντέλο δεν είναι κατάλληλο προκειμένου να βγάλουμε συμπεράσματα σχετικά με τη συμπεριφορά των παρατηρήσεων.

- $E(\varepsilon_t) = 0$, για κάθε t

- $Var(\varepsilon_t) = Var(y_t) = \sigma^2$, για κάθε t

  Αν δεν ικανοποιείται η παραπάνω υπόθεση τα τυχαία σφάλματα δεν ικανοποιούν την υπόθεση της ομοσκεδαστικότητας (*homoscedasticity*). Αν λοιπόν η διακύμανση των τιμών του τυχαίου σφάλματος δεν είναι σταθερή δημιουργείται το πρόβλημα της ετεροσκεδαστικότητας (*heteroscedasticity*).

- $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$, για κάθε $i \neq j$

  Η παραπάνω υπόθεση σημαίνει ότι τα τυχαία σφάλματα είναι ασυσχέτιστα μεταξύ τους και η παραβίασή της προκαλεί το πρόβλημα της αυτοσυσχέτισης (*autocorrelation*), ή εναλλακτικά σειριακής συσχέτισης (*serial correlation*).

- $\varepsilon_t \approx N(0, \sigma^2)$, για κάθε t

  Δηλαδή οι τιμές του τυχαίου σφάλματος ακολουθούν την κανονική κατανομή. Ουσιαστικά οι τρεις πρώτες προϋποθέσεις περιλαμβάνονται σε αυτή και απλά διατυπώθηκαν ξεχωριστά προκειμένου να δοθεί μεγαλύτερη έμφαση σε αυτές.

- $\mathrm{cov}(e_t, x_{ti}) = 0$, για κάθε t και για όλα τα $i = 1, 2, \ldots k$

  Δηλαδή οι τιμές του τυχαίου σφάλματος δεν πρέπει να συσχετίζονται με τις ανεξάρτητες μεταβλητές x_1, x_2, … , x_k.

- $\operatorname{cov}\left(x_i, x_j\right) = 0$, για κάθε $i \neq j$

  Σε περίπτωση που έχουμε παραπάνω από μία ανεξάρτητες μεταβλητές, πρέπει αυτές να είναι ανεξάρτητες μεταξύ τους. Η παραβίαση αυτής της υπόθεσης δημιουργεί το πρόβλημα της πολυσυγγραμμικότητας (multicollinearity).

Όσον αφορά τις πέντε πρώτες υποθέσεις, οι τιμές των τυχαίων σφαλμάτων ε_t είναι άγνωστες, γεγονός που μας υποχρεώνει να τις εξετάσουμε υποχρεωτικά μέσω των υπολοίπων e_t (*residuals*) χρησιμοποιώντας τον τύπο:

$$e_t = y_t - \hat{y}_t$$

Η εξέταση των παραπάνω προϋποθέσεων πραγματοποιείται είτε μέσω ελέγχων (όπως για παράδειγμα (SchweserNotes, 2013) Breusch-Pagan, ή Spearman), είτε συνηθέστερα μέσω συγκεκριμένων γραφικών παραστάσεων, οι οποίοι θα παρουσιαστούν στη συνέχεια.

3.4.5 Εξέταση υπολοίπων

Έχοντας λάβει υπόψη τις παραπάνω προϋποθέσεις, οι οποίες πρέπει να ισχύουν προκείμενου το μοντέλο να είναι κατάλληλο για να πραγματοποιήσει προβλέψεις, θα εξετάσουμε τα υπόλοιπα χρησιμοποιώντας (Οικονόμου & Καρώνη, 2010) τρεις ομάδες γραφικών παραστάσεων:

- Διάγραμμα διασποράς των υπολοίπων e_t συναρτήσει των εκτιμημένων $\hat{y}_t$.
  (Αν το συγκεκριμένο διάγραμμα δεν εμφανίζει καλή εικόνα μπορούμε να βγάλουμε το συμπέρασμα ότι παραβιάζεται η υπόθεση της ύπαρξης ομοσκεδαστικότητας. Παράλληλα μπορούμε να εντοπίσουμε και πιθανά απόμακρα σημεία, τα οποία ίσως χρειαστεί να αφαιρεθούν.)

- Διάγραμμα των υπολοίπων e_t συναρτήσει του χρόνου, καθώς και διάγραμμα διασποράς των e_t επί των e_{t-1}.
  (Αν τα συγκεκριμένα διαγράμματα δεν έχουν καλές εικόνες συμπεραίνουμε ότι τα τυχαία σφάλματα δεν είναι ασυσχέτιστα μεταξύ τους και επομένως υπάρχει το πρόβλημα της αυτοσυσχέτισης.)
- Γραφική παράσταση για την εξέταση της Κανονικής κατανομής των υπολοίπων (*Normal Probability Plot*)
  (Στο συγκεκριμένο διάγραμμα συγκρίνουμε τα υπόλοιπα e_t με την κανονική κατανομή. Αν το συγκεκριμένο διάγραμμα δεν έχει καλή εικόνα τότε τα

υπόλοιπα e_t δεν ακολουθούν την κανονική κατανομή. Σε αυτή τη περίπτωση ίσως χρειάζεται κάποιος μετασχηματισμός της εξαρτημένης μεταβλητής y_t , ή η αφαίρεση κάποιων παρατηρήσεων που δημιουργούν απομονωμένα σημεία στη γραφική μας παράσταση.)

- Διάγραμμα διασποράς των υπολοίπων e_t συναρτήσει των ανεξάρτητων μεταβλητών x_1, x_2, … , x_k.
  (Στα συγκεκριμένα διαγράμματα πρέπει να εμφανίζονται τυχαία διάσπαρτα σημεία, να μην υπάρχει δηλαδή ένας συστηματικός τρόπος συμπεριφοράς των σημείων. Ειδάλλως, αν σε κάποια από τις γραφικές παραστάσεις παρατηρήσουμε ότι τα υπόλοιπα εξαρτώνται από τις τιμές που λαμβάνει η αντίστοιχη ανεξάρτητη μεταβλητή, ενδεχομένως να απαιτείται μετασχηματισμός της μεταβλητής αυτής.)

Κεφάλαιο 4. Η αποτύπωση της Ελληνικής οικονομικής κρίσης στο Twitter

4.1 Εισαγωγή

Κατά τη διάρκεια του 2010 γνωστοποιείται ότι τα επίπεδα του δημοσιονομικού ελλείμματος της Ελλάδας για το 2009 ξεπερνούν κατά πολύ αυτά που θα το καθιστούσαν βιώσιμο. Η άμεσες επιπτώσεις δεν άργησαν να φανούν, καθώς η ελληνική κυβέρνηση αδυνατούσε να δανειστεί από τις αγορές με επιτόκια σε λογικά πλαίσια, προκειμένου να ανταπεξέλθει στη χρηματοδότηση του τρέχοντος δημοσιονομικού ελλείμματος και την αναχρηματοδότηση του χρέους. Η κυβέρνηση πραγματοποιεί προσπάθειες με απώτερο στόχο την ανατροπή του το αρνητικού κλίματος και την μείωση των επιτοκίων, μειώνοντας τις δαπάνες του δημοσίου, χωρίς όμως επιτυχία. Ο κίνδυνος στάσης πληρωμών του Ελληνικού δημοσίου είναι πλέον ορατός, με αποτέλεσμα στις 23 Απριλίου του 2010, δια στόματος πρωθυπουργού, η Ελλάδα να καταφύγει στη βοήθεια του Διεθνούς Νομισματικού Ταμείου, της Ευρωπαϊκής Επιτροπής και της Ευρωπαϊκής Κεντρικής Τράπεζας, που αναλαμβάνουν τη συγκρότηση του μηχανισμού βοήθειας για τη χώρα. Τα επόμενα χρόνια η Ελληνική οικονομία αδυνατεί να ισορροπήσει, με αποτέλεσμα να ψηφίζονται απανωτά μέτρα λιτότητας και περικοπές, οδηγώντας τον Ελληνικό λαό στην ανεργία και στην οικονομική και ψυχολογική εξουθένωση.

Δεν είναι λοιπόν δύσκολο λοιπόν να αντιληφθούμε ότι τα οικονομικά προβλήματα και οι δυσχέρειες που εμφανίζονται τα τελευταία χρόνια, τόσο στην Ελλάδα όσο και στο εξωτερικό, συνοδεύονται, όπως είναι φυσικό, από έντονο άγχος που μπορεί να οδηγήσει σε σοβαρότερα ψυχολογικά ζητήματα και σε αύξηση των προβλημάτων ψυχικής υγείας. Η οικονομική κρίση έχει ήδη ωθήσει χιλιάδες πολίτες στην ανεργία, τα ποσοστά της οποίας αυξάνονται συνεχώς. Εδώ και χρόνια, είναι γνωστή η άμεση συσχέτιση μεταξύ ανεργίας και συναισθηματικών διαταραχών. Έρευνες δείχνουν ότι στις εταιρίες που έγιναν περικοπές μισθών ή απολύσεις, διπλασιάστηκαν τα συμπτώματα καταθλίψεων και σε ορισμένες περιπτώσεις μάλιστα καταγράφηκαν ακόμη πιο έντονα συμπτώματα σε αυτούς που ζούσαν υπό την απειλή της απόλυσης, παρά σε όσους πραγματικά απολύθηκαν.

Καθώς οι Έλληνες πολίτες βιώνουν όλο και πιο έντονα την ανασφάλεια σε μια δυσοίωνη οικονομική περίοδο, παρουσιάζονται σημαντικές επιπτώσεις στη ψυχική διάθεση. Το ερώτημα που προκύπτει και το οποίο θα διερευνηθεί στο συγκεκριμένο κεφάλαιο είναι αν τα σημάδια της αρνητικής ψυχολογίας των τελευταίων ετών, αντικατοπτρίζονται και στο χώρο του διαδικτύου.

Τα Social media καθιστούν εύκολη και άμεση τη συλλογή μεγάλων ποσοτήτων δεδομένων, που παράγονται μέσα από την επικοινωνία των χρηστών. Η διαδικασία αυτή αποτελεί ισχυρό εργαλείο, στα χέρια των αναλυτών και των κοινωνιολόγων, έτσι ώστε να αποκτούν πρόσβαση στο δημόσιο αίσθημα και στην κοινή γνώμη. Η συγκεκριμένη διεργασία, μέχρι πριν λίγα χρόνια, ήταν ικανή να επιτευχθεί μόνο μέσω των δημοσκοπήσεων. Σε αυτό το σημείο πρέπει να υπογραμμιστεί ότι μία μελέτη, που πραγματοποιείται μέσω της χρήσης των μέσων κοινωνικής δικτύωσης, εμπεριέχει αξιοσημείωτα πλεονεκτήματα σε σύγκριση με μία παραδοσιακή δημοσκοπική έρευνα, αλλά παράλληλα υποκρύπτει και κάποιες σημαντικές δυσκολίες.

Πιο συγκεκριμένα μέσω των μέσων κοινωνικής δικτύωσης καθίσταται δυνατή η πρόσβαση σε εξαιρετικά μεγάλους σε μέγεθος πληθυσμούς. Παράλληλα η διαδικασία αυτή μπορεί να πραγματοποιηθεί με τρομακτική εξοικονόμηση χρόνου και χρήματος, καθώς για παράδειγμα μπορούμε με μία απλή αναζήτηση, σε κάποιο social media, να αποκτήσουμε δίοδο σε δεκάδες απόψεις χρηστών, ενώ για παράδειγμα κατά τη διάρκεια μιας τηλεφωνικής έρευνας να απαιτούνται αρκετές ώρες προκειμένου να σχηματίσουμε ένα επαρκές δείγμα και κατάλληλο για έρευνα. Παράλληλα γίνεται εφικτός ο στιγμιαίος υπολογισμός των κοινωνικών τάσεων, γνωστός και ως "*nowcasting*" [a], που σημαίνει ότι οι αναλυτές είναι ικανοί να προσδιορίσουν τη κατάσταση της κοινωνίας σε κάθε δεδομένη χρονική στιγμή, διαδικασία η οποία ήταν αδύνατο να επιτευχθεί με τις συνηθισμένες δημοσκοπικές έρευνες.

Αντίθετα η ανίχνευση των κοινωνικών τάσεων σε πραγματικό χρόνο, μέσω των social media, παρουσιάζει και αρκετά εμπόδια. Ειδικότερα, προκειμένου να βγουν συμπεράσματα που αφορούν τη κοινωνία μέσα από τις δημοσιοποιήσεις των χρηστών, απαιτούνται πολύπλοκες τεχνικές text mining, προκειμένου να αποκρυπτογραφηθεί και να φανερωθεί η ουσία και το νόημα του κειμένου, που υποθάλπεται πίσω από τις λέξεις. Επίσης μία άλλη σημαντική δυσκολία, η οποία επιβάλλεται να ληφθεί υπ' όψιν κατά την εξαγωγή πληροφοριών από τα μέσα κοινωνικής δικτύωσης, είναι ότι το

[a] Όρος που δημιουργήθηκε προκειμένου να γίνει εμφανής η αντίθεση με τον όρο "*forecasting*" που χρησιμοποιείται σε διαδικασίες πρόβλεψης (οικονομικές, μετεωρολογικές, κοκ).

δείγμα που εξετάζεται δεν αποτελεί αντιπροσωπευτικό μέρος του γενικού πληθυσμού. Για παράδειγμα τα social media είναι ιδιαίτερα δημοφιλή στις νεαρές ηλικίες, με αποτέλεσμα σε μία έρευνα να μην είναι όσο ορατές θα έπρεπε να είναι οι απόψεις των μεγαλύτερων ατόμων σε ηλικία. Ωστόσο τονίζεται ότι τα σφάλματα που δημιουργούνται από τα παραπάνω εμπόδια είναι ικανά να αντιμετωπιστούν, καθώς όπως έχει προαναφερθεί, τα δείγματα που μπορούμε να συλλέξουμε μέσα από τα social media είναι υπέρογκα.

Στη συνέχεια θα εφαρμόσουμε τα εργαλεία αυτά, προκειμένου να εξετάσουμε αν η αρνητικότητα των Ελλήνων, που οφείλεται στην έντονη οικονομική κρίση των τελευταίων ετών, καθρεφτίζεται και στο χώρο του Twitter. Παράλληλα θα διερευνηθεί κατά τη διάρκεια ποιων περιόδων οι Έλληνες χρήστες του δημοφιλούς social media ήταν, με βάση τα λεγόμενα τους, χαρούμενοι, δυστυχισμένοι, ή εξαγριωμένοι και θα επιχειρηθεί να γίνει μία αντιστοίχιση με τα γεγονότα της επικαιρότητας, κατά την αντίστοιχη χρονική περίοδο.

4.2 Η εμφάνιση και η ανάλυση της “ευτυχίας” στο Twitter

Όπως έχει ήδη αναφερθεί, προκειμένου να μελετηθεί το συναίσθημα της ευτυχίας στο Twitter, αναζητήθηκαν 15 οικογένειες λέξεων, οι οποίες περιέχουν με τη σειρά τους 368 λέξεις που είναι συνώνυμες, ή παραπέμπουν σε κάποιο ευχάριστο συναίσθημα. Οι λέξεις αυτές αναζητηθήκαν μέσα στο χρονικό διάστημα των 236 εβδομάδων, προκειμένου να καταγραφεί ο αριθμός των Tweets που τις περιέχουν. Έτσι σε συνδυασμό με το πλήθος των ουδέτερων λέξεων, ανά δεδομένη ημέρα, είμαστε σε θέση να σχηματίσουμε μία χρονοσειρά, η οποία σκιαγραφεί την πορεία της σχετικής συχνότητας των εμφανίσεων των λέξεων που παραπέμπουν στο συναίσθημα της ευτυχίας. Στην παραπάνω εικόνα εμφανίζονται οι δεκαπέντε οικογένειες λέξεων που αναζητήθηκαν, το μέγεθος της γραμματοσειράς των οποίων είναι συνάρτηση της συχνότητας με την οποία αυτές εμφανίστηκαν στο Twitter.

Εικόνα 4.1 - Οι 15 οικογένειες λέξεων, σχετικές με την ευτυχία, οι οποίες αναζητήθηκαν μέσα στο χρονικό διάστημα των 1638 ημερών

Στη συνέχεια θα κατασκευαστεί, θα απεικονιστεί και θα αναλυθεί η χρονοσειρά $R=\{R_1 , R_2 , \ldots , R_{1638}\}$ η οποία φανερώνει τη μεταβολή του συναισθήματος της ευτυχίας (*Raw joy signal*) μέσα στο Twitter, μέσα στο χρονικό διάστημα των 236 εβδομάδων που μελετήθηκαν. Παράλληλα θα πραγματοποιήσουμε και μία εξομάλυνση 15 ημερών (*smoothing*) της χρονοσειράς R, δημιουργώντας έτσι μία δεύτερη χρονοσειρά S, χρησιμοποιώντας τη μέθοδο της κινητής μέσης τιμής (*moving average*). Η διαδικασία αυτή πραγματοποιείται προκειμένου να αποκτήσουμε μία γενικότερη εικόνα των μεταβολών που πραγματοποιούνται κατά τη πάροδο του χρόνου, καθώς και επίσης για να είμαστε σε θέση να παρατηρήσουμε καλύτερα τις αυξομειώσεις και τις τάσεις της χρονοσειράς μας, μέσω της εξομάλυνσης των ακραίων τιμών.

Θέτοντας $J=\{J^{(1)}, J^{(2)}, \ldots, J^{(368)}\}$ το σύνολο των λέξεων, που είναι συνυφασμένες με το συναίσθημα της ευτυχίας (*Joy*), καθώς και με όμοιο τρόπο $N=\{N^{(1)}, N^{(2)}, N^{(3)}, N^{(4)}\}$ τις ουδέτερες (*Neutral*) λέξεις που αναζητήσαμε, πραγματοποιούμε τα παρακάτω βήματα προκειμένου να επιτευχθεί ο σχηματισμός της χρονοσειράς μας:

- Θέτουμε $J_t^{(n)}$ το πλήθος των Tweets που περιέχουν την *n*-οστή λέξη, η οποία δηλώνει ευτυχία (*Joy*), κατά τη διάρκεια της *t* ημέρας.
- Θέτουμε $N_t^{(n)}$ το πλήθος των Tweets που περιέχουν την *n*-οστή ουδέτερη (*Neutral*) λέξη, κατά τη διάρκεια της *t* ημέρας.
- Υπολογίζουμε τις παρακάτω σχετικές συχνότητες εμφάνισης των λέξεων που δηλώνουν ευτυχία, σε σχέση με τις ουδέτερες λέξεις, για κάθε μία από τις 1638 ημέρες που μελετάμε:

$$Y_t = \frac{\sum_{n=1}^{368} J_t^{(n)}}{\sum_{n=1}^{4} N_t^{(n)}} \text{, για κάθε } t=1,\ldots,1638$$

- Ολοκληρώνουμε τη κατασκευή της χρονοσειράς $R=\{R_1, R_2, \ldots, R_{1638}\}$ εκφράζοντας τις παραπάνω ποσότητες σε τυποποιημένη μορφή, μέσω του γνωστού τύπου:

$$R_t = \frac{Y_t - \mu}{\sigma} \text{, για κάθε } t=1,\ldots,1638$$

Όπου μ η μέση τιμή και σ η τυπική απόκλιση του δείγματος, δηλαδή:

$$\mu = \frac{1}{1638}\sum_{t=1}^{1638} Y_t \text{ και } \sigma = \sqrt{\frac{1}{1638}\sum_{t=1}^{1638}\left(Y_t - \mu\right)^2}$$

- Τέλος, εξομαλύνουμε την *R*, κατασκευάζοντας μία δεύτερη χρονοσειρά $S=\{S_8, S_9, \ldots, S_{1631}\}$, της οποίας οι τιμές υπολογίζονται με τη μέθοδο της κινητής μέσης τιμής (*moving average*) ως εξής:

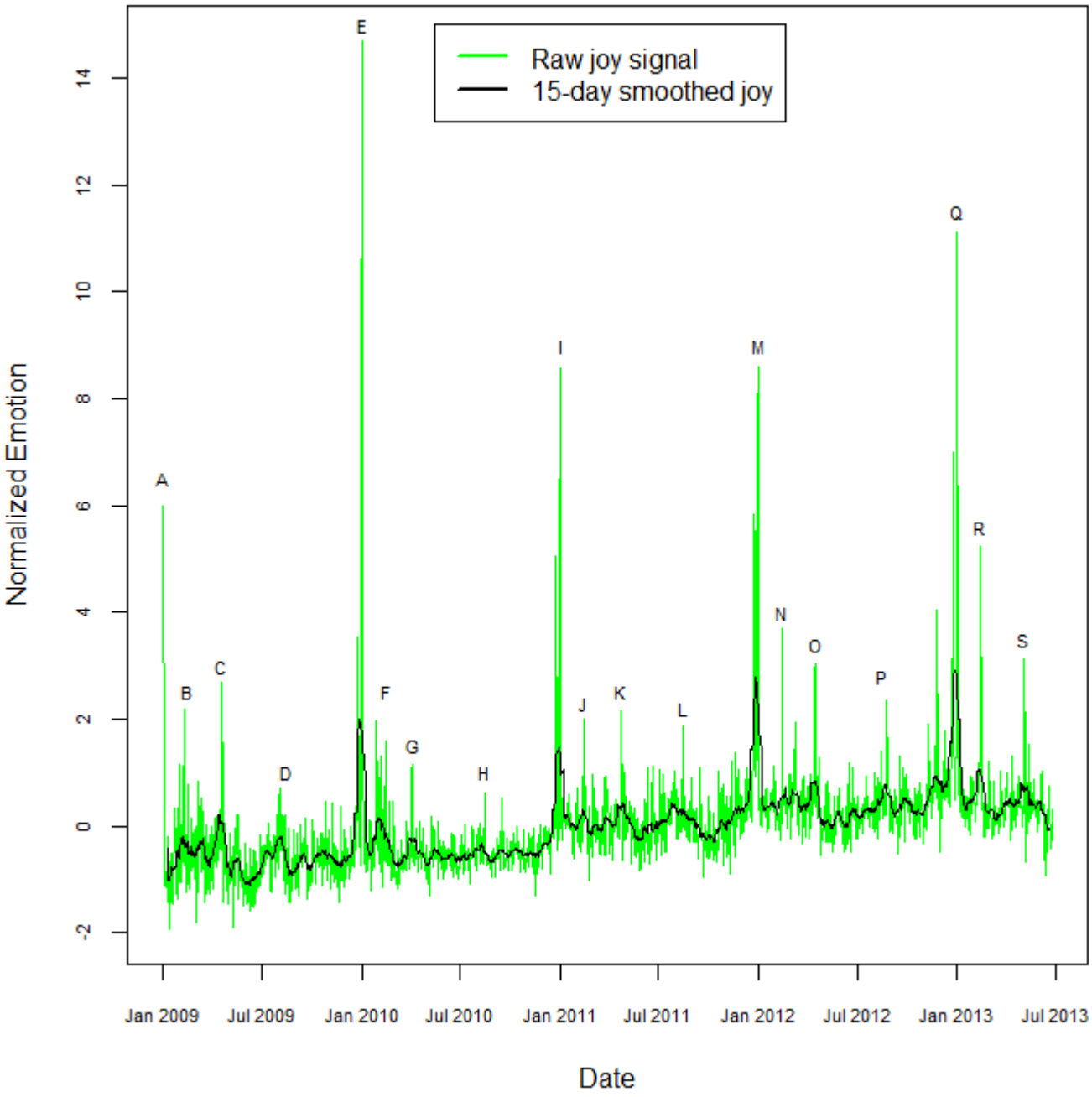
$$S_t = \frac{1}{15}\sum_{i=-7}^{7} R_{t+i} \text{, για κάθε } t=8,\ldots,1631$$

ή αλλιώς: $S_t = \frac{1}{15}R_{t-7} + \ldots + \frac{1}{15}R_{t-1} + \frac{1}{15}R_t + \frac{1}{15}R_{t+1} + \ldots + \frac{1}{15}R_{t+7}$

Προφανώς η εξομαλυμένη χρονοσειρά S χάνει συνολικά 14 παρατηρήσεις (επτά στην αρχή της και επτά στο τέλος της), κάτι που όμως δεν περιορίζει την μακροσκοπική και γενικότερη εικόνα της χρονοσειράς που θέλουμε να αποκτήσουμε.

Στο παρακάτω διάγραμμα αποτυπώνονται οι χρονοσειρές R και S που σχηματίστηκαν προηγουμένως και μας φανερώνουν τη μεταβολή του συναισθήματος της ευτυχίας μέσα στο Twitter ,από την 1η Ιανουαρίου του

2009, έως και την 26η Ιουνίου του 2013. Η γραφική παράσταση της R (πράσινο χρώμα) μας δείχνει την ημερήσια διακύμανση της ευτυχίας στο Twitter, ενώ η αντίστοιχη της S (μαύρο χρώμα) τις μεταβολές του δεκαπενθήμερου μέσου όρου της.



Διάγραμμα 4.2 - Time Series for Joy in Twitter Content (1638 days)

Στη παραπάνω γραφική παράσταση μπορούμε εύκολα να παρατηρήσουμε μία έντονη περιοδικότητα από έτος σε έτος. Πιο συγκεκριμένα τα τοπικά μέγιστα κάθε άλλο παρά τυχαία είναι, καθώς η πλειονότητα τους εμφανίζεται

κατά περιόδους διακοπών. Προκειμένου κάτι τέτοιο να γίνει κατανοητό, σημειώθηκαν 19 σημεία (A έως S) τα οποία αναλύονται παρακάτω:

- Τα σημεία **A**, **E**, **I**, **M** και **Q** αποτελούν τις διακοπές των Χριστουγέννων των ετών 2009, έως και 2013 αντίστοιχα. Στα σημεία αυτά, όπως φαίνεται στο διάγραμμα, εμφανίζονται τα εντονότερα μέγιστα της χρονοσειράς. Και στις πέντε περιπτώσεις η μεγαλύτερη τιμή εμφανίζεται κατά την Πρωτοχρονιά, ενώ παράλληλα ιδιαίτερα μεγάλες τιμές εμφανίζονται μία εβδομάδα νωρίτερα, στις 25 Δεκεμβρίου, την ημέρα δηλαδή που γιορτάζονται τα Χριστούγεννα.
- Τα σημεία **B**, **F**, **J**, **N** και **R** αφορούν την ημέρα του Αγίου Βαλεντίνου των ετών 2009, έως και 2013 αντίστοιχα.
- Τα σημεία **C**, **G**, **K**, **O** και **S** αντιστοιχίζονται στις διακοπές του Πάσχα των ετών 2009, έως και 2013 αντίστοιχα. Όπως μπορούμε εύκολα να παρατηρήσουμε στη γραφική μας παράσταση, κατά τις συγκεκριμένες χρονικές περιόδους αυξάνεται έντονα στο συναίσθημα της ευτυχίας, με απόγειο την Κυριακή του Πάσχα.
- Τα σημεία **D**, **H**, **L** και **P** περιέχονται μέσα στις καλοκαιρινές διακοπές των ετών 2009, έως και 2012 αντίστοιχα και ειδικότερα αντιστοιχίζονται στη γιορτή του Δεκαπενταύγουστου, που αποτελεί την κορύφωση των διακοπών των Ελλήνων, με το συντριπτικά μεγαλύτερο ποσοστό αυτών να βρίσκεται μακριά από την καθημερινότητά του. Παράλληλα τονίζεται ότι εκτός από το έντονο μέγιστο της χρονοσειράς R που παρατηρείται και στα τέσσερα σημεία, ανήμερα δηλαδή της αργίας του Δεκαπενταύγουστου, μπορούμε να διακρίνουμε επίσης και μία γενικότερη αύξηση, κατά τους καλοκαιρινούς μήνες, στην εμφάνιση του συναισθήματος της ευτυχίας στο Twitter. Κάτι τέτοιο διακρίνεται κυρίως από την τάση της χρονοσειράς S, η οποία κατά τη διάρκεια των καλοκαιρινών διακοπών αυξάνει έντονα τις τιμές της, ενώ αντίθετα κατά τους μήνες πριν και μετά από αυτές φθίνει σε μεγάλο βαθμό.

4.3 Η εμφάνιση και η ανάλυση της “απογοήτευσης” στο Twitter

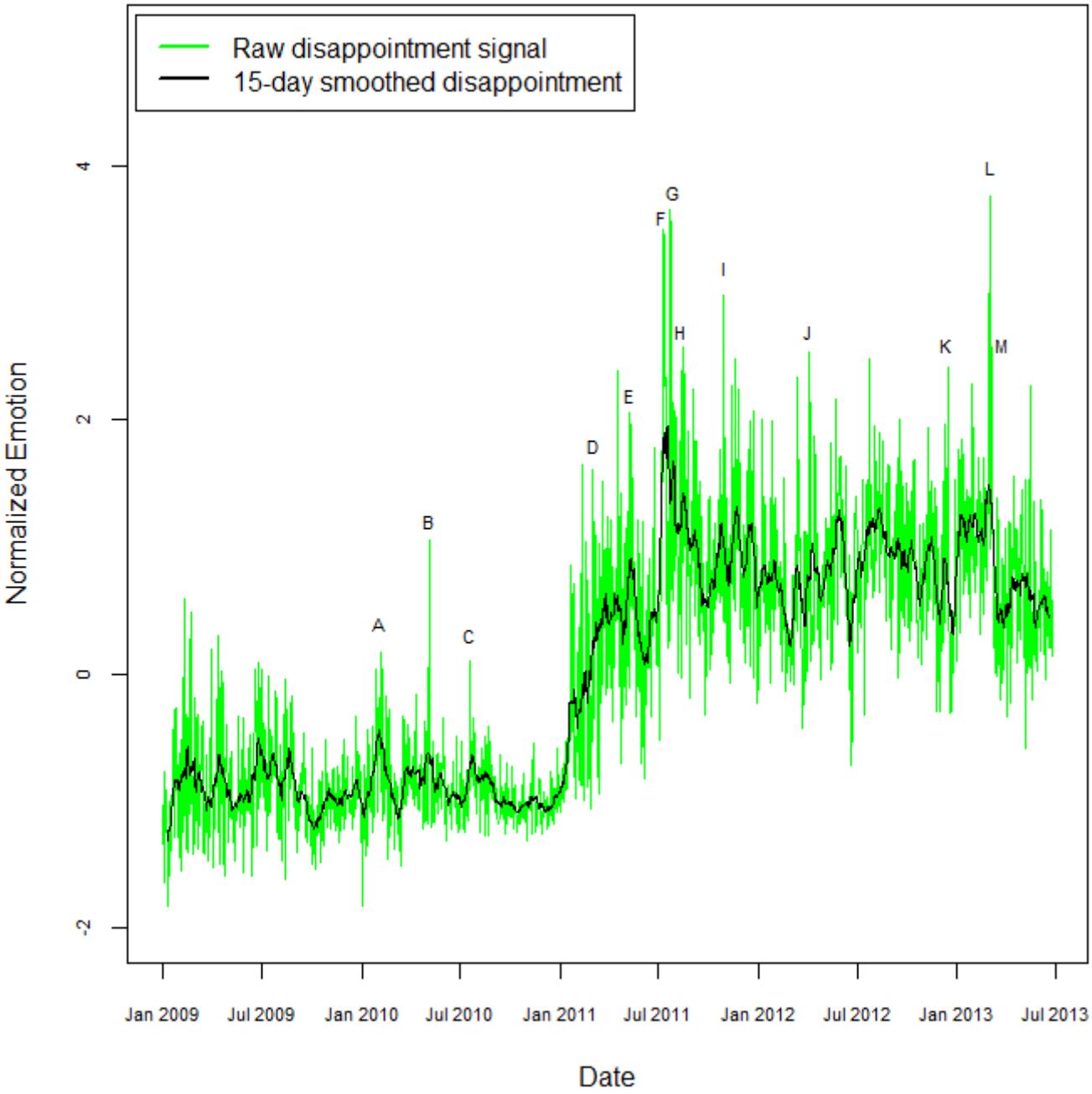
Όπως έχει ήδη αναφερθεί, προκειμένου να μελετηθεί το συναίσθημα της απογοήτευσης στο Twitter, αναζητήθηκαν 23 οικογένειες λέξεων, οι οποίες περιέχουν με τη σειρά τους 289 λέξεις που είναι συνώνυμες, ή παραπέμπουν σε κάποιο δυσάρεστο συναίσθημα. Στην παρακάτω εικόνα εμφανίζονται οι είκοσι τρείς οικογένειες λέξεων που αναζητήθηκαν, το μέγεθος της γραμματοσειράς των οποίων είναι συνάρτηση της συχνότητας με την οποία αυτές εμφανίστηκαν στο Twitter.

Εικόνα 4.3 - Οι 23 οικογένειες λέξεων, σχετικές με την απογοήτευση, οι οποίες αναζητήθηκαν μέσα στο χρονικό διάστημα των 1638 ημερών

Στη συνέχεια, θέτοντας $D=\{D^{(1)}, D^{(2)}, \ldots, D^{(289)}\}$ το σύνολο των λέξεων, που είναι ταυτισμένες με το συναίσθημα που μελετάμε, το οποίο είναι αυτό της απογοήτευσης (*Disappointment*), σχηματίζουμε με όμοιο τρόπο τις χρονοσειρές R και S, χρησιμοποιώντας τον τύπο:

$$Y_t = \frac{\sum_{n=1}^{289} D_t^{(n)}}{\sum_{n=1}^{4} N_t^{(n)}} \text{, για κάθε } t=1,\ldots,1638$$

Έτσι ολοκληρώνουμε την κατασκευή της χρονοσειράς $R=\{R_1, R_2, \ldots, R_{1638}\}$, για το συναίσθημα της απογοήτευσης, εκφράζοντας τις παραπάνω ποσότητες Y_t σε τυποποιημένη μορφή. Για την κατασκευή της εξομαλυμένης χρονοσειράς $S=\{S_8, S_9, \ldots, S_{1631}\}$, χρησιμοποιούμε τη μέθοδο της κινητής μέσης τιμής (*moving average*). Στη συνέχεια παρουσιάζονται οι γραφικές παραστάσεις των δύο χρονοσειρών R (πράσινο χρώμα) και S (μαύρο χρώμα) που σχηματίσαμε, οι οποίες μας φανερώνουν τη μεταβολή του συναισθήματος της απογοήτευσης μέσα στο Twitter ,από την 1η Ιανουαρίου του 2009, έως και την 26η Ιουνίου του 2013.



Διάγραμμα 4.4 - Time Series for Disappointment in Twitter Content (1638 days)

Σε αντίθεση με την προηγούμενη παράγραφο και το αντίστοιχο διάγραμμα της εμφάνισης της ευτυχίας στο Twitter, στη παραπάνω γραφική παράσταση μπορούμε άμεσα να παρατηρήσουμε μία ραγδαία αύξηση των Tweets που εμπεριέχουν δυστυχία, από τα τέλη του 2010 και ύστερα. Κάτι τέτοιο οφείλεται στο γεγονός ότι η αρνητικότητα και η απογοήτευση των Ελλήνων, που οφείλεται στην έντονη οικονομική κρίση των τελευταίων ετών, καθρεφτίζεται και στο χώρο του Twitter. Επίσης τα τοπικά μέγιστα κάθε άλλο παρά τυχαία είναι, καθώς η πλειονότητα τους εμφανίζεται κατά περιόδους έντονης οικονομικής αστάθειας και κυρίως δυσάρεστων κοινωνικών

φαινομένων που αποτελούν εκβάσεις αυτής. Προκειμένου το συμπέρασμα αυτό να γίνει πιο συγκεκριμένο, σημειώθηκαν 13 σημεία (Α έως Μ), από το 2010 και μετά, τα οποία θα αναλυθούν παρακάτω:

- Σημείο **Α** (*2 έως 5 Φεβρουαρίου 2010*):
  Από τα μέσα Ιανουαρίου παρατηρείται μία αυξητική τάση των Tweets που φανερώνουν δυστυχία, εμφανίζοντας τη μέγιστη τιμή τους την 3η Φεβρουαρίου. Αυτό οφείλεται στο γεγονός ότι βρισκόμαστε χρονικά λίγες μέρες πριν την ανακοίνωση της προσφυγής στον μηχανισμό στήριξης (23η Απριλίου), με αποτέλεσμα τα ΜΜΕ και γενικότερα η κοινωνία, να αναφέρονται στο οικονομικό τέλμα, στο οποίο η χώρα έχει περιέλθει. Από τα μέσα Ιανουαρίου του 2010 ξεκίνησε μία συνεχής άνοδος των *spreads*, που αναφερόταν συχνά στον τύπο ως ράλι ή κούρσα των *spreads*. Στις 21 Ιανουαρίου του 2010, το spread των 10ετών ομολόγων ξεπέρασε (Έθνος, 2010) τις 300 μονάδες, ενώ στη συνέχεια ακολούθησε (Έθνος, 2010) ξέφρενη πορεία ξεπερνώντας ακόμα και τις 1000 μονάδες τον Απρίλιο του 2010. Το αποτέλεσμα είναι, κάτω από τη πίεση αυτή, στις 2 Φεβρουαρίου να φουντώνουν οι φήμες για το πρώτο πακέτο οικονομικών μέτρων (Ελευθεροτυπία, 2010). Παράλληλα στις 3 Φεβρουαρίου του 2010 εμφανίζονται δημοσιεύματα με στοιχεία (The Best News, 2010) του ΟΑΕΔ, σύμφωνα με τα οποία μέσα στον Ιανουάριο του 2010 οι άνεργοι έφτασαν τους 770.615, ενώ σε διάστημα μόλις ενός έτους (Ιανουάριος 2009 - Ιανουάριος 2010) προστέθηκαν στο σύνολο των ανέργων επιπλέον 208.098 άτομα, αριθμός που αποτελεί ρεκόρ όλων των εποχών και με αναγωγή στο συνολικά απασχολούμενο εργατικό δυναμικό (4,5 εκατ. άτομα) εκτινάσσει το ποσοστό της πραγματικής ανεργίας στο 17,1%.
- Σημείο **Β** (*5 και 6 Μαΐου 2010*):
  Στις 6 Μαΐου ψηφίζεται από τη βουλή το μνημόνιο που υπέγραψε η κυβέρνηση, με την Ευρωπαϊκή Ένωση και το Διεθνές Νομισματικό Ταμείο, για τα μέτρα που θα ληφθούν, προκειμένου να ενεργοποιηθεί ο μηχανισμός στήριξης. Μέσα σε τρεις μόλις μήνες έχουν ανακοινωθεί τρία πακέτα οικονομικών μέτρων. Έτσι λοιπόν , μία μέρα πριν τη ψήφιση του μνημονίου, στις 5 Μαΐου, κηρύχθηκε πανελλαδική απεργία, κατά των νέων μέτρων λιτότητας και της εμπλοκής της Ελλάδας στο ΔΝΤ. Η πορεία της Αθήνας ήταν μία από τις μεγαλύτερες που έχει καταγραφεί στην Ελλάδα στην πρόσφατη ιστορία της, ενώ μαζικότατες διαδηλώσεις πραγματοποιήθηκαν στη Θεσσαλονίκη και σε όλες τις μεγάλες πόλεις της περιφέρειας. Η πορεία της Αθήνας σημαδεύτηκε από τον θάνατο τριών εργαζομένων της τράπεζας *Marfin Bank* (η μία μάλιστα εκ των οποίων ήταν έγκυος). Το γεγονός έγινε πρώτο θέμα τόσο στα εγχώρια και στα ξένα μέσα ενημέρωσης, όσο και στα

social media, γεγονός που, σε συνδυασμό με τη ψήφιση του μνημονίου, δικαιολογεί απόλυτα την εκτίναξη του συναισθήματος της θλίψης στο Twitter.

- Σημείο **C** (*19 Ιουλίου 2010*):
  Επίθεση σε βάρος του δημοσιογράφου Σωκράτη Γκιόλα, γεγονός (Ζώτου, 2010) που πυροδότησε την ανάρτηση εκατοντάδων σχετικών μηνυμάτων στα μέσα κοινωνικής δικτύωσης. Η ευθύνη αναλαμβάνεται από οργάνωση, με προκήρυξή (Τα Νέα, 2010) της σε εφημερίδα, στις 27 Ιουλίου του 2010.
- Σημείο **D** (*3 Μαρτίου 2011*):
  Πυρά (Tvxs, 2013) σε βάρος δύο νεαρών αστυνομικών, με αποτέλεσμα μεγάλος όγκος μηνυμάτων να κατακλύζει τα social media, μέσω των οποίων οι χρήστες εκφράζουν την απογοήτευσή τους για την εκτόξευση της εγκληματικότητας κατά τη διάρκεια των τελευταίων ετών.
- Σημείο **E** (*10 Μαΐου 2011*):
  Επίθεση, με δυσάρεστη έκβαση, σε 44χρονο στο κέντρο της Αθήνας, λίγο πριν γίνει πατέρας, που αποσκοπεί στη ληστεία (News247, 2011) της κάμεράς του. Για ακόμα μία φορά οι αντιδράσεις στο Twitter για πρόβλημα της εγκληματικότητας των ελληνικών πόλεων, είναι μεγάλες.
- Σημείο **F** (*11 Ιουλίου 2011*):
  Σημειώνεται (Ελευθεροτυπία, 2011) έκρηξη στη ναυτική βάση "*Ευάγγελος Φλωράκης*" της Εθνικής Φρουράς Κύπρου, 20 χιλιόμετρα ανατολικά της Λεμεσού. Χιλιάδες Κύπριοι και Έλληνες χρήστες του διαδικτύου άλλαξαν τις εικόνες τους στα διάφορα κοινωνικά δίκτυα, σε κατάμαυρες εικόνες προς ένδειξη σεβασμού προς τα θύματα της έκρηξης.
- Σημείο **G** (*23 Ιουλίου 2011*):
  Την προηγούμενη ημέρα λαμβάνει χώρα (The Telegraph, 2011) διπλή τρομοκρατική επίθεση στη Νορβηγία, η οποία αποτελεί τη φονικότερη στη χώρα μετά το τέλος του Δευτέρου Παγκοσμίου Πολέμου (77 θύματα, κυρίως έφηβοι). Η είδηση μονοπώλησε τόσο τα ΜΜΕ ολόκληρου του κόσμου, όσο και τα μέσα κοινωνικής δικτύωσης, με αποτέλεσμα και οι Έλληνες χρήστες να εκφράζουν τη θλίψη και το αποτροπιασμό τους για το συμβάν.
- Σημείο **H** (*6 έως 13 Αυγούστου 2011*):
  Μεταξύ 6 και 11 Αυγούστου (αλλά και με περιστατικά που λαμβάνουν χώρα πολύ αργότερα από το συγκεκριμένο χρονικό διάστημα) ξεσπούν στο Λονδίνο αλλά και σε ολόκληρη την Αγγλία εξεγέρσεις, που ονομάστηκαν (Economist, 2011) "*BlackBerry riots*". Η ονομασία αυτή πηγάζει από το γεγονός ότι οι διαδηλωτές χρησιμοποιούσαν τα κινητά τους τηλέφωνα και τα social media (κυρίως το Twitter), προκειμένου να οργανώνονται, κάτι που μας αποδεικνύει για ακόμα μία φορά την επιρροή τους στη σύγχρονη κοινωνία. Με αφορμή τις εξεγέρσεις και την έκβαση αυτών (BBC News, 2011), το Twitter σε ολόκληρη την Αγγλία εστιάζει στο θέμα (μάλιστα τη Δευτέρα 8 Αυγούστου, στο Ηνωμένο

Βασίλειο, οι 4 εκ των 170 επισκέψεων στο διαδίκτυο αφορούσαν (Tonkin, Pfeiffer, & Tourte, 2012) το Twitter), φαινόμενο που δεν άργησε να επεκταθεί και στη χώρα μας.

- Σημείο **I** (*27 έως 30 Οκτωβρίου 2011*):
  Σε συμφωνία, για κούρεμα του Ελληνικού χρέους κατά 50% και πρόσθετο πακέτο βοήθειας προς την Ελλάδα ύψους 130 δις ευρώ, έρχονται τα ξημερώματα της 27ης Οκτωβρίου η ευρωζώνη και οι τράπεζες (Το Βήμα, 2011). Παράλληλα την επομένη οι εορτασμοί της 28ης Οκτωβρίου μετατρέπονται (Ελευθεροτυπία, 2011)σε πανελλήνια διαδήλωση. Αποκορύφωμα αποτέλεσε η ματαίωση της στρατιωτικής παρέλασης στη Θεσσαλονίκη, μετά την εισβολή χιλιάδων διαδηλωτών στο χώρο της λεωφόρου Μεγάλου Αλεξάνδρου.
- Σημείο **J** (*4 Απριλίου 2012*):
  Τα διεθνή ΜΜΕ και τα μέσα κοινωνικής δικτύωσης, κάνουν λόγο για αλματώδη αύξηση των αυτοκτονιών στην Ελλάδα, με αφορμή δυσάρεστο συμβάν στη πλατεία Συντάγματος (Έθνος, 2012).
- Σημείο **K** (*14 Δεκεμβρίου 2012*):
  Το “*Μακελειό του Κονέκτικατ*”, με τραγικό απολογισμό 28 θύματα συγκλονίζει τις ΗΠΑ καθώς και ολόκληρο τον κόσμο (ΣΚΑΪ, 2012). Η Google αποφάσισε να συμμετάσχει στην αγρυπνία που έγινε στο Κονέκτικατ, βάζοντας ένα μικρό λεπτό κερί στην κεντρική της σελίδα (24wro, 2012).
- Σημείο **L** (*1 και 2 Μαρτίου 2013*):
  Δυσάρεστη έκβαση για τους δύο εκ των πέντε φοιτητών του ΤΕΙ Λάρισας κατά την προσπάθειά τους να ζεσταθούν με ένα αυτοσχέδιο μαγκάλι (Tvxs, 2014).
- Σημείο **M** (*5 και 6 Μαρτίου 2013*):
  Θάνατος του πρόεδρου της Βενεζουέλας Ούγκο Τσάβες (Tvxs, 2013), σε ηλικία 58 ετών. Μόνο στη Λατινική Αμερική δημοσιοποιούνται 2000 tweets ανά δευτερόλεπτο, που σχετίζονται (Radio Netherlands Worldwide, 2013) με το δυσάρεστο γεγονός.

4.4 Η εμφάνιση και η ανάλυση της "αγανάκτησης" στο Twitter

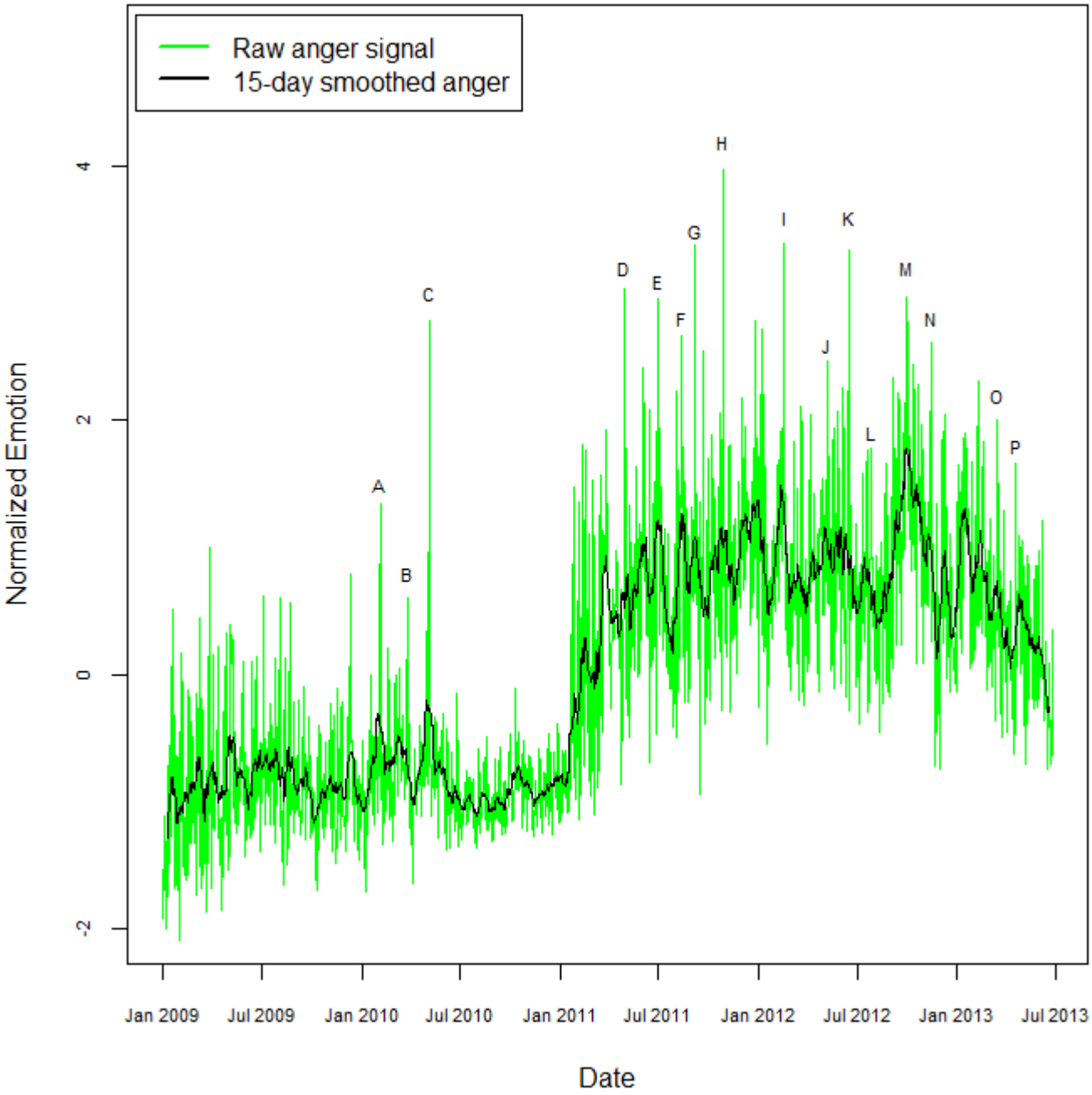
Όπως έχει ήδη αναφερθεί, προκειμένου να μελετηθεί το συναίσθημα της αγανάκτησης στο Twitter, αναζητήθηκαν 23 οικογένειες λέξεων, οι οποίες περιέχουν με τη σειρά τους 270 λέξεις. Οι λέξεις αυτές είναι συνώνυμες με την αγανάκτηση, ή παραπέμπουν σε αυτή από την κοινωνικοπολιτική οπτική γωνία. Στη διπλανή εικόνα εμφανίζονται οι είκοσι τρείς οικογένειες λέξεων που αναζητήθηκαν, το μέγεθος της γραμματοσειράς των οποίων είναι συνάρτηση της συχνότητας με την οποία αυτές εμφανίστηκαν στο Twitter.

Εικόνα 4.5 - Οι 23 οικογένειες λέξεων, σχετικές με την αγανάκτηση, οι οποίες αναζητήθηκαν μέσα στο χρονικό διάστημα των 1638 ημερών

Στη συνέχεια, θέτοντας $A=\{A^{(1)}, A^{(2)}, \ldots, A^{(270)}\}$ το σύνολο των λέξεων, που είναι ταυτισμένες με το συναίσθημα που μελετάμε, το οποίο είναι αυτό της αγανάκτησης (*Anger*), σχηματίζουμε με όμοιο τρόπο τις χρονοσειρές R και S, χρησιμοποιώντας τον εξής τύπο:

$$Y_t = \frac{\sum_{n=1}^{270} A_t^{(n)}}{\sum_{n=1}^{4} N_t^{(n)}} \text{, για κάθε } t=1,\ldots,1638$$

Έτσι ολοκληρώνουμε την κατασκευή της χρονοσειράς $R=\{R_1, R_2, \ldots, R_{1638}\}$, για το συναίσθημα της αγανάκτησης, εκφράζοντας τις παραπάνω ποσότητες Y_t σε τυποποιημένη μορφή, ενώ για την κατασκευή της εξομαλυμένης χρονοσειράς $S=\{S_8, S_9, \ldots, S_{1631}\}$, χρησιμοποιούμε τη μέθοδο της κινητής μέσης τιμής (*moving average*). Στη συνέχεια παρουσιάζονται οι γραφικές απεικονίσεις των δύο χρονοσειρών R (πράσινο χρώμα) και S (μαύρο χρώμα) που σχηματίστηκαν, οι οποίες μας φανερώνουν τη μεταβολή του συναισθήματος της αγανάκτησης μέσα στο Twitter ,από την 1η Ιανουαρίου του 2009, έως και την 26η Ιουνίου του 2013.



Διάγραμμα 4.6 - Time Series for Anger in Twitter Content (1638 days)

Ομοίως με την προηγούμενη παράγραφο και το αντίστοιχο διάγραμμα της εμφάνισης της απογοήτευσης στο Twitter, στη παραπάνω γραφική παράσταση διακρίνουμε μία απότομη αύξηση των Tweets που εμπεριέχουν αγανάκτηση, από τα μέσα του 2010 και ύστερα. Το αίτιο του φαινομένου αυτού είναι και πάλι η έντονη οικονομική κρίση των τελευταίων ετών, που μαστίζει την ελληνική κοινωνία. Προκειμένου το συμπέρασμα αυτό να γίνει πιο συγκεκριμένο, σημειώθηκαν 15 σημεία (Α έως Ο), από το 2010 και μετά, τα οποία θα αναλυθούν στη συνέχεια:

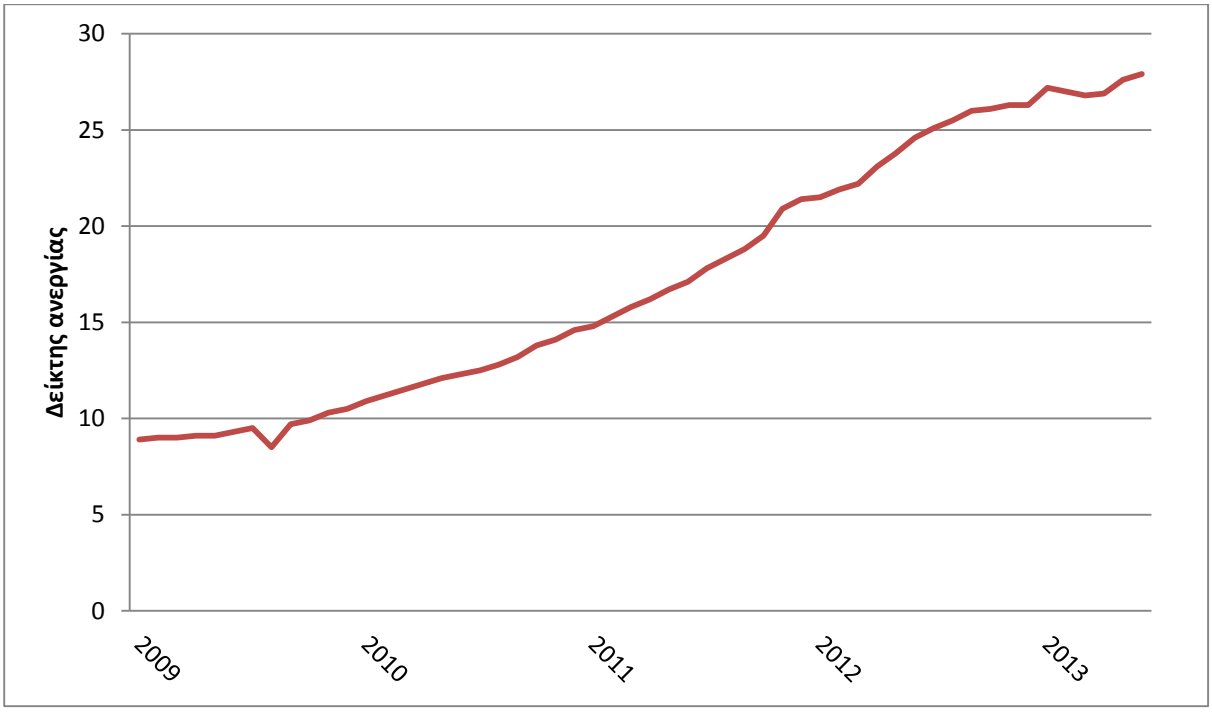
- Σημείο **A** (*2 έως 5 Φεβρουαρίου 2010*):
  Πρώτο πακέτο οικονομικών μέτρων (Βλ. ανάλυση του σημείου **A**, στο αντίστοιχο διάγραμμα απογοήτευσης).
- Σημείο **B** (*25 και 26 Μαρτίου 2010*):
  Ρατσιστικά συνθήματα ακούγονται στο κέντρο της Αθήνας από βατραχανθρώπους των Ειδικών Δυνάμεων, την ώρα που παρήλαυναν, στο πλαίσιο της επίσημης στρατιωτικής παρέλασης της 25^{ης} Μαρτίου (Ελευθεροτυπία, 2010).
- Σημείο **C** (*5 και 6 Μαΐου 2010*):
  Ψήφιση του πρώτου μνημονίου και υπόθεση Marfin (Βλ. ανάλυση του σημείου **B**, στο αντίστοιχο διάγραμμα απογοήτευσης).
- Σημείο **D** (*25 Μαΐου 2011*):
  Για δεκάδες ημέρες, με εκκίνηση την συγκέντρωση της 25ης Μαΐου, πολλές χιλιάδες Ελλήνων "Αγανακτισμένων" διαδήλωσαν ειρηνικά σε πολλές πόλεις της Ελλάδας, με σημείο τριβής τα τεκταινόμενα στην ελληνική οικονομία.
- Σημείο **E** (*28 έως 30 Ιουνίου 2011*):
  Κατά την 28^{η} Ιουλίου (πρώτη ημέρα πανελλήνιας απεργίας και κινητοποιήσεων) μεγάλο πλήθος κόσμου κατέφθασε στην Αθήνα, με στόχο τον αποκλεισμό της Βουλής, εν όψει της ψηφοφορίας του μεσοπρόθεσμου πλαισίου δημοσιονομικής πολιτικής 2012-2015, ενώ παράλληλα ογκώδεις συγκεντρώσεις πραγματοποιούνται και σε όλες τις μεγάλες πόλεις της χώρας. Την επόμενη ημέρα η Βουλή ενέκρινε το μεσοπρόθεσμο σχέδιο, ενώ την ίδια ώρα στην πλατεία Συντάγματος πραγματοποιούνται διαδηλώσεις, ο απολογισμός των οποίων ήταν δεκάδες τραυματίες (Ελευθεροτυπία, 2011).
- Σημείο **F** (*6 έως 13 Αυγούστου 2011*):
  Εξεγέρσεις του Λονδίνου (Βλ. ανάλυση του σημείου **H**, στο αντίστοιχο διάγραμμα απογοήτευσης).
- Σημείο **G** (*6 Σεπτεμβρίου 2011*):
  Στη λήψη νέων αποφάσεων (News247, 2011), οι οποίες σχετίζονται με τα κλειστά επαγγέλματα, τους μισθούς και την εφεδρεία στο δημόσιο, κατέληξε το υπουργείο οικονομικών, που συνεδρίασε ενόψει της εκταμίευσης της 6^{ης} δόσης του δανείου προς την Ελλάδα.
- Σημείο **H** (*28 Οκτωβρίου 2011*):
  Κούρεμα του Ελληνικού χρέους και διαδηλώσεις κατά τους εορτασμούς της 28^{ης} Οκτωβρίου (Βλ. ανάλυση του σημείου **I**, στο αντίστοιχο διάγραμμα απογοήτευσης).
- Σημείο **I** (*10 έως 17 Φεβρουαρίου 2012*):
  Στις 12 Φεβρουαρίου τέθηκε στη Βουλή για ψήφιση το δεύτερο μνημόνιο το οποίο και υπερψηφίστηκε (Το Βήμα, 2012).
- Σημείο **J** (*6 έως 8 Μαΐου 2012*):

Διεξαγωγή εκλογών (6 Μαΐου), από τις οποίες δεν προκύπτει κυβέρνηση (Public Issue, 2012).

- Σημείο **K** (*15 έως 17 Ιουνίου 2012*):
  Στις 17 Ιουνίου διεξάγονται νέες εκλογές, οι οποίες οδηγούν στη δημιουργία κυβέρνησης συνασπισμού (Η Καθημερινή, 2012).
- Σημείο **L** (*27 Ιουλίου 2012*):
  Πραγματοποιείται η τελετή έναρξης των Θερινών Ολυμπιακών Αγώνων στο Λονδίνο, οι οποίοι παραλληλίζονται με αυτούς της Αθήνας του 2004, οι οποίοι κόστισαν 9 δισεκατομμύρια ευρώ στο ελληνικό δημόσιο, καθιστώντας τους τους ακριβότερους αγώνες που διοργανώθηκαν μέχρι τότε.
- Σημείο **M** (*30 Σεπτεμβρίου έως 3 Οκτωβρίου 2012*):
  Στις 3 Οκτωβρίου ο κατάλογος Ελλήνων καταθετών στην τράπεζα HSBC της Ελβετίας, γνωστός και ως λίστα Λαγκάρντ, έπειτα από μια “περίεργη” διαδρομή σε υπουργικά και πολιτικά γραφεία, βρίσκεται τελικά στα χέρια της δικαιοσύνης, ενώ παράλληλα φουντώνουν οι υποψίες περί συγκάλυψης φοροφυγάδων και αλλοίωσης της λίστας κατά τη χρονική περίοδο που αυτή ήταν εξαφανισμένη (Newsbomb, 2012).
- Σημείο **N** (*5 έως 14 Νοεμβρίου 2012*):
  Στις 5 Νοεμβρίου κατατίθεται (Πρώτο Θέμα, 2012) στη Βουλή, και δύο μέρες αργότερα υπερψηφίζεται, το μεσοπρόθεσμο πλαίσιο δημοσιονομικής στρατηγικής 2013-2016.
- Σημείο **O** (*16 Μαρτίου 2013*):
  Η οικονομική κρίση της Κυπριακής Δημοκρατίας που ξεκίνησε το 2012, με κύριο αίτιο την μεγάλη έκθεση των κυπριακών τραπεζών σε ελληνικά ομόλογα, γίνεται πλέον πιο εμφανής από ποτέ. Πιο συγκεκριμένα στις 16 Μαρτίου η τρόικα (Ευρωπαϊκή Επιτροπή – Ευρωπαϊκή Κεντρική Τράπεζα – Διεθνές Νομισματικό Ταμείο) επέρχεται σε συμφωνία με την Κύπρο για δάνειο ύψους 10 δισεκατομμυρίων ευρώ. Επίσης αποφασίζεται να αντληθούν 5,8 δισεκατομμύρια ευρώ από κούρεμα των καταθέσεων (RTE News, 2013).
- Σημείο **P** (*17 και 18 Απριλίου 2013*):
  Έντονες αντιδράσεις δημιουργούνται για τις συνθήκες εργασίας των μεταναστών στην Ελλάδα, με αφορμή την δολοφονική επίθεση κατά εκατοντάδων αλλοδαπών εργατών, σε χωράφια φράουλας στη Μανωλάδα Ηλείας, όταν αυτοί ζήτησαν τα δεδουλευμένα έξι μηνών, από τους τρεις Έλληνες επιστάτες. Στα διεθνή ΜΜΕ η κατακραυγή της Ελλάδας, με αφορμή το περιστατικό, παίρνει μεγάλες διαστάσεις (Tvxs, 2013). Παράλληλα στο διαδίκτυο ξεκινάει εκστρατεία (Tvxs, 2013), η οποία καλεί σε μποϊκοτάζ των συγκεκριμένων φραουλών, ενώ συγκεκριμένα στο Twitter, τα Hashtags #bloodstrawberries, #boycott και #manolada είναι εκ των δημοφιλέστερων εκείνων των ημερών.

4.5 Συσχέτιση της ανεργίας με την απογοήτευση - αποστροφή

Σε αυτή τη παράγραφο θα εξετάσουμε αν τα δεδομένα που συλλέξαμε παραπάνω συσχετίζονται με το έντονο κοινωνικό πρόβλημα της ανεργίας, που μαστίζει την σημερινή ελληνική κοινωνία. Στο παρακάτω διάγραμμα φαίνεται η μηνιαία ανεργία, μέσα στο χρονικό διάστημα που μελετάμε, σύμφωνα με τα στοιχεία (Ελληνική Στατιστική Αρχή, 2013) που προκύπτουν από τα δελτία τύπου που δημοσιοποιούνται κάθε μήνα από την ΕΛ.ΣΤΑΤ. (Ελληνική Στατιστική Αρχή).



Διάγραμμα 4.7 - Δείκτης ανεργίας (Ιανουάριος 2009 – Ιούνιος 2013)

Θεωρώντας λοιπόν ως την ανεξάρτητη μεταβλητή μας x το δείκτη ανεργίας και ως εξαρτημένη y τον μηνιαίο μέσο όρο[a] των εμφανίσεων των δύο αρνητικών συναισθημάτων (απογοήτευση και αγανάκτηση) στο Twitter, θα υπολογίσουμε το δειγματικό συντελεστή συσχέτισης του Pearson (*Pearson's correlation coefficient*), ο οποίος εκφράζει το βαθμό της γραμμικής συσχέτισης μεταξύ των δύο μεταβλητών:

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

[a] Δηλαδή τους μηνιαίους μέσους όρους των Y_t που δημιουργήθηκαν στις παραγράφους *4.4* και *4.5*.

όπου $S_{xy}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$, $S_{xx}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$, $S_{yy}=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$

Τα x_i και y_i αντιπροσωπεύουν τις 54 παρατηρήσεις, των μεταβλητών x και y αντίστοιχα, για κάθε έναν από τους 54 μήνες που περιέχεται στο χρονικό διάστημα που μελετάμε. Έτσι λοιπόν, χρησιμοποιώντας τα δεδομένα μας καταλήγουμε στα εξής αποτελέσματα:

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=0.23\ ,\ \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=17.26\ ,\ S_{xy}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)=27.57$$

$$S_{xx}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2=2317.13\ ,\ S_{yy}=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=0.44$$

Και τελικά:

$$r_{xy}=\frac{S_{xy}}{\sqrt{S_{xx}\cdot S_{yy}}}=0.865$$

Σε αυτό το σημείο σημειώνεται ότι όσο πιο κοντά βρίσκεται η τιμή του δειγματικού συντελεστή συσχέτισης του Pearson στο 1 ή στο -1 τόσο ισχυρότερη είναι η συσχέτιση μεταξύ δύο μεταβλητών, ενώ αντίστροφα αν βρίσκεται κοντά στο 0, τότε δεν υπάρχει. Άρα με βάση τη τιμή του συντελεστή συσχέτισης Pearson, παρατηρούμε ότι υπάρχει ισχυρή γραμμική συσχέτιση μεταξύ των δύο μεταβλητών. Ωστόσο όμως πρέπει να τονίσουμε ότι δεν αρκεί να υπολογίσουμε μόνο τον δειγματικό συντελεστή συσχέτισης προκειμένου να βγάλουμε ασφαλή συμπεράσματα για τη συμμεταβολή των δύο μεταβλητών. Ο λόγος που συμβαίνει αυτό είναι επειδή οι παρατηρήσεις x_i και y_i αντιπροσωπεύουν μία συγκεκριμένη χρονική περίοδο με αποτέλεσμα να εμπεριέχεται ο κίνδυνος, η γραμμική συσχέτιση που αποδείχθηκε ότι υπάρχει, να είναι τυχαία. Γι' αυτόν το λόγο πραγματοποιούμε τον παρακάτω έλεγχο, θέτοντας ρ_{xy} τον πληθυσμιακό, πραγματικό συντελεστή συσχέτισης:

$$H_0:\rho_{xy}=0 \text{ έναντι της } H_1:\rho_{xy}\neq 0$$

Υποθέτοντας ότι οι δύο μεταβλητές είναι κανονικά κατανεμημένες πραγματοποιούμε έναν έλεγχο t προκείμενου να απορρίψουμε την μηδενική υπόθεση (*null hypothesis*), χρησιμοποιώντας την παρακάτω ελεγχοσυνάρτηση:

$$t=\frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Όπου, $n-2=52$ είναι οι βαθμοί ελευθερίας (*degrees of freedom*) καθώς το δείγμα μας περιέχει 54 παρατηρήσεις. Επίσης τονίζεται ότι η παραπάνω ελεγχοσυνάρτηση υπό την H_0 ακολουθεί την κατανομή t_{n-2}.

Θα εξετάσουμε αν ο δειγματικός συντελεστής συσχέτισης που υπολογίστηκε παραπάνω είναι στατιστικά σημαντικός (Βλ. Παράρτημα Β):

$$t=\frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}=\frac{0.865\sqrt{52}}{\sqrt{1-0.865^2}}=12.416$$

$$(p\text{-}\tau\iota\mu\acute{\eta}<0.0001)$$

Άρα η μηδενική υπόθεση απορρίπτεται και το συμπέρασμα που βγάζουμε είναι ότι η συσχέτιση, μεταξύ της ανεργίας και της εμφάνισης των συναισθημάτων της απογοήτευσης και της αγανάκτησης στο Twitter, είναι στατιστικά σημαντική.

Κεφάλαιο 5. Αποτύπωση της οικονομικής δραστηριότητας στο Google

5.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο παρατηρήθηκε ότι ένα κοινωνικοπολιτικό φαινόμενο αποτυπώνεται σε μεγάλο βαθμό στα μέσα κοινωνικής δικτύωσης. Αυτό συμβαίνει διότι στις μέρες μας τείνουμε ολοένα και περισσότερο να ενημερωνόμαστε, για τα θέματα που αφορούν την καθημερινότητά μας, μέσα από το διαδίκτυο αλλά και παράλληλα να δημοσιεύουμε τις απόψεις μας στα διάφορα social media. Το αποτέλεσμα είναι να δημιουργείται ένας μεγάλος όγκος δεδομένων, ο οποίος αν αξιοποιηθεί και αναλυθεί κατάλληλα μπορεί να μας οδηγήσει σε συμπεράσματα, όπως αυτά του προηγούμενου κεφαλαίου.

Το ερώτημα όμως που άμεσα προκύπτει είναι αν οι πληροφορίες, που αντλούμε από τα μέσα κοινωνικής δικτύωσης, είναι ικανές να μας βοηθήσουν στην ανάλυση πιο πολύπλοκων μοντέλων, όπως για παράδειγμα η πρόβλεψη της οικονομικής δραστηριότητας μιας επιχείρησης. Αξιοσημείωτο είναι το γεγονός ότι οι επενδυτές για να πολλαπλασιάσουν τα κέρδη τους, οι οικονομολόγοι με σκοπό να συγκεντρώσουν στοιχεία για τις έρευνές τους, αλλά και οι δημοσιογράφοι προκειμένου να ενημερώσουν την κοινωνία, παρακολουθούν εκατοντάδες μηνιαία οικονομικά δελτία, τα οποία εκδίδουν οι κυβερνήσεις και οι διάφορες εταιρείες. Ωστόσο τα συγκεκριμένα στοιχεία πολλές φορές αποτελούν και τροχοπέδη στις παραπάνω διεργασίες επιβραδύνοντας τες σε μεγάλο βαθμό, καθώς δημοσιεύονται μέρες ή ακόμα και μήνες αργότερα από την χρονική περίοδο την οποία αφορούν, ενώ παράλληλα είναι εξαιρετικά σύνηθες το γεγονός αρκετούς μήνες αργότερα να δημοσιεύονται αναθεωρημένες εκδόσεις τους, κάτι που οδηγεί σε λανθασμένες υποθέσεις και άρα σε εσφαλμένα συμπεράσματα.

Αντίθετα τα μέσα κοινωνικής δικτύωσης μπορούν να αποτελέσουν ένα εργαλείο το οποίο εμπεριέχει μία εξαιρετικά μεγάλη πηγή δεδομένων η οποία μάλιστα ανανεώνεται συνεχώς και συμπεριλαμβάνει ελάχιστα περιθώρια σφαλμάτων. Πιο συγκεκριμένα η Google μέσω της υπηρεσίας της Google Trends παρέχει στους χρήστες ημερήσιες εκθέσεις σχετικές με τον όγκο των αναζητήσεων που πραγματοποιήθηκαν στη μηχανή αναζήτησής της. Αν

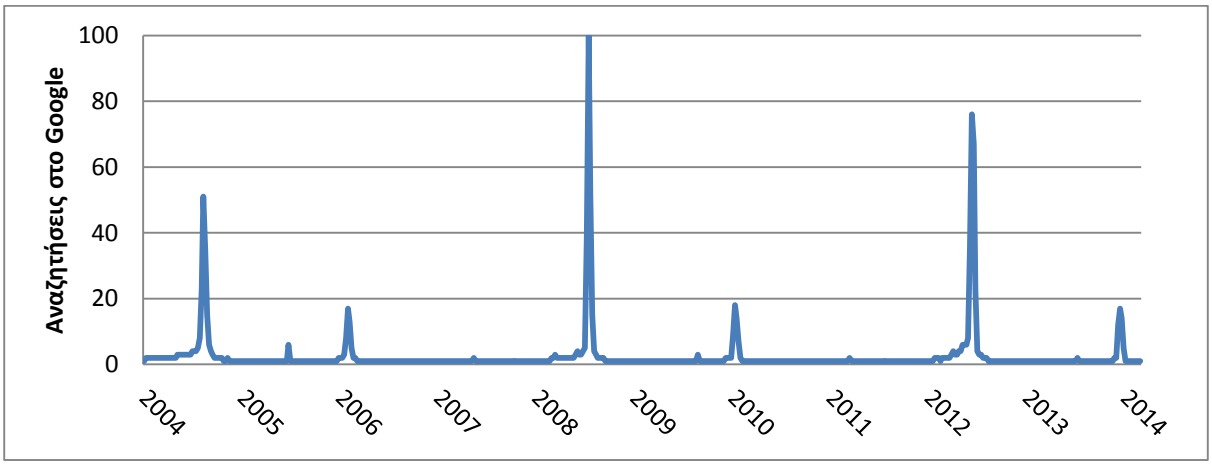
επομένως τα στοιχεία αυτά αξιοποιηθούν κατάλληλα θα είναι ικανά να χρησιμοποιηθούν προκειμένου να αποτυπωθεί η πορεία διαφόρων οικονομικών δραστηριοτήτων, όπως για παράδειγμα η οικονομία μιας χώρας, η πρόβλεψη των πωλήσεων μιας επιχείρησης, κτλ.

Έτσι λοιπόν στο συγκεκριμένο κεφάλαιο θα υποθέσουμε ότι οι αναζητήσεις μιας βιομηχανίας στο Google, ίσως συσχετίζονται με το τρέχον επίπεδο της οικονομικής της δραστηριότητας. Ως εκ τούτου, αν η συγκεκριμένη υπόθεση αποδειχθεί αληθής, ο αριθμός των αναζητήσεων μίας επιχείρησης στο διαδίκτυο ενδέχεται να είναι χρήσιμος προκειμένου να προβλεφθούν οι πωλήσεις της εταιρείας στο άμεσο μέλλον. Παράλληλα πρέπει να υπογραμμιστεί ότι αν η συγκεκριμένη υπόθεση είναι αληθής δεν σημαίνει ότι τα "χτυπήματα" μίας αυτοκινητοβιομηχανίας για παράδειγμα στο διαδίκτυο μέσα σε ένα τρίμηνο, είναι ικανά προκειμένου να προβλέψουν τις πωλήσεις αυτοκινήτων για το επόμενο έτος, καθώς οι πωλήσεις αυτές εξαρτώνται και από άλλους πολύ ισχυρούς παράγοντες, όπως για παράδειγμα οι τιμές του πετρελαίου, το ΑΕΠ (ακαθάριστο εγχώριο προϊόν) των χωρών που πωλούνται τα αυτοκίνητα, η εποχικότητα των πωλήσεων (για παράδειγμα το καλοκαίρι αυξάνονται οι πωλήσεις κάμπριο αυτοκινήτων), και ούτω καθεξής. Αντίθετα αυτό που ισχυριζόμαστε είναι ότι ο όγκος των αναζητήσεων στο Google μιας συγκεκριμένης μάρκας αυτοκινήτων μέσα σε κάποιο χρονικό διάστημα, ίσως είναι ανάλογος με τον αριθμό των πωλήσεων της ίδιας μάρκας μέσα στο ίδιο χρονικό διάστημα και ίσως είναι ικανός για να κάνουμε μία μικρή πρόβλεψη για το προσεχές μέλλον.

Ο στόχος μας επομένως σε αυτό το κεφάλαιο θα είναι να παρουσιάσουμε τις παγκόσμιες αναζητήσεις της αυτοκινητοβιομηχανίας Α στο Google και να τις συγκρίνουμε με τις πωλήσεις της. Στη συνέχεια θα εξετάσουμε αν αυτά τα δύο μεγέθη σχετίζονται μεταξύ τους. Αν κάτι τέτοιο αποδειχθεί αληθές, θα αποτελέσει ισχυρό εργαλείο στα χέρια μας, προκειμένου να μπορέσουμε να πραγματοποιήσουμε μικρές μελλοντικές προβλέψεις των πωλήσεων της εταιρείας, χρησιμοποιώντας τους αριθμούς που λαμβάνουμε μέσω της υπηρεσίας Google Trends (η κατασκευή μοντέλου πρόβλεψης πωλήσεων θα πραγματοποιηθεί στο επόμενο κεφάλαιο).

5.2 Συλλογή δεδομένων της αυτοκινητοβιομηχανίας Α

Όπως έχει ήδη αναφερθεί το Google Trends δεν παρουσιάζει τον απόλυτο αριθμό αναζητήσεων μιας λέξης, αλλά εμφανίζει, συναρτήσει του χρόνου, το λόγο των αναζητήσεων σε μία δεδομένη γεωγραφική περιοχή, ανά το πλήθος των συνολικών αναζητήσεων στην ίδια περιοχή. Επίσης η Google πολλαπλασιάζει όλους τους λόγους με έναν αριθμό έτσι ώστε ο μέγιστος λόγος να γίνει ίσος με 100, για να μπορούμε να παρατηρούμε την ποσοστιαία απόκλιση των υπόλοιπων δεδομένων σε σχέση με αυτόν. Για παράδειγμα στο παρακάτω διάγραμμα εμφανίζονται οι εβδομαδιαίες παγκόσμιες αναζητήσεις των λέξεων "*Olympic Games*" στο Google από τον Ιανουάριο του 2004 έως τον Απρίλιο του 2014. Έτσι λοιπόν η συγκεκριμένη γραφική παράσταση εμπεριέχει 539 τιμές (όσες και οι εβδομάδες μέσα στο συγκεκριμένο χρονικό διάστημα), κάθε μία από τις οποίες ισούται με τον αριθμό των αναζητήσεων των λέξεων "*Olympic Games*", ανά το σύνολο των παγκόσμιων αναζητήσεων κατά την αντίστοιχη εβδομάδα, πολλαπλασιασμένη με έναν αριθμό προκειμένου η μέγιστη εκ των 539 τιμών να ισούται με 100.



Διάγραμμα 5.1 - Αναζητήσεις των λέξεων "Olympic Games" παγκοσμίως (Ιανουάριος 2004 – Απρίλιος 2014)

Παρατηρούμε επομένως ότι η μέγιστη τιμή που είναι ίση με 100 εμφανίζεται κατά τη δεύτερη εβδομάδα του Αυγούστου του 2008 (Ολυμπιακοί Αγώνες του Πεκίνου), ενώ για παράδειγμα η αντίστοιχη τιμή κατά την τρίτη εβδομάδα του Αυγούστου του 2004 (Ολυμπιακοί Αγώνες της Αθήνας) είναι ίση με 67. Έτσι λοιπόν, συγκρίνοντας αυτούς τους δύο αριθμούς, μπορούμε να αντιληφθούμε το ποσοστό απόκλισης των αναζητήσεων που πραγματοποιήθηκαν κατά τη διάρκεια των δύο συγκεκριμένων διοργανώσεων. Παράλληλα όπως αναμενόταν το διάγραμμα χαρακτηρίζεται από την έντονη περιοδικότητα του, καθώς παρατηρούνται τρία ισχυρά μέγιστα κατά τη διάρκεια του Αυγούστου των ετών 2004, 2008 και 2012, τα οποία αντιστοιχούν στους Ολυμπιακούς

Αγώνες της Αθήνας, του Πεκίνου και του Λονδίνου αντίστοιχα, καθώς και άλλα τρία λιγότερο ισχυρά μέγιστα κατά τη διάρκεια του Φεβρουαρίου των ετών 2006, 2010 και 2014, τα οποία οφείλονται στους χειμερινούς Ολυμπιακούς Αγώνες που πραγματοποιήθηκαν στο Τορίνο, στο Βανκούβερ και στο Σότσι αντίστοιχα.

Ο σκοπός μας στο συγκεκριμένο κεφάλαιο είναι να ερευνήσουμε αν οι αναζητήσεις μίας συγκεκριμένης αυτοκινητοβιομηχανίας (Α στην περίπτωση μας) στο διαδίκτυο (και συγκεκριμένα στη μηχανή αναζήτησης της Google), συσχετίζονται με τα ύψη των πωλήσεων της εταιρείας.

Με ανάλογο τρόπο όπως στο παραπάνω παράδειγμα πραγματοποιήθηκε η συλλογή των δεδομένων που αφορούν τις αναζητήσεις στο Google της λέξης Α (το όνομα δηλαδή της αυτοκινητοβιομηχανίας) από τον Ιανουάριο του 2006 έως και τον Μάρτιο του 2014. Το Google Trends δίνει τη δυνατότητα στον χρήστη να αποθηκεύσει στον υπολογιστή του τα στοιχεία αυτά σε μορφή CSV, γεγονός που καθιστά εφικτή την έρευνα τους.

Εν συνεχεία μέσω των επίσημων τριμηνιαίων αναφορών του ομίλου V συλλέξαμε τα νούμερα πωλήσεων των αυτοκινήτων Α σε ολόκληρο τον κόσμο, από το πρώτο τρίμηνο του 2006 έως και το πρώτο τρίμηνο του 2014 (σημειώνεται ότι το 99.55% της αυτοκινητοβιομηχανίας Α ανήκει στη V Group από το 1966). Οι αναφορές (Volkswagen Group, 2014) αυτές εκδίδονται κάθε τρεις μήνες, περιέχουν όλα τα οικονομικά στοιχεία της επιχείρησης και είναι διαθέσιμες στο κοινό μέσω της επίσημης σελίδας της εταιρείας στο διαδίκτυο. Στη συνέχεια παρατίθενται τα δεδομένα που συγκεντρώθηκαν.

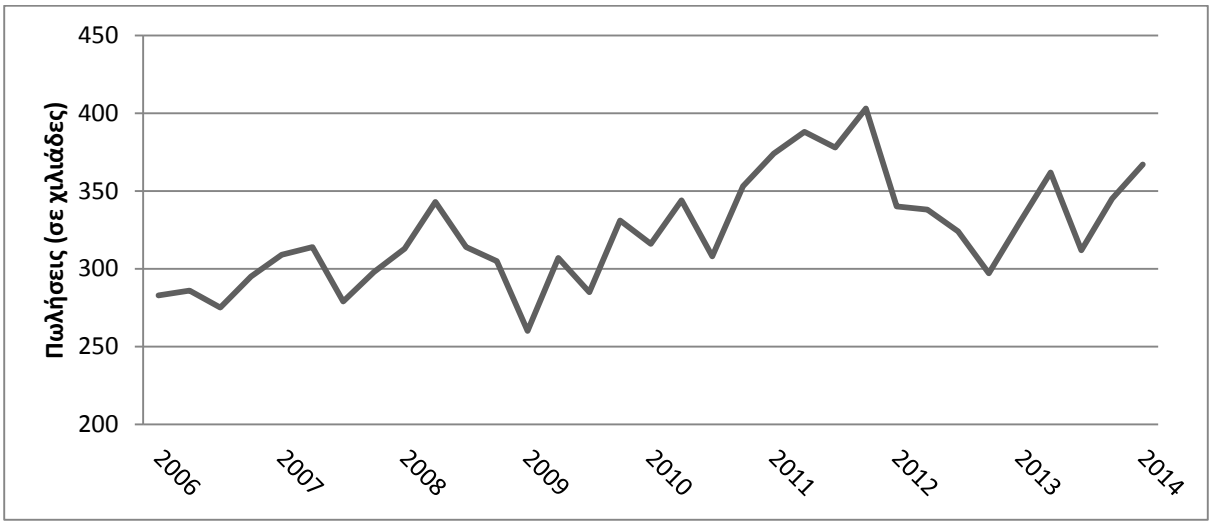
	2006	2007	2008	2009	2010	2011	2012	2013	2014
1ο τρίμηνο	283	309	313	260	316	374	340	330	367
2ο τρίμηνο	286	314	343	307	344	388	338	362	-
3ο τρίμηνο	275	279	314	285	308	378	324	312	-
4ο τρίμηνο	295	298	305	331	353	403	297	345	-

Πίνακας 5.2 - Παγκόσμιες πωλήσεις της Α (σε χιλιάδες)

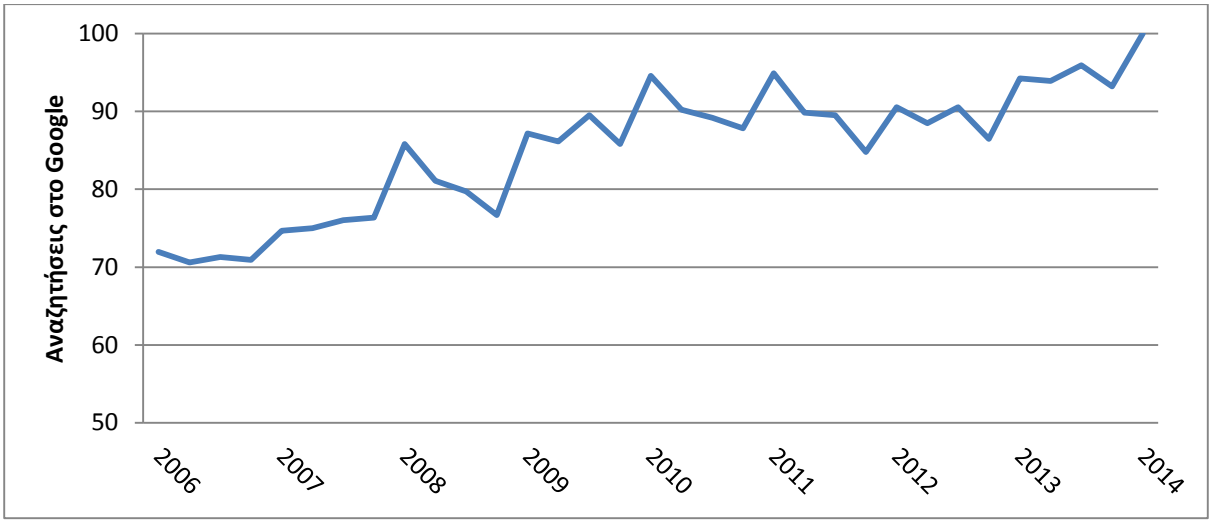
	2006	2007	2008	2009	2010	2011	2012	2013	2014
1ο τρίμηνο	72	75	86	87	95	95	91	94	100
2ο τρίμηνο	71	75	81	86	90	90	89	94	-
3ο τρίμηνο	71	76	80	90	89	90	91	96	-
4ο τρίμηνο	71	76	77	86	88	85	86	93	-

Πίνακας 5.3 - Αναζητήσεις της λέξης “Α” παγκοσμίως

Παρακάτω παρατίθενται οι γραφικές παραστάσεις των πωλήσεων και των αναζητήσεων της αυτοκινητοβιομηχανίας Α στο Google, αντίστοιχα. Οι τιμές που περιέχονται στα διαγράμματα είναι τριμηνιαίες, με εύρος από το πρώτο τρίμηνο του 2006 μέχρι και το πρώτο τρίμηνο του 2014, που σημαίνει ότι συνολικά τα δύο δείγματα μας περιέχουν από 33 τιμές το καθένα, όσα δηλαδή είναι και τα τρίμηνα που περιέχονται μέσα στο χρονικό διάστημα που θα μελετήσουμε.



Διάγραμμα 5.4 - Παγκόσμιες πωλήσεις της Α (Ιανουάριος 2006 – Μάρτιος 2014)

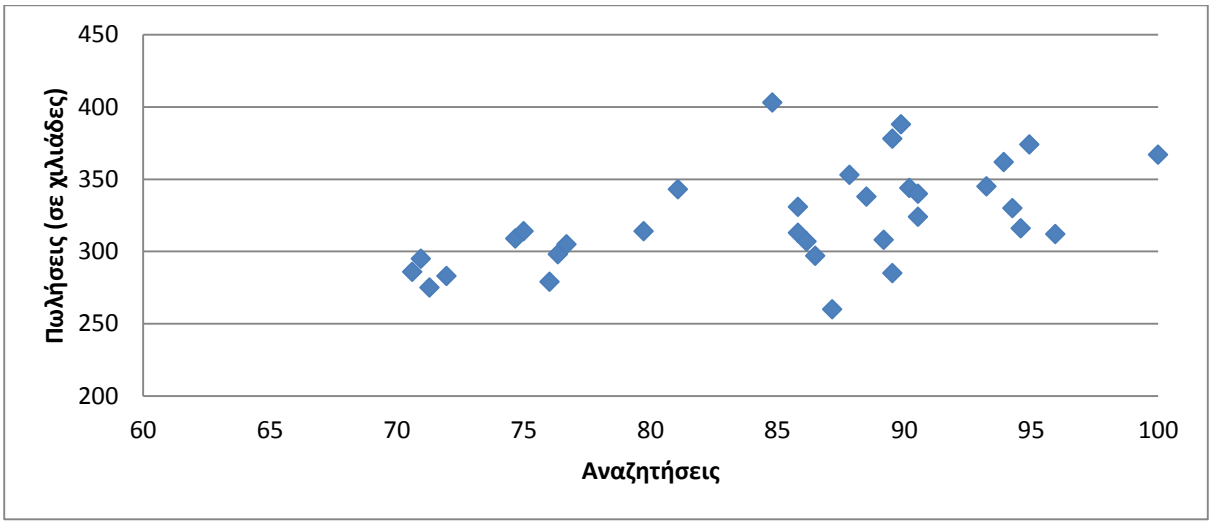


Διάγραμμα 5.5 - Αναζητήσεις της λέξης "Α" παγκοσμίως (Ιανουάριος 2006 – Μάρτιος 2014)

5.3 Συσχέτιση των δύο μεταβλητών

Με βάση τα παραπάνω είναι λογικό, να αναζητήσουμε μέτρα τα οποία να μπορούν να εκφράσουν και να ποσοτικοποιήσουν μία πιθανή συμμεταβολή των δύο χαρακτηριστικών που εξετάζουμε. Το ερώτημα επομένως που άμεσα προκύπτει είναι αν συσχετίζονται οι αναζητήσεις της αυτοκινητοβιομηχανίας στο διαδίκτυο με τις πωλήσεις της. Επίσης αν κάτι τέτοιο ισχύει πόσο ισχυρή είναι η συσχέτιση μεταξύ των δύο μεταβλητών;

Θεωρώντας λοιπόν ως την ανεξάρτητη μεταβλητή μας x τις αναζητήσεις και ως εξαρτημένη y τις πωλήσεις, ένας απλός τρόπος για να αποκτήσουμε μια πρώτη ιδέα για το αν και πώς οι δυο μεταβλητές συσχετίζονται, είναι να κατασκευάσουμε το διάγραμμα διασποράς (*Scatter Diagram*), απεικονίζοντας τα ζεύγη των παρατηρήσεων σε αυτό.



Διάγραμμα 5.6 - Διάγραμμα Διασποράς

Με μία πρώτη ματιά στο διάγραμμα διασποράς φαίνεται ότι οι αναζητήσεις της Α στο διαδίκτυο σχετίζονται, σε κάποιο βαθμό, με τις πωλήσεις της. Υπάρχει δηλαδή μία ανάλογη σχέση μεταξύ των δύο μεταβλητών. Ωστόσο παρατηρούνται και κάποιες ακραίες τιμές στις παρατηρήσεις, οι οποίες μας οδηγούν στο να ερωτόμαστε πόσο ισχυρή είναι αυτή η συσχέτιση. Για παράδειγμα το πρώτο τρίμηνο του 2009 παρατηρήθηκαν οι χαμηλότερες πωλήσεις της εταιρείας (260 χιλιάδες αυτοκίνητα) μέσα στο χρονικό διάστημα που εξετάζουμε, οι οποίες όμως δεν απεικονίστηκαν, με απολύτως ανάλογο τρόπο, και σε μεγάλη πτώση στα ποσοστά αναζητήσεων της επιχείρησης στο Google. Ο κύριος λόγος που συνέβη αυτό είναι η διεθνής οικονομική κρίση εκείνης της χρονικής περιόδου, η οποία όπως είναι λογικό δεν είναι απαραίτητο ότι θα εκφραστεί και στο χώρο του διαδικτύου.

Πόσο ισχυρή είναι λοιπόν αυτή η συσχέτιση; Προκειμένου να απαντηθεί το συγκεκριμένο ερώτημα θα υπολογίσουμε το δειγματικό συντελεστή συσχέτισης του Pearson (*Pearson's correlation coefficient*), ο οποίος εκφράζει το βαθμό της γραμμικής συσχέτισης μεταξύ των δύο μεταβλητών:

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

όπου $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Χρησιμοποιώντας τα δεδομένα μας καταλήγουμε στα παρακάτω αποτελέσματα:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 323.52 \ , \ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 85.25 \ , \ S_{xy} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = 5105.86$$

$$S_{xx} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = 2204.77 \ , \ S_{yy} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = 38692.24$$

Και άρα:

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = 0.553$$

Ενδεικτικά αναφέρεται ότι:

- Αν $r = \pm 1$ υπάρχει **τέλεια γραμμική** συσχέτιση.
- Αν $-1 < r \leq -0{,}8$ ή $0{,}8 \leq r < 1$ υπάρχει **πολύ ισχυρή γραμμική** συσχέτιση.
- Αν $-0.8 < r \leq -0{,}7$ ή $0{,}7 \leq r < 0{,}8$ υπάρχει **ισχυρή γραμμική** συσχέτιση.
- Αν $-0{,}7 < r \leq -0{,}5$ ή $0{,}5 \leq r < 0{,}7$ υπάρχει **μέση γραμμική** συσχέτιση.
- Αν $-0{,}5 < r \leq -0{,}3$ ή $0{,}3 \leq r < 0{,}5$ υπάρχει **ασθενής γραμμική** συσχέτιση.
- Αν $-0{,}3 < r \leq 0{,}3$ **δεν υπάρχει γραμμική** συσχέτιση.

Το συμπέρασμα επομένως που μπορούμε άμεσα να βγάλουμε είναι ότι με βάση το δείγμα μας υπάρχει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών. Παράλληλα όμως πρέπει να υπογραμμιστεί ότι δεν αρκεί να εξετάσουμε μόνο τον δειγματικό συντελεστή συσχέτισης προκειμένου να βγάλουμε ασφαλή συμπεράσματα για τη συμμεταβολή των δύο μεγεθών. Αυτό συμβαίνει διότι οι παρατηρήσεις x_i και y_i αντιπροσωπεύουν μία

συγκεκριμένη χρονική περίοδο με αποτέλεσμα να εμπεριέχεται ο κίνδυνος, η γραμμική συσχέτιση που αποδείχθηκε ότι υπάρχει, να είναι τυχαία. Πραγματοποιούμε επομένως τον παρακάτω έλεγχο, θέτοντας ρ_{xy} τον πληθυσμιακό, πραγματικό συντελεστή συσχέτισης:

$$H_0 : \rho_{xy} = 0 \text{ έναντι της } H_1 : \rho_{xy} \neq 0$$

Υποθέτοντας ότι οι δύο μεταβλητές είναι κανονικά κατανεμημένες πραγματοποιούμε έναν έλεγχο *t* προκείμενου να απορρίψουμε την μηδενική υπόθεση (*null hypothesis*), χρησιμοποιώντας την παρακάτω ελεγχοσυνάρτηση:

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Όπου, $n-2=31$ είναι οι βαθμοί ελευθερίας (*degrees of freedom*) καθώς το δείγμα μας περιέχει 33 παρατηρήσεις. Επίσης τονίζεται ότι η παραπάνω ελεγχοσυνάρτηση υπό την H_0 ακολουθεί την κατανομή t_{n-2}.

Θα εξετάσουμε αν ο δειγματικός συντελεστής συσχέτισης που υπολογίστηκε παραπάνω είναι στατιστικά σημαντικός (Βλ. Παράρτημα Γ):

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0.553\sqrt{31}}{\sqrt{1-0.553^2}} = 3.694$$

$$(p\text{-}\tau\iota\mu\acute{\eta} < 0.001)$$

Άρα η μηδενική υπόθεση απορρίπτεται και συμπεραίνουμε ότι η συσχέτιση μεταξύ των δύο μεταβλητών είναι στατιστικά σημαντική.

Κεφάλαιο 6. Πρόβλεψη της οικονομικής δραστηριότητας μέσω του Google

6.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο αποδείξαμε ότι οι αναζητήσεις της εταιρείας Α στο Google, συσχετίζονται με τις πωλήσεις της, γεγονός που μας οδηγεί στο να ερωτόμαστε αν μπορούμε μέσω αυτών των αριθμών, που διατίθενται από τη δημοφιλή τη μηχανή αναζήτησης, να πραγματοποιούμε προβλέψεις πωλήσεων. Κάτι τέτοιο δεν ακούγεται παράλογο, καθώς στις μέρες μας μία από τις πρώτες ενέργειες, αν όχι η πρώτη, που ένας πιθανός αγοραστής θα πραγματοποιήσει προτού αγοράσει ένα αυτοκίνητο, είναι να ψάξει και να μελετήσει στο διαδίκτυο εικόνες, βίντεο , άρθρα, κριτικές και άλλα πολλά, που σχετίζονται με το αυτοκίνητο αυτό.

Όπως έχουμε ήδη τονίσει αν η απάντηση στο παραπάνω ερώτημα αποδειχθεί καταφατική, με αποτέλεσμα να καταφέρουμε να κατασκευάσουμε ένα μοντέλο πρόβλεψης πωλήσεων, τότε οι προβλέψεις αυτές θα αφορούν το άμεσο μέλλον και όχι οποιαδήποτε στιγμή. Αυτό συμβαίνει διότι οι πωλήσεις μιας εταιρείας εξαρτώνται από πολύ ισχυρούς παράγοντες, οι οποίοι ανά πάσα στιγμή είναι ικανοί να μεταβληθούν σε πολύ μεγάλο βαθμό, όπως για παράδειγμα μία ενδεχόμενη οικονομική κρίση, η οποία προφανώς θα έχει σαν αποτέλεσμα την πτώση των πωλήσεων αυτοκινήτων.

Αν θεωρήσουμε λοιπόν ότι έχουμε στα χέρια μας τις μηνιαίες πωλήσεις μιας αυτοκινητοβιομηχανίας, από τον Ιανουάριο του 2004 έως τον Απρίλιο του 2014, τότε οι προβλέψεις μας θα αφορούν λίγους μήνες εκτός του χρονικού μας εύρους. Θα είμαστε σε θέση λοιπόν να προβλέψουμε με σχετική ασφάλεια τις πωλήσεις του Μαΐου του 2014, αλλά όσο περισσότερο απομακρυνόμαστε χρονικά, τόσο πιο επικίνδυνο θα είναι να υποπέσουμε σε σφάλμα. Επίσης ένας ακόμα ασφαλέστερος τρόπος να αξιοποιήσουμε ένα τέτοιο μοντέλο, είναι να πραγματοποιήσουμε μία πρόβλεψη για το παρόν χρησιμοποιώντας, αντί για τις μηνιαίες, τις εβδομαδιαίες[a] αναζητήσεις της εταιρείας στο Google. Έτσι λοιπόν αξιοποιώντας και τις αναζητήσεις που

[a] Εκτός από τα δεδομένα που αφορούν τις μηνιαίες αναζητήσεις ενός όρου, το Google Trends μας δίνει επίσης πρόσβαση και στα αντίστοιχα εβδομαδιαία.

πραγματοποιήθηκαν κατά την πρώτη εβδομάδα του Μαΐου, είμαστε σε θέση να προβλέψουμε με ακόμα μεγαλύτερη ασφάλεια τις συνολικές πωλήσεις του Μαΐου. Έτσι, αυτό που ισχυριζόμαστε είναι ότι ο όγκος των αναζητήσεων στο Google μιας συγκεκριμένης μάρκας αυτοκινήτων μέσα σε κάποιο χρονικό διάστημα, ίσως μας βοηθήσει να προβλέψουμε τις πωλήσεις της αντίστοιχης εταιρείας για το προσεχές μέλλον, ή το παρόν.

Συνοπτικά λοιπόν ο στόχος μας είναι να δείξουμε ότι, ομοίως με το προηγούμενο κεφάλαιο, οι αναζητήσεις της αυτοκινητοβιομηχανίας Η στο Google συμμεταβάλλονται με τις πωλήσεις της εταιρείας. Στη συνέχεια θα προχωρήσουμε ένα βήμα παραπέρα κατασκευάζοντας ένα μοντέλο, το οποίο θα είναι ικανό να πραγματοποιεί μελλοντικές προβλέψεις των πωλήσεων της εταιρείας, χρησιμοποιώντας τους αριθμούς που λαμβάνουμε μέσω της υπηρεσίας Google Trends.

6.2 Συλλογή δεδομένων της αυτοκινητοβιομηχανίας Η

Με τρόπο όμοιο σε σχέση με αυτόν του προηγούμενου κεφαλαίου, πραγματοποιήθηκε η συλλογή των δεδομένων που αφορούν τις αναζητήσεις στο Google της λέξης Η (του ονόματος δηλαδή της αυτοκινητοβιομηχανίας) από τον Ιανουάριο του 2004 έως και τον Απρίλιο του 2014. Όπως έχει αναφερθεί, το Google Trends δίνει τη δυνατότητα στον χρήστη να αποθηκεύσει στον υπολογιστή του τα στοιχεία αυτά σε μορφή CSV, γεγονός που καθιστά εφικτή την έρευνα τους. Στο συγκεκριμένο κεφάλαιο συγκεντρώθηκαν τα δεδομένα που αφορούν τις αναζητήσεις της Η στο Google, μόνο στις Ηνωμένες Πολιτείες Αμερικής.

Εν συνεχεία μέσω των επίσημων μηνιαίων αναφορών της Η Motor America, συγκεντρώθηκαν τα νούμερα πωλήσεων των αυτοκινήτων Η στις ΗΠΑ, από τον Ιανουάριο του 2004 έως και τον Απρίλιο του 2014. Οι αναφορές (Hyundai News, 2014) αυτές, οι οποίες είναι διαθέσιμες στο κοινό μέσω της επίσημης σελίδας της εταιρείας στο διαδίκτυο, εκδίδονται κάθε μήνα από το 1986 και μετά και περιέχουν όλα τα οικονομικά στοιχεία της επιχείρησης.

Το γεγονός ότι οι αριθμοί των πωλήσεων της Η βρίσκονται σε μηνιαία βάση (και όχι σε τριμηνιαία όπως αυτοί της Α) μας βοηθάει ιδιαίτερα, καθώς το μοντέλο που θα κατασκευάσουμε θα είναι ακριβέστερο, λόγω του αυξημένου αριθμού παρατηρήσεων που διαθέτουμε. Παράλληλα θα είμαστε σε θέση να εντοπίσουμε και να αντιμετωπίσουμε ευκολότερα κάποια φαινόμενα, όπως η εποχικότητα των πωλήσεων, τα οποία στη περίπτωση που δεν διαθέτουμε μηνιαία στοιχεία τείνουν να αμβλύνονται και να μην προσδιορίζονται σωστά. Για παράδειγμα το γεγονός ότι οι πωλήσεις κατά τον μήνα Δεκέμβριο αυξάνονται λόγω των επερχόμενων Χριστουγεννιάτικων εορτών, δεν φαίνεται σε μεγάλο βαθμό σε τριμηνιαίες αναφορές, ενώ αντίθετα στις μηνιαίες η συγκεκριμένη πληροφορία είναι αρκετά ευδιάκριτη.

Τέλος το γεγονός ότι πλέον δεν θα εξετάσουμε τις παγκόσμιες πωλήσεις (όπως στην περίπτωση της Α, στο προηγούμενο κεφάλαιο), αλλά θα εστιάσουμε στην αγορά των ΗΠΑ, απλουστεύει κάπως την κατασκευή του μοντέλου μας. Αυτό συμβαίνει διότι η οικονομία και η πορεία των επιχειρήσεων μίας δεδομένης χώρας εξαρτάται από δεκάδες παράγοντες, με αποτέλεσμα αν πραγματοποιήσουμε μία μελέτη η οποία αφορά ολόκληρο τον κόσμο, οι παράγοντες αυτοί καθώς και οι αλληλεπιδράσεις τους, για κάθε μία από τις εκατοντάδες χώρες, να δημιουργούν ένα χαοτικό σύμπλεγμα, το οποίο είναι δύσκολο να μοντελοποιηθεί και να ερμηνευτεί. Για παράδειγμα αν στην Κίνα μία αυτοκινητοβιομηχανία ανακοινώσει για κάποιο χρονικό διάστημα

μία προσφορά, γεγονός που θα αυξήσεις τις εγχώριες πωλήσεις, δεν συνεπάγεται ότι θα αυξηθούν και οι πωλήσεις, μέσα στο ίδιο χρονικό διάστημα, της εταιρείας και στην Ευρώπη. Έτσι λοιπόν δεν είναι δύσκολο να καταλάβουμε ότι αν περιορίσουμε τη μελέτη μας σε μία συγκεκριμένη χώρα, είμαστε σε θέση να κατανοήσουμε με μεγαλύτερη ευχέρεια τους διάφορους παράγοντες που είναι ικανοί να επηρεάσουν το μοντέλο μας.

Στη συνέχεια παρατίθενται τα δεδομένα που συγκεντρώθηκαν, δηλαδή οι πωλήσεις και οι αναζητήσεις της αυτοκινητοβιομηχανίας Η στο Google, αντίστοιχα. Συνολικά τα δύο δείγματά μας περιέχουν από 124 τιμές το καθένα, όσοι δηλαδή είναι και οι μήνες μέσα στο χρονικό διάστημα που θα μελετήσουμε.

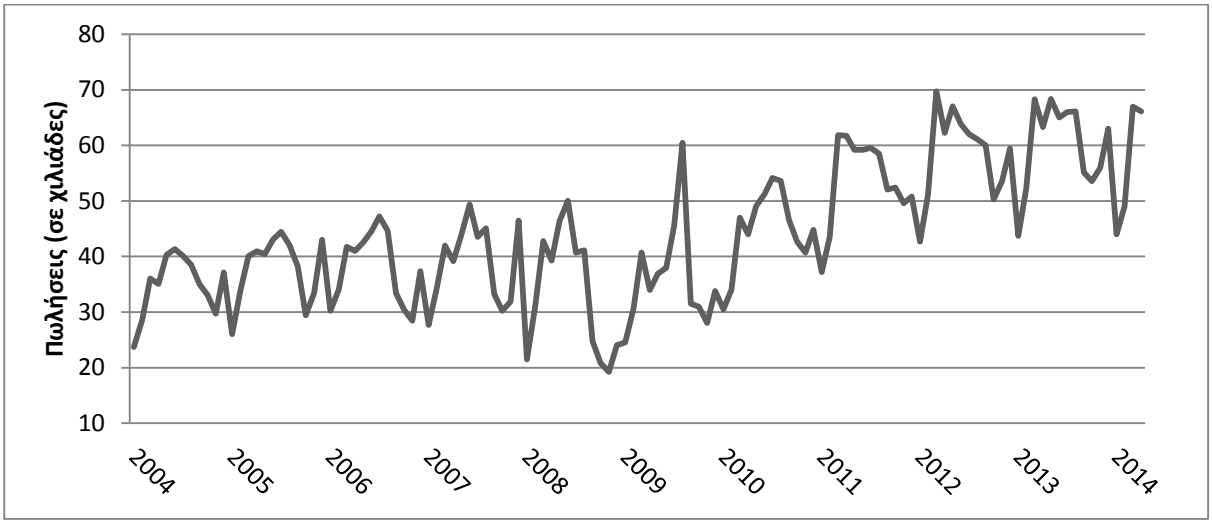
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Ιανουάριος	23738	26009	30208	27721	21452	24512	30503	37214	42694	43713	44005
Φεβρουάριος	28531	33940	34014	34500	31090	30621	34004	43533	51151	52311	49003
Μάρτιος	36024	40063	41766	41984	42796	40721	47002	61873	69728	68306	67005
Απρίλιος	35045	40958	41025	39137	39280	33952	44023	61754	62264	63315	66107
Μάιος	40337	40416	42514	43885	46415	36937	49045	59214	67019	68358	-
Ιούνιος	41336	43051	44508	49368	50033	37943	51205	59209	63813	65007	-
Ιούλιος	40125	44431	47205	43511	40703	45553	54106	59561	62021	66005	-
Αύγουστος	38513	42113	44635	45087	41130	60467	53603	58505	61099	66101	-
Σεπτέμβριος	35025	38214	33384	33214	24765	31511	46556	52051	60025	55102	-
Οκτώβριος	33111	29413	30479	30232	20820	31005	42656	52402	50271	53555	-
Νοέμβριος	29682	33383	28417	31883	19221	28045	40723	49610	53487	56005	-
Δεκέμβριος	37148	43021	37365	46487	24037	33797	44802	50765	59435	63005	-

Πίνακας 6.1 - Πωλήσεις της Η στις ΗΠΑ (Ιανουάριος 2004 – Απρίλιος 2014)

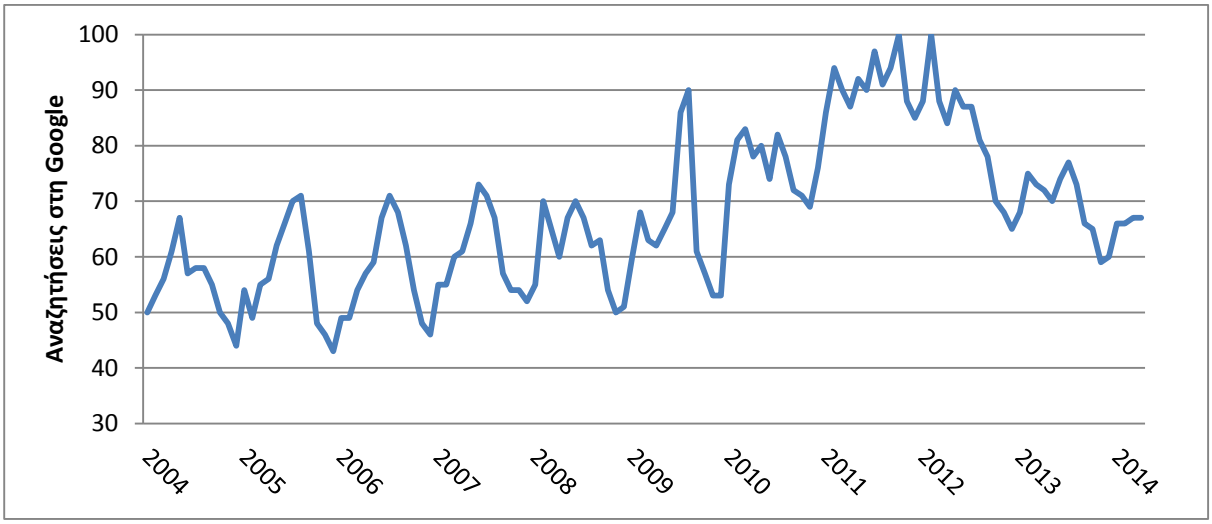
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Ιανουάριος	50	54	49	55	55	60	73	86	88	68	66
Φεβρουάριος	53	49	49	55	70	68	81	94	100	75	66
Μάρτιος	56	55	54	60	65	63	83	90	88	73	67
Απρίλιος	61	56	57	61	60	62	78	87	84	72	67
Μάιος	67	62	59	66	67	65	80	92	90	70	-
Ιούνιος	57	66	67	73	70	68	74	90	87	74	-
Ιούλιος	58	70	71	71	67	86	82	97	87	77	-
Αύγουστος	58	71	68	67	62	90	78	91	81	73	-
Σεπτέμβριος	55	61	62	57	63	61	72	94	78	66	-
Οκτώβριος	50	48	54	54	54	57	71	100	70	65	-
Νοέμβριος	48	46	48	54	50	53	69	88	68	59	-
Δεκέμβριος	44	43	46	52	51	53	76	85	65	60	-

Πίνακας 6.2 - Αναζητήσεις της λέξης “ Η” στις ΗΠΑ (Ιανουάριος 2004 – Απρίλιος 2014)

Παρακάτω παρατίθενται οι γραφικές παραστάσεις των πωλήσεων και των αναζητήσεων της αυτοκινητοβιομηχανίας Η στο Google, συναρτήσει του χρόνου. Σημειώνεται ότι οι τιμές που περιέχονται στα διαγράμματα εστιάζονται στις ΗΠΑ και είναι μηνιαίες, με εύρος από τον Ιανουάριο του 2004 έως και τον Απρίλιο του 2014 (Βλ. Πίνακες 6.1 και 6.2 αντίστοιχα).



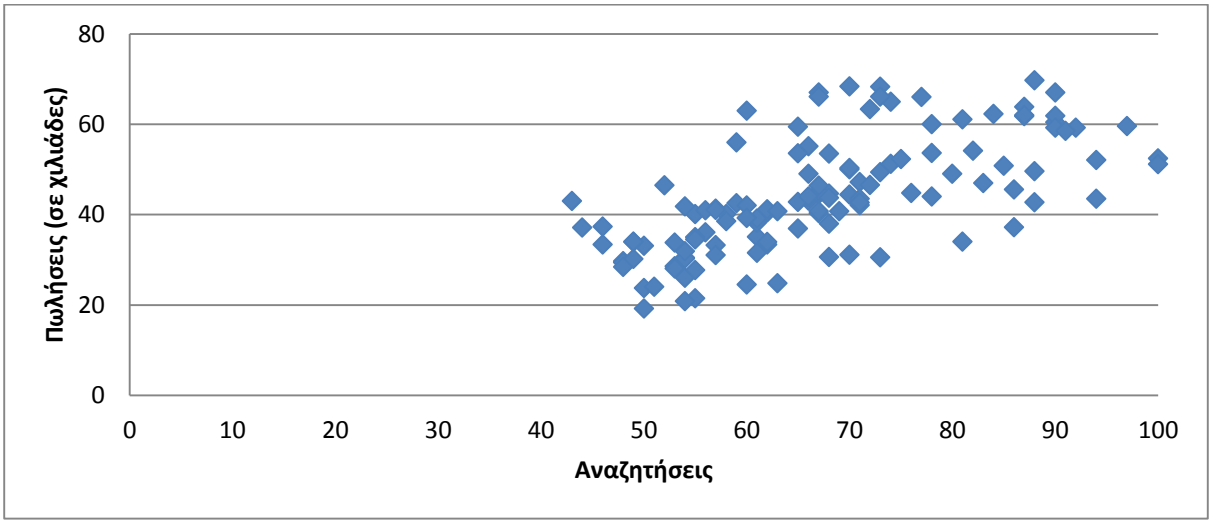
Διάγραμμα 6.3 - Πωλήσεις της Η στις ΗΠΑ (Ιανουάριος 2004 – Απρίλιος 2014)



Διάγραμμα 6.4 - Αναζητήσεις της λέξης " Η" στις ΗΠΑ (Ιανουάριος 2004 – Απρίλιος 2014)

6.3 Συσχέτιση των δύο μεταβλητών

Θεωρώντας ως την ανεξάρτητη μεταβλητή μας x τις αναζητήσεις και ως εξαρτημένη y τις πωλήσεις αυτοκινήτων της Η στις ΗΠΑ, θα κατασκευάσουμε το διάγραμμα διασποράς (*Scatter Diagram*), απεικονίζοντας τα ζεύγη των παρατηρήσεων σε αυτό.



Διάγραμμα 6.5 - Διάγραμμα Διασποράς

Στο παραπάνω διάγραμμα διασποράς φαίνεται ότι οι αναζητήσεις της Η στο Google σχετίζονται, σε κάποιο σημαντικό βαθμό, με τις πωλήσεις που η εταιρεία πραγματοποίησε. Επειδή ωστόσο παρατηρούνται και κάποιες ακραίες τιμές στις παρατηρήσεις, προκειμένου να προσδιορίσουμε το πόσο ισχυρή είναι αυτή η συσχέτιση μεταξύ των δύο μεταβλητών, θα υπολογίσουμε το δειγματικό συντελεστή συσχέτισης του Pearson:

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

όπου $S_{xy} = \sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})$, $S_{xx} = \sum_{t=1}^{n} (x_t - \bar{x})^2$, $S_{yy} = \sum_{t=1}^{n} (y_t - \bar{y})^2$

και $\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t$, $\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$

Χρησιμοποιώντας τα δεδομένα μας καταλήγουμε στο παρακάτω αποτέλεσμα:

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = 0.659$$

Το συμπέρασμα επομένως που μπορούμε άμεσα να βγάλουμε είναι ότι με βάση το δείγμα μας υπάρχει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών. Παράλληλα όμως πρέπει να υπογραμμιστεί ότι δεν αρκεί να εξετάσουμε μόνο τον δειγματικό συντελεστή συσχέτισης προκειμένου να βγάλουμε ασφαλή συμπεράσματα για τη συμμεταβολή των δύο μεγεθών. Αυτό συμβαίνει διότι οι παρατηρήσεις x_i και y_i αντιπροσωπεύουν μία συγκεκριμένη χρονική περίοδο με αποτέλεσμα να εμπεριέχεται ο κίνδυνος, η γραμμική συσχέτιση που αποδείχθηκε ότι υπάρχει, να είναι τυχαία. Πραγματοποιούμε επομένως τον παρακάτω έλεγχο, θέτοντας ρ_{xy} τον πληθυσμιακό, πραγματικό συντελεστή συσχέτισης:

$$H_0 : \rho_{xy} = 0 \text{ έναντι της } H_1 : \rho_{xy} \neq 0$$

Υποθέτοντας ότι οι δύο μεταβλητές είναι κανονικά κατανεμημένες πραγματοποιούμε έναν έλεγχο t προκείμενου να απορρίψουμε την μηδενική υπόθεση (*null hypothesis*), χρησιμοποιώντας την παρακάτω ελεγχοσυνάρτηση:

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Όπου, $n–2=122$ είναι οι βαθμοί ελευθερίας (*degrees of freedom*) καθώς το δείγμα μας περιέχει 124 παρατηρήσεις. Επίσης τονίζεται ότι η παραπάνω ελεγχοσυνάρτηση υπό την H_0 ακολουθεί την κατανομή t_{n-2}.

Θα εξετάσουμε αν ο δειγματικός συντελεστής συσχέτισης που υπολογίστηκε παραπάνω είναι στατιστικά σημαντικός (Βλ. Παράρτημα Γ):

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0.659\sqrt{122}}{\sqrt{1-0.659^2}} = 9.689$$

$$(p\text{-}\tau\iota\mu\acute{\eta} < 0.0001)$$

Άρα η μηδενική υπόθεση απορρίπτεται και συμπεραίνουμε ότι η συσχέτιση μεταξύ των δύο μεταβλητών είναι στατιστικά σημαντική.

6.4 Απλό γραμμικό μοντέλο

Στη συνέχεια θα προσαρμόσουνε ένα απλό γραμμικό μοντέλο στα δεδομένα των πινάκων 6.1 και 6.2 και ειδικότερα θεωρώντας ως ανεξάρτητη μεταβλητή x τις αναζητήσεις της Η στο Google (και συγκεκριμένα στις ΗΠΑ) και ως εξαρτημένη y τις πωλήσεις των αυτοκινήτων της στις ΗΠΑ, θα έχουμε:

$$y_t = b_0 + b_1 x_t + \varepsilon_t$$

Το μοντέλο προσαρμόζεται σε: $\hat{y}_t = \hat{b}_0 + \hat{b}_1 x_t$

Όπου τα ε_t είναι (Οικονόμου & Καρώνη, 2010) τα τυχαία σφάλματα και e_t τα υπόλοιπα ή αλλιώς κατάλοιπα (*residuals*) τα οποία είναι ίσα με:

$$e_t = y_t - \hat{y}_t = y_t - \hat{b}_0 - \hat{b}_1 x_t$$

Στη συνέχεια θα εκτιμηθούν οι άγνωστες πληθυσμιακές παράμετροι b_0 και b_1 με τη βοήθεια των εκτιμητριών[a], που θα προκύψουν από την παλινδρόμηση. Η εκτίμηση των παραμέτρων b_0 και b_1 του υποδείγματος θα πραγματοποιηθεί με τη μέθοδο ελαχίστων τετραγώνων (*least squares method*).

Η ανάλυση των δεδομένων και η εκτίμηση της ανάλυσης παλινδρόμησης πραγματοποιήθηκε με χρήση του στατιστικού πακέτου R και φαίνεται στο Παράρτημα Γ. Όπως θα δούμε στη συνέχεια ο έλεγχος της ερμηνευτικής ικανότητας του μοντέλου διακρίνεται σε δύο επίπεδα: Στον έλεγχο της στατιστικής σημαντικότητας των επιμέρους παραμέτρων (στη αυτήν την περίπτωση b_0 και b_1), καθώς και στον έλεγχο της ερμηνευτικής ικανότητας του μοντέλου συνολικά (μέσω του συντελεστή προσδιορισμού R^2).

Παράλληλα θα εξετάσουμε αν παραβιάζεται κάποια από τις προϋποθέσεις (με εξαίρεση τις δύο τελευταίες) οι οποίες αναφέρθηκαν παραπάνω και πρέπει να ισχύουν ώστε το μοντέλο να είναι κατάλληλο, προκειμένου να βγάλουμε συμπεράσματα σχετικά με τη συμπεριφορά των παρατηρήσεων. Τονίζεται ότι, επειδή βρισκόμαστε στην περίπτωση του απλού γραμμικού μοντέλου (έχουμε δηλαδή μία ανεξάρτητη μεταβλητή), η εξέταση του πέμπτου ισχυρισμού είναι περιττή, καθώς συμπίπτει με αυτή του δεύτερου[b], ενώ όσον αφορά τον έκτο, είναι προφανώς περιττό να εξετάσουμε αν υπάρχει πολυσυγγραμμικότητα, αφού το μοντέλο περιλαμβάνει μόνο μία ανεξάρτητη μεταβλητή.

---

[a] *Δηλαδή των* $\hat{b}_0$ και $\hat{b}_1$

[b] Καθώς τα εκτιμημένα $\hat{y}_t$ συνδέονται γραμμικά με τα x_t

6.4.1 Έλεγχος της στατιστικής σημαντικότητας των επιμέρους παραμέτρων

- Έλεγχος υπόθεσης *t* για το συντελεστή b_1:

Οι υποθέσεις για τον συντελεστή b_1 είναι:

$$H_0 : b_1 = 0 \text{ έναντι της } H_1 : b_1 \neq 0$$

Πραγματοποιούμε έναν έλεγχο *t* προκείμενου να απορρίψουμε την μηδενική υπόθεση (*null hypothesis*), χρησιμοποιώντας την παρακάτω ελεγχοσυνάρτηση (Οικονόμου & Καρώνη, 2010), που ακολουθεί την κατανομή t_{n-2}:

$$t = \frac{\hat{b}_1}{\sqrt{Var(\hat{b}_1)}} \sim t_{n-2}$$

όπου $Var(\hat{b}_1)$ είναι η εκτιμημένη διασπορά του $\hat{b}_1$ και ισούται με:

$$Var(\hat{b}_1) = \frac{S^2}{\sum_{t=1}^{n}\left(x_t - \bar{x}\right)^2} \text{ , για } S^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)$$

Όπως φαίνεται στα αποτελέσματα του Παραρτήματος Γ έχουμε

$$t = \frac{\hat{b}_1}{\sqrt{Var(\hat{b}_1)}} = 9.689$$

$$(p\text{-τιμή} < 0.0001)$$

Άρα συμπεραίνουμε ότι ο συντελεστής της μεταβλητής που αναπαριστά τις αναζητήσεις της Η στο Google (και συγκεκριμένα στις ΗΠΑ) είναι στατιστικά σημαντικός.

- Σταθερός όρος b_0:

Όπως έχει ήδη αναφερθεί μία ενδεχόμενη αφαίρεση του σταθερού όρου b_0 είναι εξαιρετικά σπάνια και οφείλεται στη φύση του φαινομένου που σκοπεύουμε να μοντελοποιήσουμε και όχι σε μία ενδεχόμενη δεκτή μηδενική υπόθεση του έλεγχου *t*. Έτσι λοιπόν δεχόμαστε ότι ο σταθερός όρος δεν θα παραληφθεί από το μοντέλο μας.

6.4.2 Έλεγχος της ερμηνευτικής ικανότητας του μοντέλου

Από τη στιγμή που έχουμε μόνο μία ανεξάρτητη μεταβλητή δεν χρειάζεται να ελέγξουμε τη σημαντικότητα της παλινδρόμησης χρησιμοποιώντας τη στατιστική συνάρτηση F, καθώς δεν θα μας δώσει κάποια επιπλέον πληροφορία σε σχέση με τον έλεγχο t που πραγματοποιήθηκε προηγουμένως για την παράμετρο b_1 (αποδεικνύεται (Οικονόμου & Καρώνη, 2010) ότι στην περίπτωση του απλού γραμμικού μοντέλου ο στατιστικός έλεγχος F και ο στατιστικός έλεγχος t είναι ισοδύναμοι).

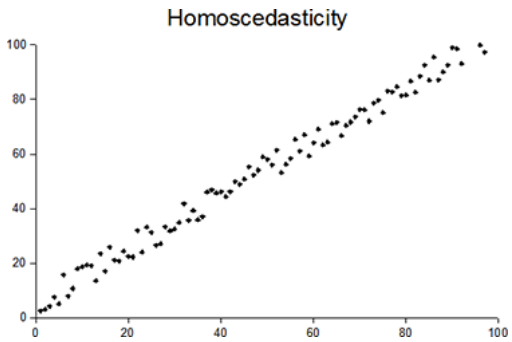
- Συντελεστής προσδιορισμού R^2:

Όπως φαίνεται από τα αποτελέσματα του Παραρτήματος Γ, στην περίπτωση μας περίπου το 43% της μεταβλητότητας των πωλήσεων αυτοκινήτων της Η στις ΗΠΑ, εξηγείται από τις αναζητήσεις, που αφορούσαν την εταιρεία, στο Google (και συγκεκριμένα στις ΗΠΑ). Ωστόσο το συγκεκριμένο ποσοστό δεν κρίνεται ιδιαίτερα ικανοποιητικό[a] προκειμένου να πραγματοποιήσουμε μελλοντικές προβλέψεις και όπως θα δούμε παρακάτω μπορεί να βελτιωθεί σε πολύ μεγάλο βαθμό.

[a] Γενικά ένα μοντέλο με συντελεστή προσδιορισμού μεγαλύτερο από 60% θεωρείται ικανοποιητικό.

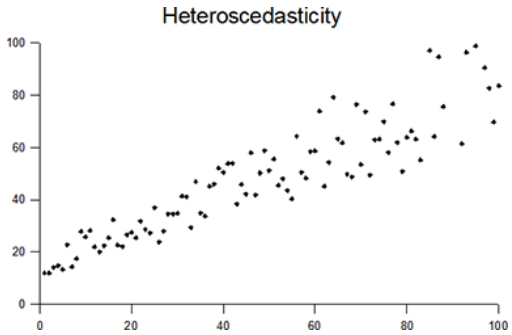
6.4.3 Έλεγχος για ύπαρξη ετεροσκεδαστικότητας

Όπως προαναφέρθηκε, μία εκ των υποθέσεων, προκειμένου να εφαρμόσουμε τη μέθοδο ελαχίστων τετραγώνων, είναι η σταθερή διακύμανση των υπολοίπων e_t (*residuals*). Αυτό ισχύει διότι υποθέτουμε ότι τα e_t , προκύπτουν από πανομοιότυπες κατανομές. Σε περίπτωση που εμφανιστεί το φαινόμενο της ετεροσκεδαστικότητας προκαλείται παραβίαση της συγκεκριμένης υπόθεσης.

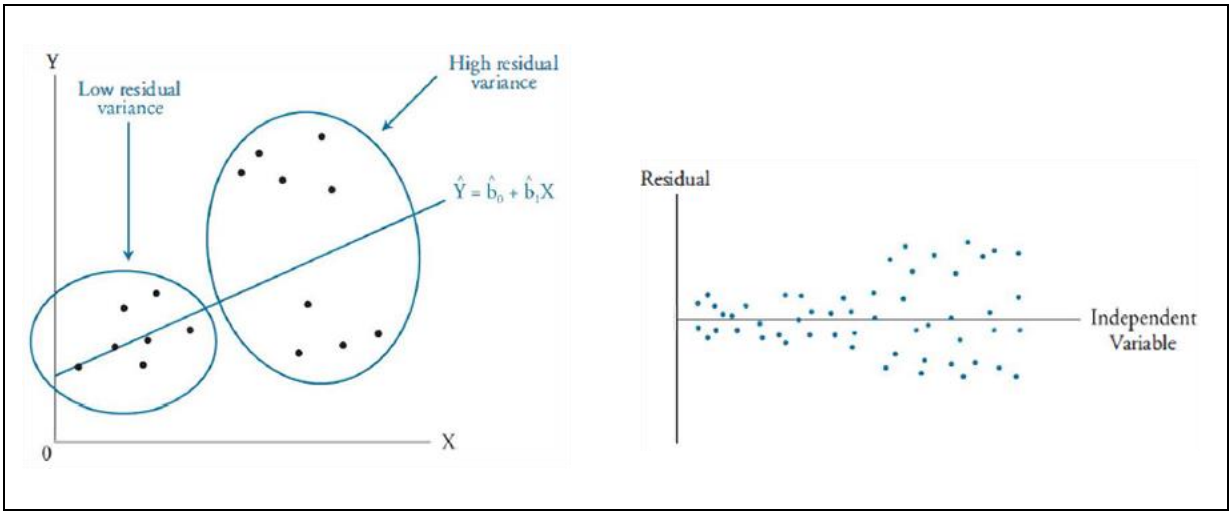


Διάγραμμα 6.6

Για παράδειγμα η ετεροσκεδαστικότητα εμφανίζεται στα μοντέλα στα οποία όσο αυξάνεται η τιμή του x αυξάνεται αντίστοιχα και η μεταβλητότητα των τιμών του y (Βλ. διάγραμμα 6.7 σε αντίθεση με το 6.6). Αυτό συμβαίνει διότι η διασπορά των υπολοίπων e_t αυξάνεται συνεχώς, καθώς μεγαλώνει η τιμή της ανεξάρτητης τιμής x. Το φαινόμενο αυτό σε ένα διάγραμμα διασποράς διακρίνεται όταν οι παρατηρήσεις σχηματίζουν ένα κωνικό σχήμα, αντί να κατανέμονται ισόμορφα.



Διάγραμμα 6.7

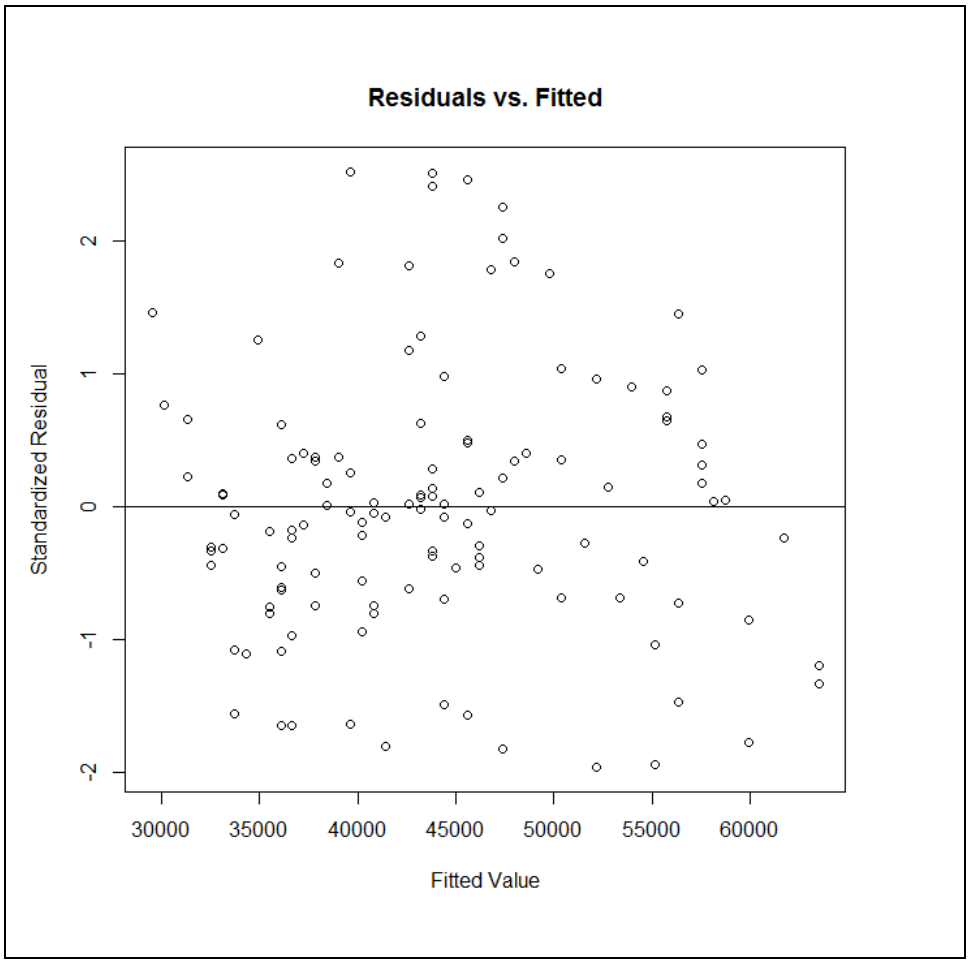


Εικόνα 6.8 – Εμφάνιση Ετεροσκεδαστικότητας (SchweserNotes, 2013)

Το πρόβλημα της ετεροσκεδαστικότητας εμφανίζεται στη περίπτωση που υπάρχει παραβίαση της υπόθεσης της σταθερής διακύμανσης των τιμών των υπολοίπων e_t. Εμπειρικά, δηλαδή μέσω γραφικού ελέγχου, μπορούμε να

εντοπίσουμε την ετεροσκεδαστικότητα, μέσω του διαγράμματος διασποράς των υπολοίπων e_t συναρτήσει των εκτιμημένων $\hat{y}_t$.

Αν στο συγκεκριμένο διάγραμμα διασποράς παρατηρηθεί ένας συστηματικός τρόπος συμπεριφοράς των σημείων, τότε είναι εξαιρετικά πιθανό να υπάρχει το πρόβλημα της ετεροσκεδαστικότητας. Η εξάρτηση αυτή μας οδηγεί στο να συμπεράνουμε ότι τα υπόλοιπα εμπεριέχουν πληροφορίες, οι οποίες δεν έχουν ληφθεί υπόψη κατά τη κατασκευή του μοντέλου. Αν αντίθετα τα σημεία είναι τυχαία διεσπαρμένα στο επίπεδο, τότε δεν υπάρχουν ενδείξεις ετεροσκεδαστικότητας και επομένως δεχόμαστε ότι υπάρχει σταθερότητα της διασποράς των τυχαίων υπολοίπων, δηλαδή ομοσκεδαστικότητα. Στη συνέχεια παρουσιάζεται το διάγραμμα διασποράς των τυποποιημένων (*Standardized*) υπολοίπων συναρτήσει των εκτιμημένων $\hat{y}_t$ του μοντέλου μας.

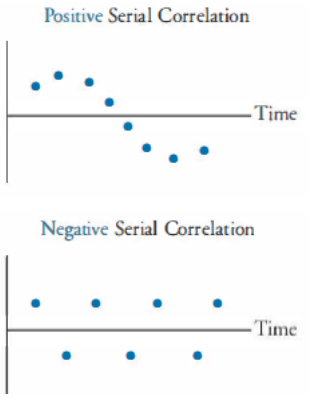


Διάγραμμα 6.9 – Υπόλοιπα σε σχέση με τις τιμές του x

Στο παραπάνω διάγραμμα φαίνεται ότι η διασπορά των υπολοίπων είναι σταθερή, καθώς κατανέμονται τυχαία πάνω και κάτω από την ευθεία που διέρχεται από το μηδέν. Επομένως μπορούμε να συμπεράνουμε ότι δεν υπάρχει ένδειξη ετεροσκεδαστικότητας, καθώς δεν υπάρχει κάποια μορφή εξάρτησης ανάμεσα στα υπόλοιπα e_t και τα εκτιμημένα $\hat{y}_t$.

6.4.4 Έλεγχος για ύπαρξη αυτοσυσχέτισης

Στην παλινδρόμηση η αυτοσυσχέτιση (σειριακή συσχέτιση) εμφανίζεται όταν τα σφάλματα παρατήρησης δεν είναι κατά ζεύγη ασυσχέτιστα και επομένως παραβιάζεται η υπόθεση περί ανεξαρτησίας των τιμών του διαταρακτικού όρου του μοντέλου. Η αυτοσυσχέτιση επομένως εμφανίζεται (SchweserNotes, 2013) όταν κάποιες ή όλες εκ των διαδοχικών τιμών του σφάλματος συσχετίζονται μεταξύ τους (Βλ. Εικόνα 6.10), όταν δηλαδή η συνδιακύμανσή τους είναι διάφορη του μηδενός:



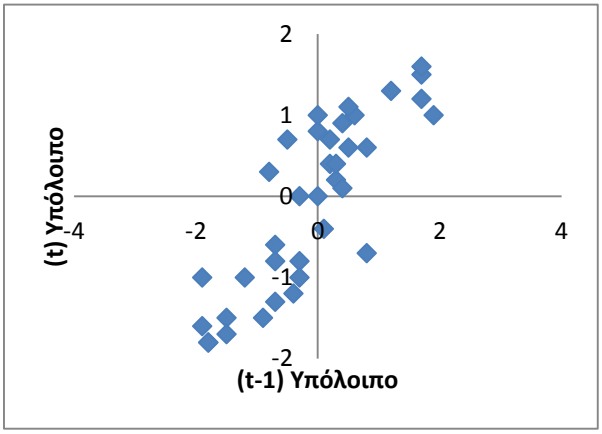
Εικόνα 6.10
Αυτοσυσχέτιση
(SchweserNotes, 2013)

$$\operatorname{cov}(\varepsilon_t, \varepsilon_{t-s}) = E(\varepsilon_t, \varepsilon_{t-s}) \neq 0 \text{ , για } s \neq 0$$

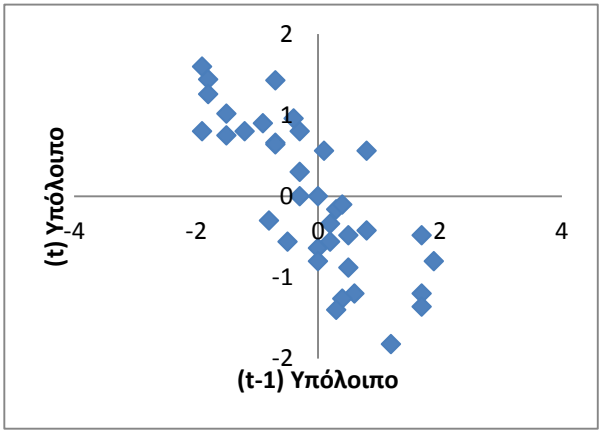
Κάποιοι λόγοι εξαιτίας των οποίων μπορεί να προκύψει αυτοσυσχέτιση είναι οι εξής:

- Η παράλειψη σημαντικών ανεξάρτητων μεταβλητών.
- Η λανθασμένη αλγεβρική εξειδίκευση.
  Για παράδειγμα η y_t σχετίζεται με την x_t μέσω μίας τετραγωνικής σχέσης:
  $$y_t = b_0 + b_1 x_t^2 + \varepsilon_t$$
  ενώ λανθασμένα υποθέσαμε ότι συνδέονται γραμμικά και εκτιμήσαμε μία ευθεία γραμμή:
  $$y_t = b_0 + b_1 x_t + \varepsilon_t$$
- Τα συστηματικά σφάλματα κατά τη διάρκεια των μετρήσεων.

Η συσχέτιση μεταξύ των υπολοίπων, τα οποία απέχουν μεταξύ τους ένα, δύο, τρία κ.ο.κ. βήματα, ονομάζεται (Draper & Smith, 1997) αντίστοιχα αυτοσυσχέτιση με βήμα-1, βήμα-2, βήμα-3 κ.ο.κ. (*lag-1, lag-2, lag-3, κ.ο.κ.*).



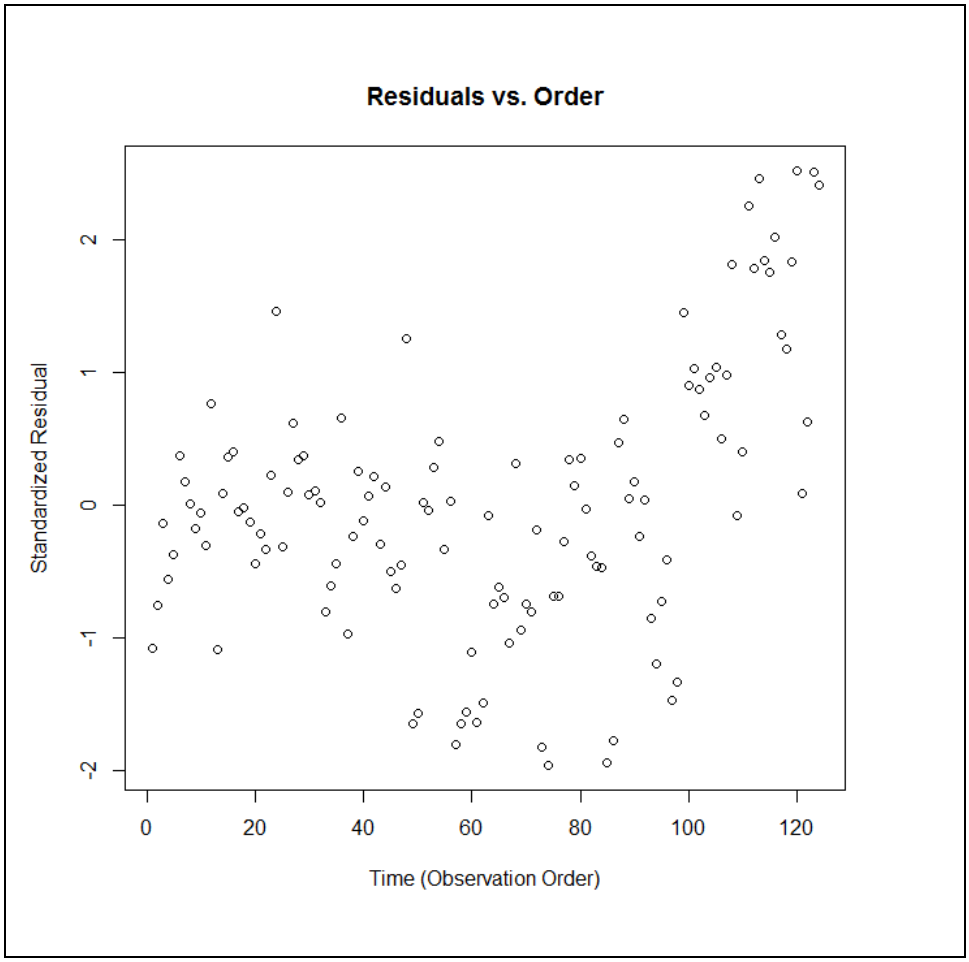
Διάγραμμα 6.11
Παράδειγμα Θετικής Σειριακής Συσχέτισης



Διάγραμμα 6.12
Παράδειγμα Αρνητικής Σειριακής Συσχέτισης

Η διαπίστωση της αυτοσυσχέτισης γίνεται κατ' αρχάς από τη γραφική παράσταση των καταλοίπων σε σχέση με το χρόνο (Βλ. Εικόνα 6.10), και εν συνεχεία από το διάγραμμα διασποράς των καταλοίπων σε σχέση με τις υστερήσεις τους (Βλ. Εικόνα 6.11 και 6.12). Έπειτα γίνεται η χρήση ειδικών ελέγχων για τη διαπίστωση του φαινομένου, όπως για παράδειγμα ο έλεγχος Durbin-Watson, μέσω του οποίου διερευνάται αν υπάρχει αυτοσυσχέτιση που προκύπτει από αυτοπαλίνδρομο υπόδειγμα πρώτου βαθμού (εφαρμόζεται δηλαδή για τη διαπίστωση σειριακής συσχέτισης με βήμα-1).

Στη συνέχεια παρουσιάζεται το διάγραμμα διασποράς των τυποποιημένων (*Standardized*) υπολοίπων σε σχέση με τη σειρά των δεδομένων, δηλαδή συναρτήσει του χρόνου (από τον 1ο έως και τον 124ο μήνα).

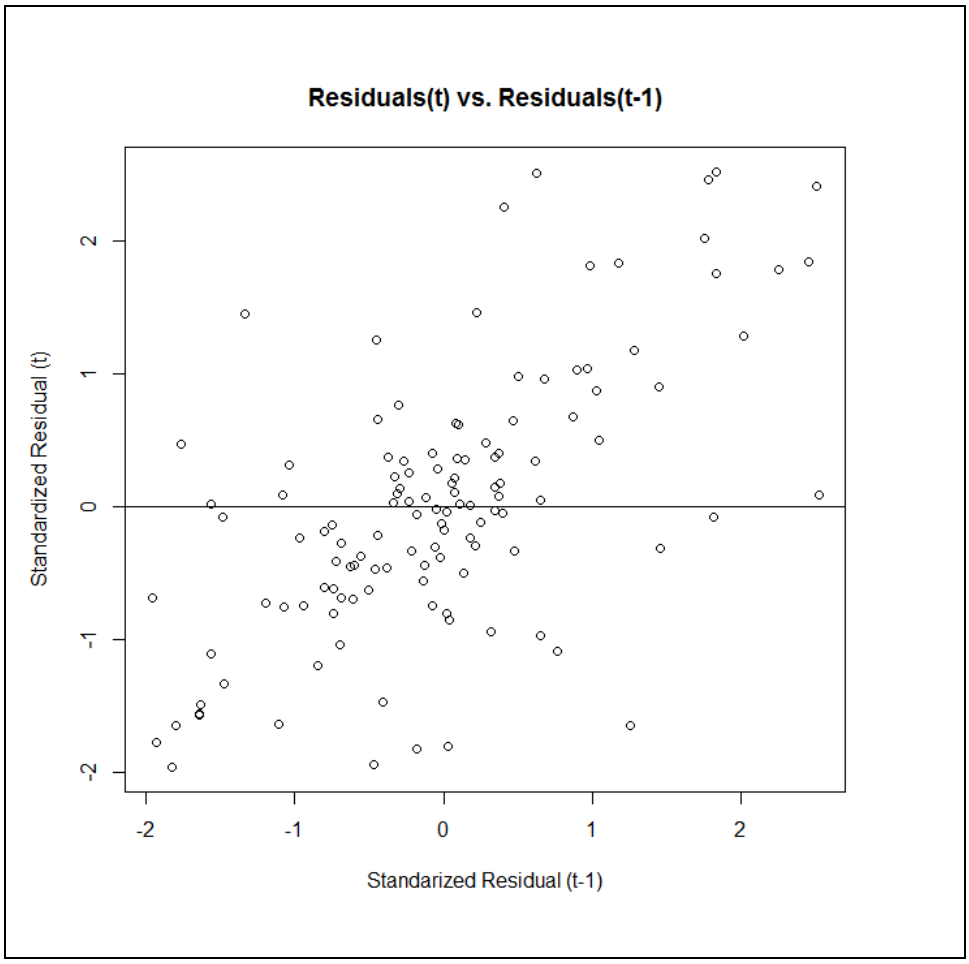


Διάγραμμα 6.13 – Υπόλοιπα σε σχέση με τον χρόνο

Στο παραπάνω διάγραμμα φαίνεται ότι υπάρχει σχέση μεταξύ των υπολοίπων και της σειράς των δεδομένων, γεγονός που παραβιάζει την υπόθεση περί ασυσχέτιστων τυχαίων σφαλμάτων. Για παράδειγμα, παρατηρούμε ότι στις τελευταίες παρατηρήσεις, δηλαδή όσο αυξάνει ο χρόνος, υπάρχει παράλληλη αύξηση και του σφάλματος. Ενώ λοιπόν αναμέναμε τα σφάλματα να κατανέμονται τυχαία γύρω από το μηδέν, παρατηρούμε ότι υπάρχει μία

χρονική εξάρτηση. Η παραβίασή αυτή μας δείχνει ότι υπάρχει το πρόβλημα της αυτοσυσχέτισης και επομένως το μοντέλο δεν ανταποκρίνεται καλά στα δεδομένα μας.

Στη συνέχεια θα παρουσιαστεί το διάγραμμα διασποράς των υπολοίπων e_i με τα αμέσως προηγούμενα, το οποίο αναμένεται να έχει τη μορφή του Διαγράμματος 6.12, προκειμένου να επιβεβαιωθεί η ύπαρξη της αυτοσυσχέτισης.

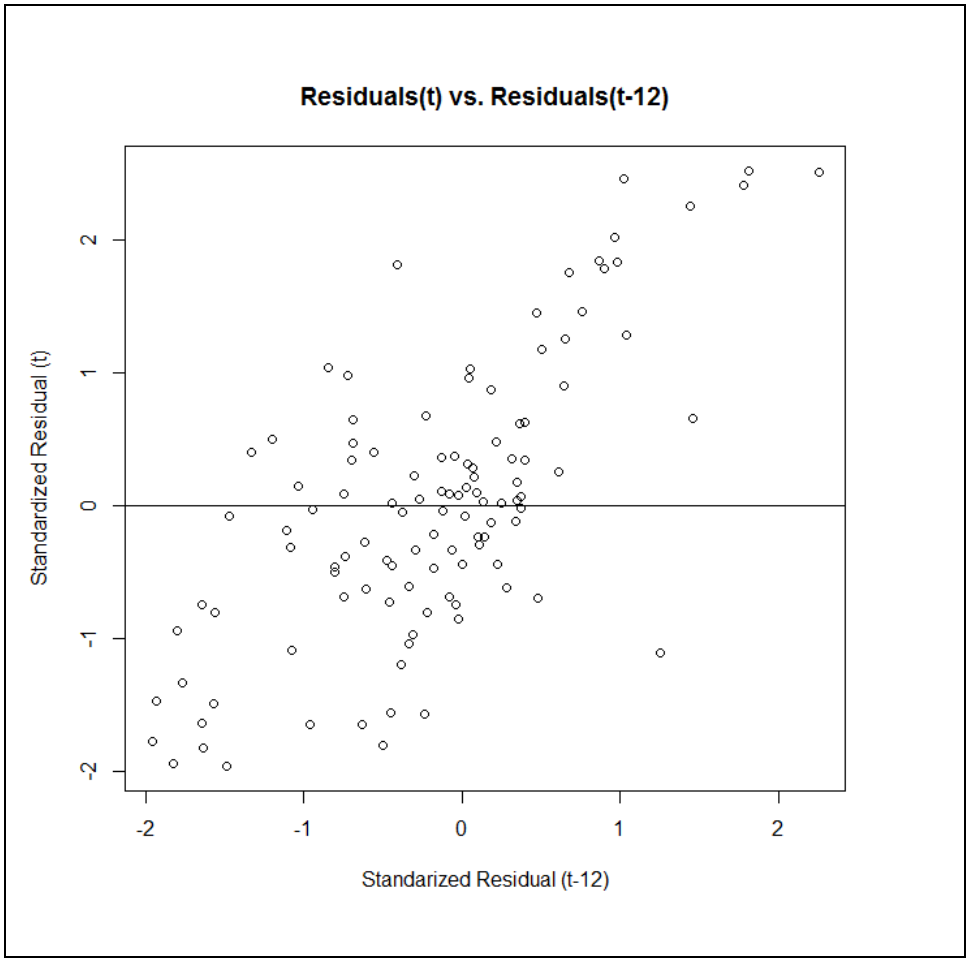


Διάγραμμα 6.14 – Διάγραμμα διασποράς των υπολοίπων e(t) σε σχέση με τα e(t-1)

Παρατηρώντας το Διάγραμμα 6.14 καταλήγουμε στο ίδιο συμπέρασμα, σύμφωνα με το οποίο υπάρχει αυτοσυσχέτιση και μάλιστα θετική. Ένας τρόπος (SchweserNotes, 2013) για να διορθώσουμε την αυτοσυσχέτιση είναι να επαναπροσδιορίσουμε το μοντέλο, ενσωματώνοντας τη χρονολογική του φύση, προσθέτοντας δηλαδή σε αυτό χρονικές υστερήσεις (*time lags*). Αυτό συμβαίνει διότι όταν οι παρατηρήσεις των μεταβλητών προέρχονται από χρονοσειρές, είναι πιθανό να εμφανίζεται χρονική υστέρηση μεταξύ των μεταβλητών. Πιο συγκεκριμένα αυτό σημαίνει ότι ίσως η σχέση μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών δεν αναφέρεται μόνο στην ίδια χρονική περίοδο, αλλά και σε παρελθοντικές.

Το συμπέρασμα λοιπόν που βγάζουμε από το Διάγραμμα 6.14 είναι ότι, κατά την κατασκευή του μοντέλου, πρέπει να λάβουμε υπόψη και τις πωλήσεις αυτοκινήτων του προηγούμενου μήνα.

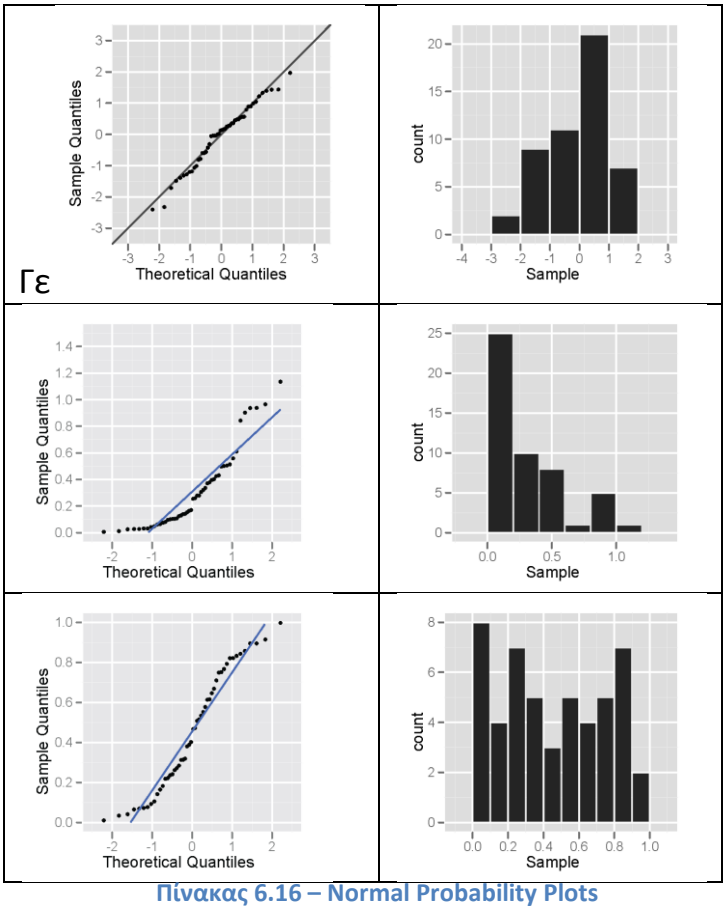
Παράλληλα στο Διάγραμμα 6.3 παρατηρείται έντονη εποχικότητα (*seasonality*) στις πωλήσεις αυτοκινήτων. Πιο συγκεκριμένα παρατηρείται μία έντονη συσχέτιση των παρατηρήσεων που απέχουν μεταξύ τους 12 βήματα. Κάτι τέτοιο δεν είναι παράλογο, καθώς για παράδειγμα κατά τον μήνα Δεκέμβριο οι πωλήσεις αυξάνονται λόγω των ερχόμενων εορτών και των προσφορών, που συνήθως οι αυτοκινητοβιομηχανίες ανακοινώνουν. Στη συνέχεια, προκειμένου να επιβεβαιωθεί η εποχικότητα των παρατηρήσεων, παρουσιάζεται το διάγραμμα διασποράς (Βλ. Διάγραμμα 2.15) των υπολοίπων που απέχουν 12 βήματα, στο οποίο φαίνεται σχετίζονται έντονα.



Διάγραμμα 6.15 – Διάγραμμα διασποράς των υπολοίπων *e*(t) σε σχέση με τα *e*(t-12)

Έτσι λοιπόν συμπεραίνουμε ότι υπάρχει αυτοσυσχέτιση, η οποία θα πρέπει να διορθωθεί. Γι' αυτόν το λόγο, στην επόμενη παράγραφο, θα πραγματοποιηθεί ο επαναπροσδιορισμός του μοντέλου, λαμβάνοντας υπόψη και τις χρονικές υστερήσεις βήματος-1 (*lag-1*) και βήματος-12 (*lag-12*).

6.4.5 Γραφική εξέταση της Κανονικής κατανομής

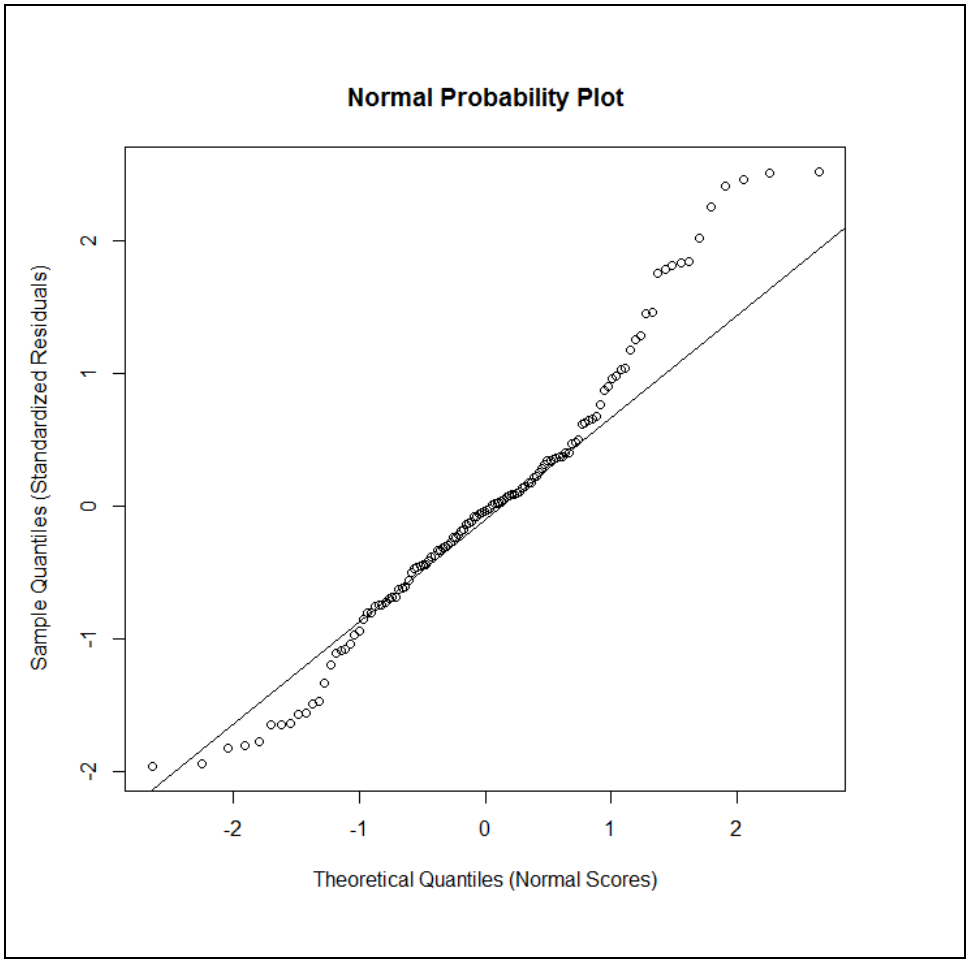


Πίνακας 6.16 – Normal Probability Plots

Προκειμένου να εξεταστεί αν οι τιμές του τυχαίου σφάλματος ακολουθούν την Κανονική κατανομή θα κατασκευαστεί το *"Normal Probability Plot"*. Αν σε αυτό τα σημεία ξεφεύγουν συστηματικά από την ευθεία γραμμή, τότε παραβιάζεται η προϋπόθεση περί Κανονικής κατανομής των σφαλμάτων. Για παράδειγμα, στις διπλανές εικόνες παρατηρούμε τρεις εντελώς διαφορετικές περιπτώσεις. Πιο συγκεκριμένα στην πρώτη περίπτωση οι παρατηρήσεις είναι αρκετά συμμετρικές με αποτέλεσμα στο διάγραμμα NPP να παρατηρείται ότι τα σημεία τείνουν να μην ξεφεύγουν από την ευθεία σε εξαιρετικά ικανοποιητικό βαθμό. Στη δεύτερη περίπτωση παρατηρούμε ότι η κατανομή των παρατηρήσεων δεν είναι συμμετρική και η συνέπεια αυτού είναι στο διάγραμμα NPP να εμφανίζεται ένα σχήμα-C, με τα κοίλα προς τα άνω. Σε περίπτωση που οι παρατηρήσεις ήταν συσσωρευμένες στο δεξί μέρος, τότε τα σημεία στο NPP θα έστρεφαν τα κοίλα κάτω. Τέλος στην τρίτη περίπτωση οι παρατηρήσεις θυμίζουν ομοιόμορφη κατανομή, με αποτέλεσμα να δημιουργείται ένα σχήμα-S στο διάγραμμα NPP.

Στη συνέχεια παρουσιάζεται το *Normal Probability Plot* του μοντέλου μας, προκειμένου να δούμε αν τα τυχαία σφάλματα ακολουθούν την κανονική κατανομή. Παράλληλα θα είμαστε σε θέση να εντοπίσουμε αν υπάρχουν απόμακρα σημεία, τα οποία ενδεχομένως επηρεάζουν το μοντέλο. Κάτι τέτοιο συνήθως γίνεται προκειμένου να αφαιρεθούν κάποιες παρατηρήσεις, οι οποίες προέκυψαν από λανθασμένες μετρήσεις. Στην περίπτωσή μας όμως παρόλο που είναι απίθανο να έχει γίνει κάποιο λάθος στις μετρήσεις (καθώς τα δεδομένα ,τόσο της Η όσο και της Google, είναι επίσημα και άρα η πιθανότητα λάθους ελάχιστη), μία ενδεχόμενη αφαίρεση παρατήρησης

μπορεί να οφείλεται για παράδειγμα σε κάποιον καθαρά οικονομικό παράγοντα (όπως μία οικονομική κρίση που μείωσε τις πωλήσεις, ή αντίστροφα μία μεγάλη προσφορά της αυτοκινητοβιομηχανίας που τις αύξησε απότομα), ο οποίος να κριθεί ότι αν δεν ληφθεί υπόψη, οι προβλέψεις του μοντέλου θα είναι ισχυρότερες και ακριβέστερες.



Διάγραμμα 6.17 – Normal Probability Plot

Η εικόνα του παραπάνω διαγράμματος παρατηρούμε ότι δεν είναι ικανοποιητική. Ένας από τους τρόπους, μέσω των οποίων αυτό μπορεί να βελτιωθεί, είναι να προστεθούν επιπλέον ανεξάρτητες μεταβλητές στο μοντέλο, κάτι που συνάδει και με το συμπέρασμά μας που αφορά τον τρόπο αντιμετώπισης της αυτοσυσχέτισης (ενσωματώνοντας δηλαδή στο μοντέλο τις χρονικές υστερήσεις).

6.5 Γενικό γραμμικό μοντέλο

6.5.1 Μέθοδος της διαδοχικής αφαίρεσης και προσδιορισμός παραμέτρων

Στη συνέχεια θα θεωρήσουμε το γενικό γραμμικό μοντέλο δεύτερης-τάξης για τα δεδομένα των πινάκων 6.1 και 6.2. Πιο συγκεκριμένα, θεωρώντας ως ανεξάρτητη μεταβλητή x τις αναζητήσεις της Η στο Google (και συγκεκριμένα στις ΗΠΑ) και ως εξαρτημένη y τις πωλήσεις των αυτοκινήτων της στις ΗΠΑ και λαμβάνοντας υπόψη τις χρονικές υστερήσεις των πωλήσεων, βήματος-1 (*lag-1*) και βήματος-12 (*lag-12*), θα έχουμε:

$$\begin{aligned} y_t = {} & b_0 + b_1 x_t + b_2 y_{t-1} + b_3 y_{t-12} + b_{11} x_t^2 + b_{22} y_{t-1}^2 + b_{33} y_{t-12}^2 + \\ & + b_{12} x_t \cdot y_{t-1} + b_{13} x_t \cdot y_{t-12} + b_{23} y_{t-1} \cdot y_{t-12} + \varepsilon_t \end{aligned}$$

Το μοντέλο προσαρμόζεται σε:

$$\begin{aligned} \hat{y}_t = {} & \hat{b}_0 + \hat{b}_1 x_t + \hat{b}_2 y_{t-1} + \hat{b}_3 y_{t-12} + b_{11} x_t^2 + \hat{b}_{22} y_{t-1}^2 + \hat{b}_{33} y_{t-12}^2 + \\ & + \hat{b}_{12} x_t \cdot y_{t-1} + \hat{b}_{13} x_t \cdot y_{t-12} + \hat{b}_{23} y_{t-1} \cdot y_{t-12} \end{aligned}$$

Όπου τα ε_t είναι (Οικονόμου & Καρώνη, 2010) τα τυχαία σφάλματα και e_t τα υπόλοιπα ή αλλιώς κατάλοιπα (*residuals*), τα οποία είναι ίσα με:

$$e_t = y_t - \hat{y}_t$$

Σε αυτό το σημείο τονίζεται ότι κατά τη διαδικασία της επιλογής μίας εξίσωσης, κατάλληλης για προβλέψεις, εμφανίζονται δύο κριτήρια τα οποία αντιτίθενται το ένα στο άλλο:

- Το μοντέλο θα πρέπει να περιλαμβάνει όσο γίνεται περισσότερες μεταβλητές έτσι ώστε οι εκτιμήσεις να είναι αξιόπιστες.
- Το μοντέλο θα πρέπει να περιλαμβάνει όσο το δυνατό λιγότερες μεταβλητές, καθώς επεξεργασία, η επεξήγησή, καθώς και η συγκέντρωση πληροφοριών για αυτές κοστίζουν.

Έτσι λοιπόν, προκειμένου να συμβιβαστούμε ανάμεσα και στα δύο κριτήρια, θα επιλέξουμε την καλύτερη εξίσωση παλινδρόμησης, εφαρμόζοντας τη διαδικασία (Draper & Smith, 1997) της διαδοχικής αφαίρεσης ή αλλιώς προς τα πίσω απαλοιφής (*Backward Elimination*). Κατά την εκκίνηση της συγκεκριμένης μεθόδου συμπεριλαμβάνονται στο μοντέλο όλες οι ανεξάρτητες μεταβλητές και στη συνέχεια αφαιρείται διαδοχικά η λιγότερο σημαντική μεταβλητή, που αντιστοιχεί στη μικρότερη αύξηση του SSE, με την

προϋπόθεση ο έλεγχος F να είναι στατιστικά μη σημαντικός. Η διαδικασία επαναλαμβάνεται μέχρι να αφαιρεθούν όλες οι μεταβλητές που έχουν μικρή συμβολή στο μοντέλο, μέχρι δηλαδή η αφαίρεση κάποιας μεταβλητής να είναι στατιστικά σημαντική.

Προτού εφαρμόσουμε τη διαδικασία της προς τα πίσω απαλοιφής πρέπει να υπογραμμιστεί ότι καλά είναι οι ανεξάρτητες μεταβλητές να εκφραστούν σε τυποποιημένη μορφή (αφαιρώντας με τον μέσο όρο και διαιρώντας με την τυπική απόκλιση). Αυτό συμβαίνει διότι αρκετά ζεύγη των μεταβλητών αυτών αναμένεται να έχουν έντονη συσχέτιση, με αποτέλεσμα να δημιουργείται πολυσυγγραμμικότητα. Ενδεικτικά έχουμε ότι η συσχέτιση μεταξύ της y_{t-1} και της y_{t-1}^2 από 0.989, ισούται με 0.245, μετά από την τυποποίηση.

Εφαρμόζοντας τη μέθοδο της διαδοχικής αφαίρεσης, με χρήση του στατιστικού πακέτου R (Βλ. Παράρτημα Δ), καταλήγουμε στο εξής μοντέλο:

$$\hat{y}_t = \hat{b}_0 + \hat{b}_1 x_t + \hat{b}_2 y_{t-1} + \hat{b}_3 y_{t-12} + \hat{b}_{22} y_{t-1}^2 + \hat{b}_{23} y_{t-1} \cdot y_{t-12}$$

Στη συνέχεια θα εκτιμηθούν οι άγνωστες πληθυσμιακές παράμετροι b_0 , b_1 , b_2, b_3, b_{22}, b_{23} με τη βοήθεια των εκτιμητριών[a] που θα προκύψουν από την παλινδρόμηση. Η εκτίμηση των παραμέτρων b_0 , b_1 , b_2, b_3, b_{22}, b_{23} του υποδείγματος θα πραγματοποιηθεί με τη μέθοδο ελαχίστων τετραγώνων (*least squares method*). Η ανάλυση των δεδομένων και η εκτίμηση της ανάλυσης παλινδρόμησης πραγματοποιήθηκε με χρήση του στατιστικού πακέτου R και φαίνεται στο Παράρτημα Δ.

Έτσι λοιπόν έχουμε τα εξής αποτελέσματα:

$$\hat{b}_0 = 44595.7 \ , \ \hat{b}_1 = 3059.1 \ , \ \hat{b}_2 = 2651.9$$

$$\hat{b}_3 = 7208.1 \ , \ \hat{b}_{22} = -1828.9 \ , \ \hat{b}_{23} = 1851.5$$

Στη συνέχεια θα πραγματοποιηθεί ο έλεγχος της ερμηνευτικής ικανότητας του μοντέλου ο οποίος, όπως έχει προαναφερθεί, διακρίνεται σε δύο επίπεδα: Στον έλεγχο της στατιστικής σημαντικότητας των επιμέρους παραμέτρων (στη αυτήν την περίπτωση b_0 , b_1 , b_2, b_3, b_{22}, b_{23}), ο οποίος ουσιαστικά πραγματοποιήθηκε με την εφαρμογή της διαδικασίας διαδοχικής αφαίρεσης, καθώς και στον έλεγχο της ερμηνευτικής ικανότητας του μοντέλου συνολικά (μέσω του Στατιστικού ελέγχου F και του συντελεστή προσδιορισμού R^2).

[a] *Δηλαδή των* $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$, $\hat{b}_3$, $\hat{b}_{22}$ και $\hat{b}_{23}$

Παράλληλα θα εξετάσουμε αν παραβιάζεται κάποιος από τους πέντε ισχυρισμούς, που αναφέρθηκαν στις προηγούμενες παραγράφους και πρέπει να ισχύουν ώστε το μοντέλο να είναι κατάλληλο προκειμένου να μπορούμε να βγάλουμε συμπεράσματα, σχετικά με τη συμπεριφορά των παρατηρήσεων. Επιπροσθέτως, αν αποδειχθεί ότι όλα τα παραπάνω δεν εμφανίζουν προβλήματα, χρησιμοποιώντας την "απόσταση Cook" (*Cook's Distance*), θα εξεταστεί αν υπάρχουν σημεία επιρροής. Τα σημεία αυτά έχουν σαν αποτέλεσμα τη διαφοροποίηση των παραμέτρων του μοντέλου, σε σύγκριση με αυτές που θα λαμβάναμε, αν οι παρατηρήσεις τους ήταν απούσες, και μία ενδεχόμενη απομάκρυνση τους ίσως βελτιώσει σημαντικά το μοντέλο.

6.5.2 Έλεγχος της ερμηνευτικής ικανότητας του μοντέλου

- Στατιστικός Έλεγχος F:
  Προκειμένου να ελέγξουμε τη σημαντικότητα της παλινδρόμησης χρησιμοποιούμε τη στατιστική συνάρτηση F. Στη συνέχεια πραγματοποιούμε τον παρακάτω έλεγχο για τους συντελεστές του μοντέλου:

  $$H_0 : b_1 = b_2 = b_3 = b_{22} = b_{23} = 0 \text{ έναντι της } H_1 : \text{τουλάχιστον ένα } b_j \neq 0$$

  Η στατιστική συνάρτηση F υπό την H_0 ακολουθεί[a] την κατανομή $F_{k,(n-k-1)}$

  Από τα αποτελέσματα του Παραρτήματος Δ, έχουμε:

  $$F = 159.4 \quad (p\text{-}\tau\iota\mu\acute{\eta} < 0.0001)$$

  Άρα συμπεραίνουμε ότι η μηδενική υπόθεση απορρίπτεται.

- Συντελεστής προσδιορισμού R^2:
  Όπως φαίνεται από τα αποτελέσματα του Παραρτήματος Δ, στην περίπτωση μας το 87.1% της μεταβλητότητας των πωλήσεων αυτοκινήτων της Η στις ΗΠΑ, εξηγείται από τις μεταβλητές μας. Το συγκεκριμένο ποσοστό κρίνεται εξαιρετικά ικανοποιητικό[b], προκειμένου να πραγματοποιήσουμε μελλοντικές προβλέψεις, υπό την προϋπόθεση βέβαια ότι το μοντέλο μας ικανοποιεί τις υποθέσεις, που αναφέρθηκαν στις προηγούμενες παραγράφους, οι οποίες θα εξεταστούν στη συνέχεια.

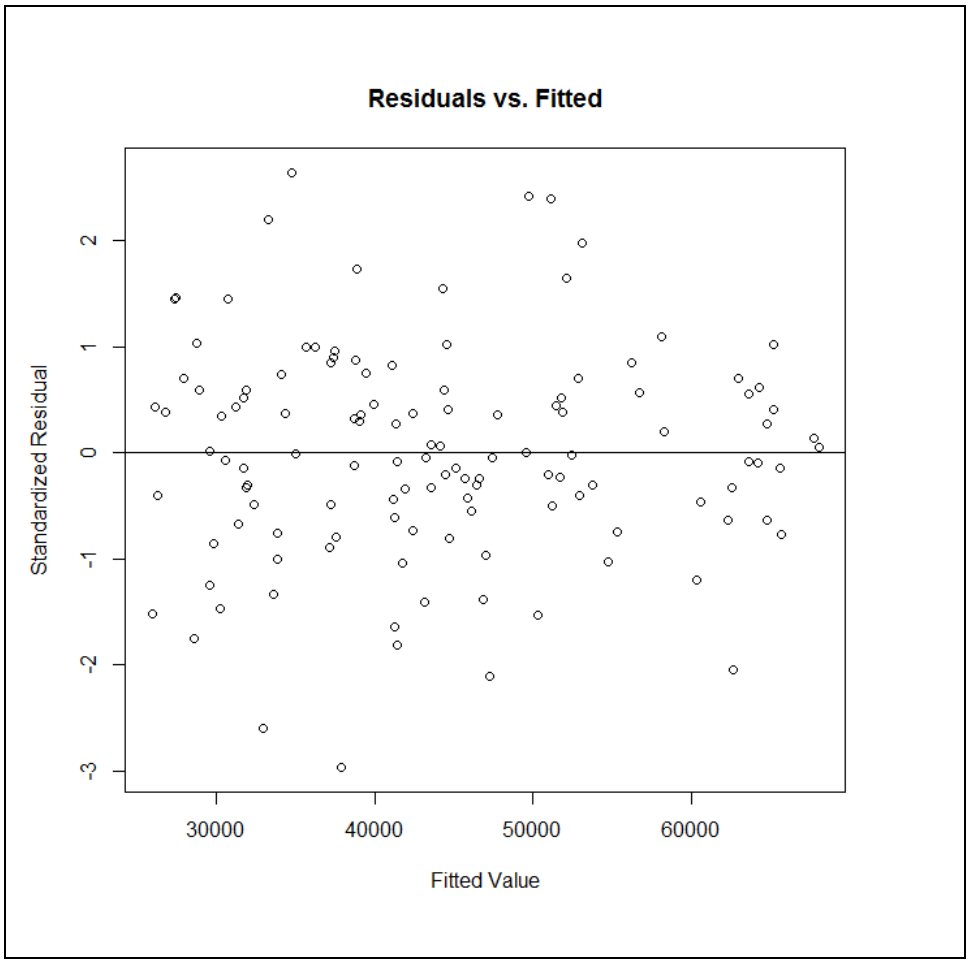
---

[a] Για $k=5$ και $n=124$.

[b] Γενικά ένα μοντέλο με συντελεστή προσδιορισμού μεγαλύτερο από 60% θεωρείται ικανοποιητικό.

6.5.3 Έλεγχος για ύπαρξη ετεροσκεδαστικότητας

Στη συνέχεια παρουσιάζεται το διάγραμμα διασποράς των τυποποιημένων (*Standardized*) υπολοίπων συναρτήσει των εκτιμημένων $\hat{y}_t$ του μοντέλου μας.

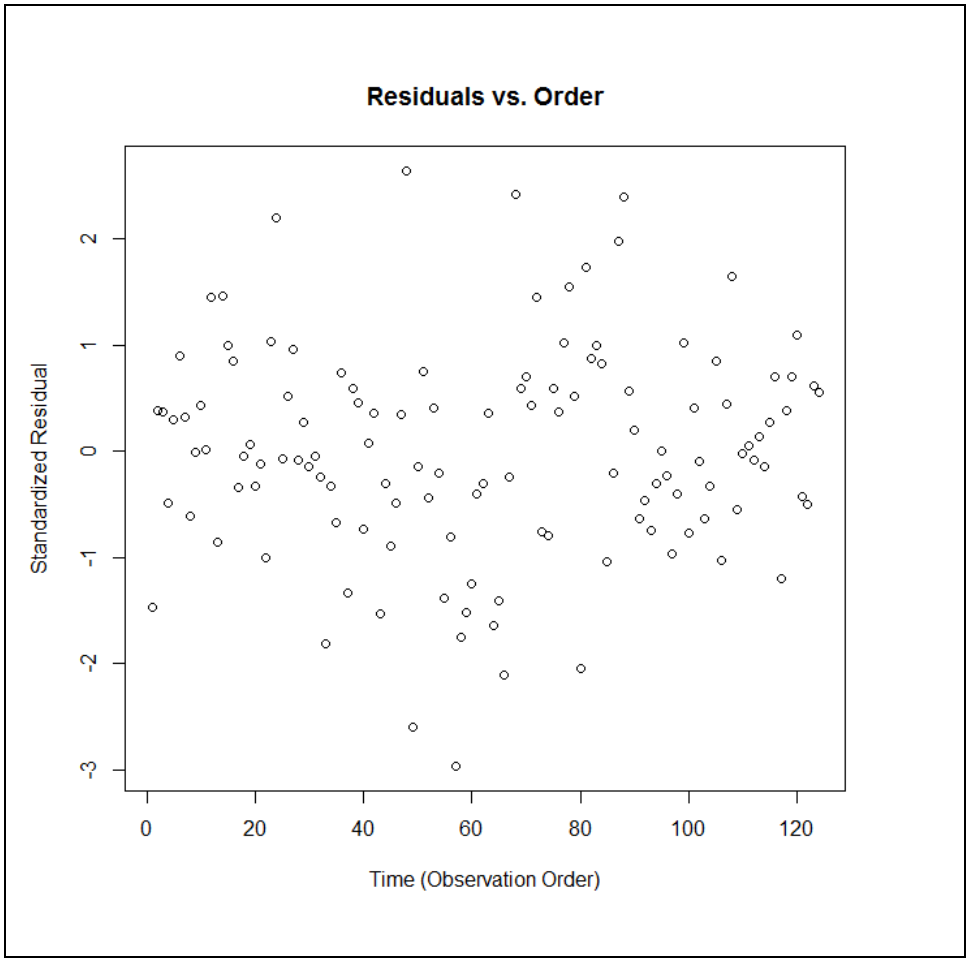


Διάγραμμα 6.18 – Υπόλοιπα σε σχέση με τις τιμές του x

Στο παραπάνω διάγραμμα φαίνεται ότι η διασπορά των υπολοίπων είναι σταθερή, καθώς κατανέμονται τυχαία πάνω και κάτω από την ευθεία που διέρχεται από το μηδέν. Επομένως μπορούμε να συμπεράνουμε ότι δεν υπάρχει ένδειξη ετεροσκεδαστικότητας, καθώς δεν υπάρχει κάποια μορφή εξάρτησης ανάμεσα στα υπόλοιπα e_t και τα εκτιμημένα $\hat{y}_t$.

6.5.4 Έλεγχος για ύπαρξη αυτοσυσχέτισης

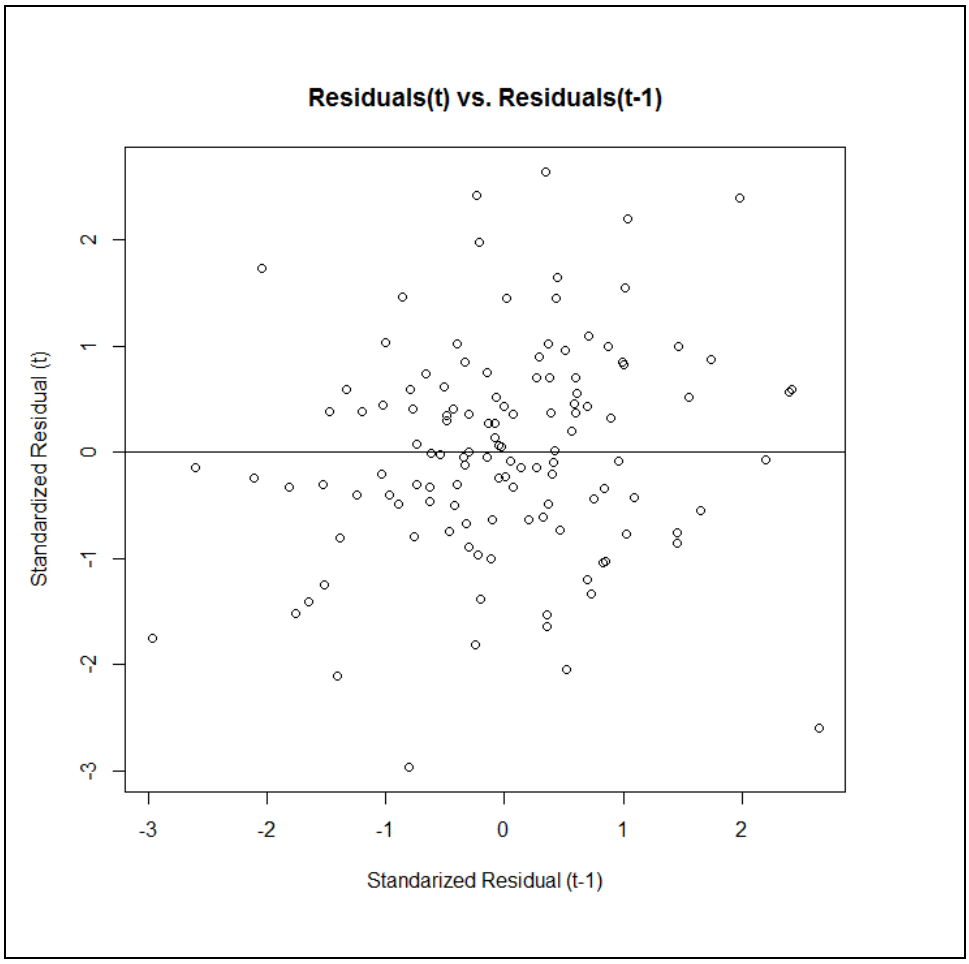
Στη συνέχεια παρουσιάζεται το διάγραμμα διασποράς των τυποποιημένων (*Standardized*) υπολοίπων σε σχέση με τη σειρά των δεδομένων, δηλαδή συναρτήσει του χρόνου (από τον 1^{ο} έως και τον 124^{ο} μήνα).



Διάγραμμα 6.19 – Υπόλοιπα σε σχέση με τον χρόνο

Στο παραπάνω διάγραμμα φαίνεται ότι δεν υπάρχει σχέση μεταξύ των υπολοίπων και της σειράς των δεδομένων, γεγονός που δεν παραβιάζει την υπόθεση περί ασυσχέτιστων τυχαίων σφαλμάτων. Έτσι λοιπόν παρατηρούμε ότι τα σφάλματα κατανέμονται τυχαία γύρω από το μηδέν και επομένως συμπεραίνουμε ότι δεν υπάρχει χρονική εξάρτηση. Σε αντίθεση με τα αποτελέσματα της προηγούμενης παραγράφου, στο συγκεκριμένο μοντέλο δεν υπάρχουν ενδείξεις για παραβίαση της προϋπόθεσης των ασυσχέτιστων σφαλμάτων, γεγονός που μας οδηγεί στο να καταλάβουμε ότι το πρόβλημα της αυτοσυσχέτισης δεν είναι παρόν. Έτσι λοιπόν λόγω της ενσωμάτωσης των χρονικών υστερήσεων στο μοντέλο η αυτοσυσχέτιση παύει να υπάρχει με αποτέλεσμα να είμαστε σε θέση να συνεχίσουμε την διερεύνησή του.

Στη συνέχεια θα παρουσιαστεί το διάγραμμα διασποράς των υπολοίπων e_i με τα αμέσως προηγούμενα, το οποίο αναμένεται να επιβεβαιώσει την απουσία του προβλήματος της αυτοσυσχέτισης.

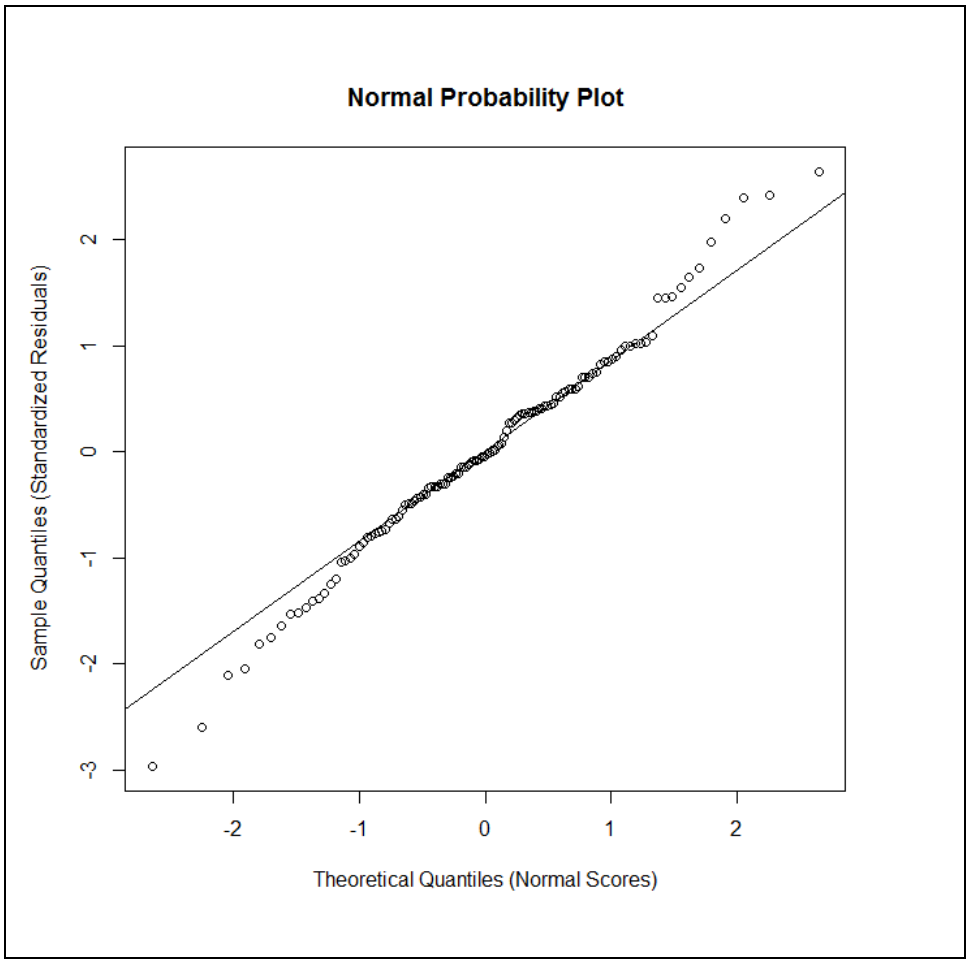


Διάγραμμα 6.20 – Διάγραμμα διασποράς των υπολοίπων e(t) σε σχέση με τα e(t-1)

Παρατηρώντας το Διάγραμμα 6.20 καταλήγουμε στο ίδιο συμπέρασμα, σύμφωνα με το οποίο δεν υπάρχει αυτοσυσχέτιση.

6.5.5 Γραφική εξέταση της Κανονικής κατανομής

Στη συνέχεια παρουσιάζεται το *Normal Probability Plot* του μοντέλου μας, προκειμένου να δούμε αν τα τυχαία σφάλματα ακολουθούν την κανονική κατανομή. Παράλληλα θα είμαστε σε θέση να εντοπίσουμε αν υπάρχουν απόμακρα σημεία, τα οποία ενδεχομένως επηρεάζουν το μοντέλο.

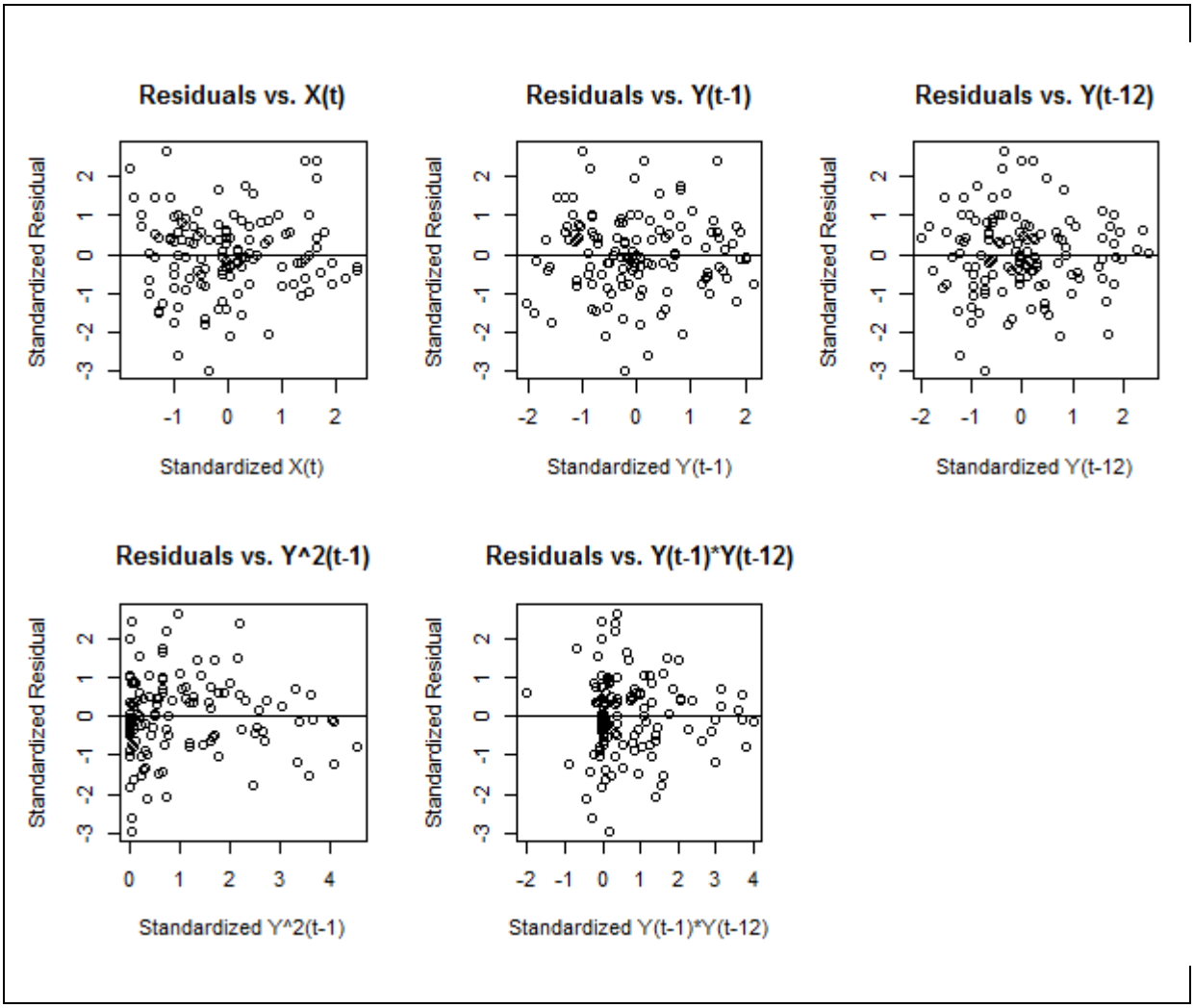


Διάγραμμα 6.21 – Normal Probability Plot

Η εικόνα του παραπάνω διαγράμματος παρατηρούμε ότι είναι ικανοποιητική. Με εξαίρεση κάποια ελάχιστα σημεία υπάρχει καλή προσαρμογή στην ευθεία, γεγονός που μας οδηγεί στο να συμπεράνουμε ότι η ενσωμάτωση ανεξάρτητων μεταβλητών στο μοντέλο, άλλαξε την αρνητική εικόνα του αντίστοιχου διαγράμματος στο απλό γραμμικό μοντέλο (Βλ. διάγραμμα 6.17). Η εικόνα του διαγράμματος μπορεί να βελτιωθεί ακόμα περισσότερο, αν εντοπιστούν και αφαιρεθούν κάποιες πιθανές έκτροπες παρατηρήσεις.

6.5.6 Έλεγχος για συσχέτιση του σφάλματος με τις ανεξάρτητες μεταβλητές

Στη συνέχεια παρουσιάζονται τα διαγράμματα διασποράς των τυποποιημένων (*Standarized*) υπολοίπων συναρτήσει των ανεξάρτητων μεταβλητών του μοντέλου μας.



Διάγραμμα 6.22 – Υπόλοιπα σε σχέση με τις ανεξάρτητες μεταβλητές

Στα τρία πρώτα, εκ των παραπάνω διαγραμμάτων, φαίνεται ότι η διασπορά των υπολοίπων είναι σταθερή, καθώς κατανέμονται τυχαία πάνω και κάτω από την ευθεία που διέρχεται από το μηδέν. Η τάση που παρουσιάζεται στα δύο τελευταία διαγράμματα δεν πρέπει να μας ανησυχεί, καθώς οφείλεται στη φύση των μεταβλητών (η πρώτη είναι το τετράγωνο της μεταβλητής y_{t-1}, ενώ η δεύτερη είναι η αλληλεπίδραση των μεταβλητών y_{t-1} και y_{t-12}). Επομένως μπορούμε να συμπεράνουμε ότι δεν παρατηρείται κάποια εμφανής σχέση μεταξύ των υπολοίπων και των πέντε ανεξάρτητων μεταβλητών.

6.5.7 Έλεγχος για ύπαρξη πολυσυγγραμμικότητας

Αρχικά πρέπει να τονιστεί ότι η πολυσυγγραμμικότητα δεν έχει έντονες επιπτώσεις, όσον αφορά την αξία ενός μοντέλου και την προσαρμογή του στην εξαρτημένη μεταβλητή. Το πρόβλημα που δημιουργεί όμως είναι ότι καθιστά δύσκολη την ερμηνεία της κάθε μίας εκ των ανεξάρτητων μεταβλητών, καθώς υπάρχει έντονη συσχέτιση μεταξύ αυτών.

Όπως αναφέρθηκε στην αρχή της παραγράφου, από τη στιγμή που επιλέξαμε να χρησιμοποιήσουμε το γενικό γραμμικό μοντέλο δεύτερης-τάξης, έπρεπε να εκφράσουμε τις μεταβλητές μας σε τυποποιημένη μορφή, καθώς αρκετά ζεύγη από αυτές εμφάνιζαν μεγάλη συσχέτιση μεταξύ τους. Αν αυτό δεν συνέβαινε, τότε είναι δεδομένο ότι θα εμφανιζόταν το πρόβλημα της πολυσυγγραμμικότητας και μάλιστα σε ιδιαίτερα μεγάλο βαθμό. Ωστόσο το γεγονός ότι τυποποιήσαμε τις ανεξάρτητες μεταβλητές του μοντέλου, δεν συνεπάγεται απαραίτητα ότι η πολυσυγγραμμικότητα παύει αυτομάτως να υπάρχει. Για το λόγο αυτό, με χρήση του στατιστικού πακέτου R (Βλ. Παράρτημα Δ), υπολογίστηκε η "ανοχή" (*tolerance*) και ο "παράγοντας μεγέθυνσης διασποράς" (*variance inflation factor – VIF*), ο οποίος μας δείχνει, όταν οι ανεξάρτητες μεταβλητές συσχετίζονται μεταξύ τους, κατά πόσο αυξάνεται η διασπορά ενός εκτιμημένου συντελεστή[a] του μοντέλου.

Έτσι λοιπόν έχουμε (Οικονόμου & Καρώνη, 2010):

$$tolerance = 1 - R_j^2 \quad \text{και} \quad VIF = \frac{1}{tolerance} = \frac{1}{1 - R_j^2}$$

Ενδείξεις ύπαρξης πολυσυγγραμμικότητας θεωρούνται οι τιμές:

$$tolerance < 0.2 \text{ , ή ισοδύναμα: } VIF > 5$$

Από παρακάτω πίνακα, στον οποίο παρουσιάζονται οι τιμές των παραγόντων μεγέθυνσης διασποράς για κάθε μία από τις ανεξάρτητες μεταβλητές μας, συμπεραίνουμε ότι δεν υπάρχει πολυσυγγραμμικότητα (Βλ. Παράρτημα Δ).

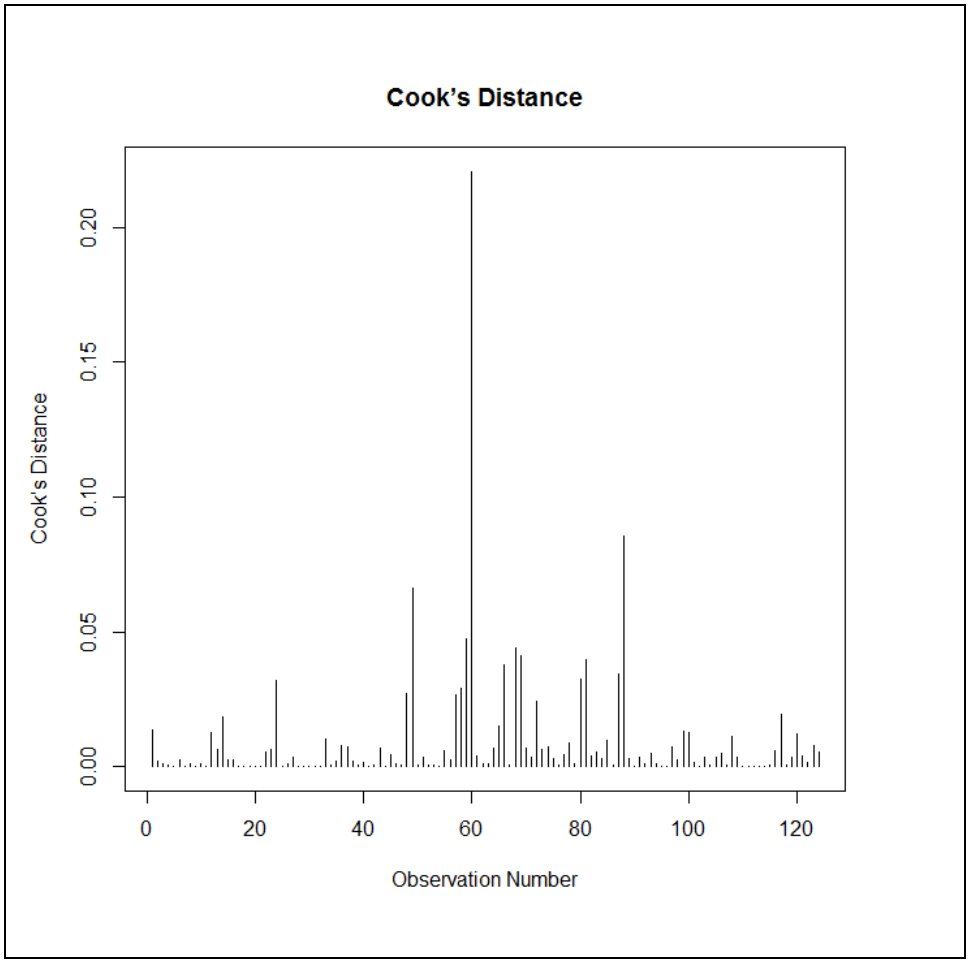
Μεταβλητή	x_t	y_{t-1}	y_{t-12}	y_{t-1}^2	$y_{t-1} \cdot y_{t-12}$
VIF	1.592540	2.487086	2.213030	2.205525	2.406208

Πίνακας 6.23 - Τιμές των παραγόντων μεγέθυνσης διασποράς

[a] *Δηλαδή της* $Var(\hat{b}_j)$

6.5.8 Έλεγχος για ύπαρξη σημείων επιρροής

Όπως έχει αναφερθεί ένα σημείο επιρροής μπορεί να προκαλέσει σημαντικές διαφοροποιήσεις στις εκτιμήσεις των παραμέτρων του μοντέλου. Παράλληλα όμως πρέπει να υπογραμμιστεί ότι στην περίπτωση που εντοπιστεί ένα σημείο επιρροής, πρέπει να εξεταστεί και η επίδραση που θα έχει η ενδεχόμενη αφαίρεσή του. Στη συνέχεια θα απεικονίσουμε την απόσταση Cook (*Cook's Distance*), για κάθε μία από τις παρατηρήσεις του δείγματός μας, με χρήση του στατιστικού πακέτου R (Βλ. Παράρτημα Δ).



Διάγραμμα 6.24 – Απόσταση Cook

Σύμφωνα με το παραπάνω διάγραμμα δεν υπάρχουν σημεία επιρροής στο δείγμα μας, καθώς με βάση τον γενικό κανόνα (Οικονόμου & Καρώνη, 2010), πρέπει να εστιάζουμε και να εξετάζουμε με προσοχή τις παρατηρήσεις για τις οποίες ισχύει:

$D_i > 1$, όπου D_i η απόσταση Cook της i παρατήρησης

Ωστόσο παρατηρούμε ότι η παρατήρηση 60 είναι απέχει σημαντικά από τις υπόλοιπες. Αυτό συμβαίνει διότι η συγκεκριμένη παρατήρηση αντιστοιχεί

στον Δεκέμβριο του 2008, κατά τη διάρκεια του οποίου η οικονομική κρίση των Ηνωμένων Πολιτειών βρίσκεται στο ανώτατο σημείο της. Πιο συγκεκριμένα αν παρατηρήσουμε τα δεδομένα μας (Βλ. Πίνακα 6.1), θα δούμε μία έντονη αύξηση των πωλήσεων κατά τον μήνα Δεκέμβριο, όλων των ετών, λόγω των εορτών των Χριστουγέννων. Αντίθετα στην παρατήρηση 60 (Δεκέμβριος 2008) δεν παρατηρείται η ίδια τάση στις πωλήσεις λόγω της οικονομικής αστάθειας της χώρας. Έτσι λοιπόν ενώ το μοντέλο θα "περίμενε" ένα απότομο ανέβασμα της εξαρτημένης μεταβλητής, λαμβάνει μία πιο "χλιαρή" μεταβολή. Το γεγονός αυτό συνεπάγεται την εμφάνιση μεγάλου σφάλματος, για τη συγκεκριμένη παρατήρηση και κατ' επέκταση μεγάλη απόσταση Cook.

Παρόλα αυτά επειδή η τιμή της απόστασης Cook, για την παρατήρηση 60, δεν είναι κοντά στη μονάδα κρίθηκε ότι δεν είναι απαραίτητη η παράληψή της, καθώς δεν θα έχει κάποια ουσιαστική συνέπεια στη βελτίωση του μοντέλου. Σαν μελλοντική διερεύνηση μπορεί να αφαιρεθεί η συγκεκριμένη παρατήρηση και να συγκριθεί το μοντέλο που προκύπτει με το αρχικό, έτσι ώστε να κριθεί αν η αφαίρεσή της συνδράμει στη βελτίωση των προβλέψεων.

6.6 Πραγματοποίηση πρόβλεψης και συμπεράσματα

Το μοντέλο στο οποίο καταλήξαμε στην παράγραφο 6.6 είναι το εξής:

$$\hat{y}_t = \hat{b}_0 + \hat{b}_1 x_t + \hat{b}_2 y_{t-1} + \hat{b}_3 y_{t-12} + \hat{b}_{22} y_{t-1}^2 + \hat{b}_{23} y_{t-1} \cdot y_{t-12}$$

Όπου:

$$\hat{b}_0 = 44595.7 \ , \ \hat{b}_1 = 3059.1 \ , \ \hat{b}_2 = 2651.9$$

$$\hat{b}_3 = 7208.1 \ , \ \hat{b}_{22} = -1828.9 \ , \ \hat{b}_{23} = 1851.5$$

Στη συνέχεια θα πραγματοποιήσουμε μία πρόβλεψη για τον Μάιο του 2014, δηλαδή για τον 125^{ο} μήνα.

Από τα στοιχεία του Google Trends, καθώς και από τον Πίνακα 6.1, από τον οποίο λαμβάνουμε τις τιμές των χρονικών υστερήσεων που απαιτούνται για τις υπόλοιπες τέσσερις μεταβλητές, έχουμε:

$$x_{125} = 68 \ , \ y_{124} = 66107 \ , \ y_{113} = 68358$$

Οι παραπάνω παρατηρήσεις θα εκφραστούν σε συνδυασμό με τις υπόλοιπες παρατηρήσεις σε τυποποιημένη μορφή, και στη συνέχεια θα προσδιοριστούν και οι τιμές των δύο εναπομεινάντων ανεξάρτητων μεταβλητών (του τετραγώνου της μεταβλητής y_{t-1}, και της αλληλεπίδρασης των μεταβλητών y_{t-1} και y_{t-12}). Με αυτόν τον τρόπο οι πέντε ανεξάρτητες μεταβλητές θα είναι έτοιμες να εισέλθουν στην εξίσωση, προκειμένου να πραγματοποιήσουμε την πρόβλεψη του 125^{ου} μήνα.

Η διαδικασία της πρόβλεψης πραγματοποιήθηκε με χρήση του στατιστικού πακέτου R (Βλ. Παράρτημα Δ). Πιο συγκεκριμένα τα αποτελέσματα που λαμβάνουμε είναι τα εξής:

- Το 95% διάστημα εμπιστοσύνης για τη μέση τιμή της y (64994 , 71033)
- Το 95% διάστημα εμπιστοσύνης για την άγνωστη τιμή της y (58563 , 77464)

Σύμφωνα με την επίσημη μηνιαία αναφορά (Hyundai News, 2014) της Η Motor America, οι πωλήσεις των αυτοκινήτων Η στις ΗΠΑ για τον μήνα Μάιο του 2014, ανέρχονται στις 70907, αριθμός που βρίσκεται μέσα στα διαστήματα εμπιστοσύνης και επαληθεύει την αξιοπιστία του μοντέλου.

Παραρτήματα

Παράρτημα Α. Ο απόηχος του κλεισίματος της ΕΡΤ στο διαδίκτυο

Α.1 Εισαγωγή

Το καλοκαίρι του 2013, η ΕΡΤ οδηγείται (In, 2013) σε κλείσιμο και πιο συγκεκριμένα, λίγο πριν τα μεσάνυχτα της 12ης Ιουνίου διακόπτεται σε ολόκληρη την Ελλάδα η εκπομπή σήματος, από τις ραδιοτηλεοπτικές κεραίες. Το αναλογικό σήμα των τηλεοπτικών καναλιών της ΕΡΤ διεκόπη στις 22:54, ενώ λίγα λεπτά αργότερα, στις 23:12, μαύρο πέφτει και στις ψηφιακές εκπομπές. Η διακοπή του ραδιοφωνικού σήματος λαμβάνει χώρα στις 23:22. Ταυτόχρονα σταματάει η εκπομπή και άλλων καναλιών που χρησιμοποιούσαν τους ίδιους αναμεταδότες.

Αυτομάτως η είδηση του επικείμενου κλεισίματος της ΕΡΤ συγκέντρωσε το ενδιαφέρον των πολιτών και των ΜΜΕ, τόσο στην Ελλάδα, όσο και στο εξωτερικό (Newsit, 2013), ενώ παράλληλα γίνεται πρώτο θέμα και στα social media (Πρώτο Θέμα, 2013). Πιο συγκεκριμένα το Facebook και το Twitter κατακλύζονται από συζητήσεις, μηνύματα, σχόλια και κριτικές των χρηστών γύρω από το θέμα. Επίσης αρκετοί από τους εργαζόμενους της ΕΡΤ, καθώς και απλοί χρήστες του διαδικτύου, κάνουν συνεχείς αναρτήσεις, που αφορούν το θέμα, στην αγγλική γλώσσα προκειμένου αυτό να γίνει γνωστό σε ολόκληρο τον κόσμο, καθώς και στα μέσα ενημέρωσης του εξωτερικού.

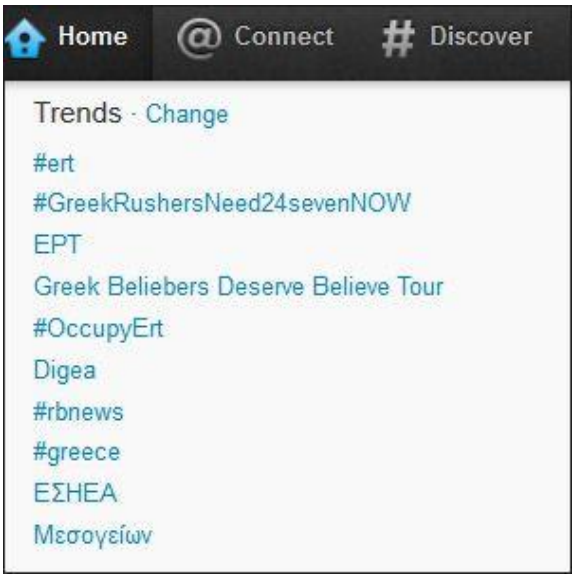
Η περίπτωση της ΕΡΤ από τα μέσα μέχρι τα τέλη του Ιουνίου μονοπώλησε το ενδιαφέρον των πολιτών, τόσο εντός όσο και εκτός συνόρων. Τα μέσα κοινωνικής δικτύωσης έπαιξαν καταλυτικό ρόλο σε αυτό, καθώς ήταν η αιτία που η είδηση διαδόθηκε με ταχύτατους ρυθμούς, ενώ παράλληλα αποτέλεσαν εστία ανταλλαγής απόψεων και πολιτικής ζύμωσης γύρω από το θέμα. Αν μάλιστα αναλογιστούμε και το γεγονός ότι οι δημοσιογράφοι βρίσκονταν σε απεργιακές κινητοποιήσεις για αρκετές μέρες μετά την 11η Ιουνίου, καταλαβαίνουμε ότι ουσιαστικά τα social media ήταν αυτά που ανέλαβαν τα ηνία της ενημέρωσης των πολιτών στα θέματα που αφορούσαν το κλείσιμο της ΕΡΤ.

Α.2 Η εμφάνιση της ΕΡΤ στο Twitter

Μόλις τα πρώτα σενάρια για το κλείσιμο της ΕΡΤ και την απόλυση των εργαζομένων της βλέπουν το φως της δημοσιότητας, εκατοντάδες μηνύματα κατακλύζουν τα μέσα κοινωνικής δικτύωσης. Πιο συγκεκριμένα το βράδυ της 11ης Ιουνίου το Hashtag[a] #ert καθώς και η ελληνική λέξη ΕΡΤ γίνονται η δεύτερη και η τρίτη δημοφιλέστερη λέξη στο Twitter παγκοσμίως, αντίστοιχα (Βλ. Εικόνα 3.1).

#	Trending Topic
1	#DirenGeziSeninleyiz
2	#ert
3	EPT
4	#ReasonsWeStoppedTalking
5	#thingsihatemost
6	#WeWantAustinMahoneInItaly
7	#LessonsFromScaryMovies
8	#ProvokatörlereUYMA
9	iTunes Radio
10	#SonyE3

Εικόνα Α.1 - Παγκόσμια Trends



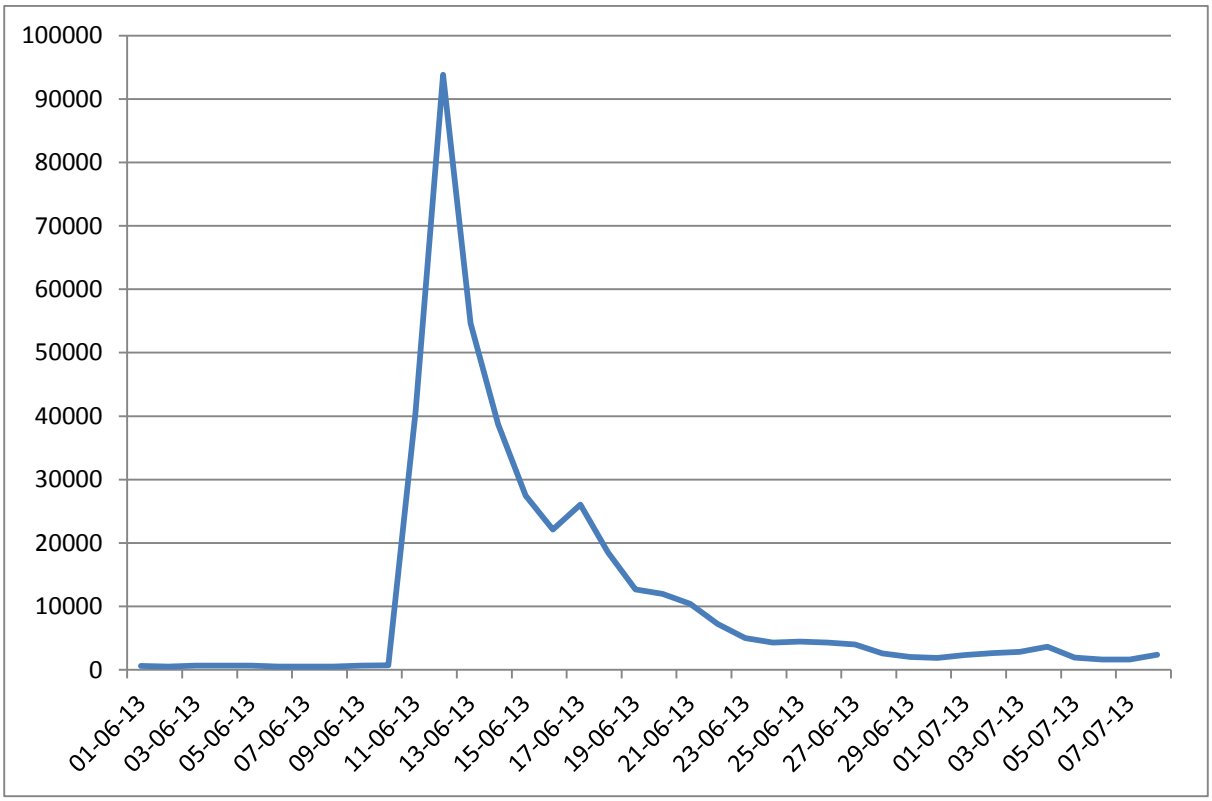
Εικόνα Α.2 - Ελληνικά Trends

Όμοια εικόνα παρατηρείται και στα μηνύματα, τα οποία προέρχονται από ελληνικούς λογαριασμούς Twitter (Βλ. Εικόνα 3.2). Έτσι λοιπόν, σύμφωνα με τα στοιχεία του Twitter, οι έξι εκ των δέκα δημοφιλέστερων λέξεων, στα μηνύματα των Ελλήνων χρηστών, σχετίζονται με τη δημόσια ραδιοτηλεόραση (και συγκεκριμένα οι λέξεις #ert, ΕΡΤ, #OccupyErt, Digea, ΕΣΗΕΑ, Μεσογείων).

[a] Το Hashtag είναι μία λέξη ή φράση με πρόθεμα το σύμβολο # και πρόκειται για ετικέτες (*metadata tags*) οι οποίες χρησιμοποιούνται στα blogs και στα social media προκειμένου να δοθεί έμφαση στο περιεχόμενο της λέξης - φράσης. Η βασικότερη ιδιότητά τους είναι ότι παρέχουν στον χρήστη ένα μέσο ομαδοποίησης μηνυμάτων, δίνοντας του τη δυνατότητα να μεταβεί σε όλα τα μηνύματα που περιέχουν το hashtag, που τον ενδιαφέρει. (Messina, 2008)

Α.3 Η ΕΡΤ στο Twitter συναρτήσει του χρόνου

Στη συγκεκριμένη μελέτη μετρήθηκε η εμφάνιση των λέξεων “ert” και “ερτ” στο Twitter από την 1η Ιουνίου, μέχρι και την 8η Ιουλίου του 2013. Στο παρακάτω διάγραμμα (Βλ. Διάγραμμα 3.3) φαίνεται ο συνολικός αριθμός μηνυμάτων, που χρήστες ανέφεραν τουλάχιστον μία εκ των παραπάνω λέξεων, συναρτήσει του χρόνου (ημερησίως).



Διάγραμμα Α.3 - Αναφορές των λέξεων ερτ και ert στο Twitter συναρτήσει του χρόνου

Προκειμένου να μελετηθεί η αύξηση των αναφορών των συγκεκριμένων λέξεων, θα ορίσουμε τις μεταβλητές E_i για $i=1,2,...,38$, οι οποίες αναπαριστούν τις ημερήσιες εμφανίσεις, των λέξεων ert και ερτ, στο Twitter μεταξύ της 1ης Ιουνίου και της 8ης Ιουλίου του 2013.

Όπως ήταν αναμενόμενο η συνάρτηση, που αφορά το πλήθος των ημερήσιων αναφορών των λέξεων, την 11η Ιουλίου αυξάνει απότομα φτάνοντας τις

E_{11}=40636 αναφορές[a] των δύο λέξεων, ενώ μέγιστο παρουσιάζεται την επόμενη μέρα, κατά τη διάρκεια της οποίας το θέμα γίνεται ακόμα πιο γνωστό, στην Ελλάδα αλλά και σε ολόκληρο τον κόσμο, με E_{12}=93826 αναφορές[b].

Στη συνέχεια ο μέσος όρος των 10 πρώτων ημερών της μελέτης (δηλαδή μέχρι και την προηγούμενη ημέρα από την ανακοίνωση του κλεισίματος) υπολογίστηκε ότι είναι ίσος με:

$$E_{1-10} = \frac{\sum_{i=1}^{10} E_i}{10} = 621,4$$

Επομένως η μεταβολή των αναφορών της 12ης Ιουλίου σε σχέση με τις πρώτες ημέρες υπολογίζεται:

$$\frac{E_{12} - E_{1-10}}{E_{1-10}} \cdot 100\% = 15000\%$$

Τέλος η συνολική μεταβολή των αναφορών του πρώτου δεκαήμερου του Ιουνίου, σε σχέση με το δεύτερο υπολογίζεται:

$$\frac{E_{11-20} - E_{1-10}}{E_{1-10}} \cdot 100\% = 5478\%$$

[a] ερτ: 19.248 , ert: 21.388

[b] ερτ: 35.770 , ert: 58.056

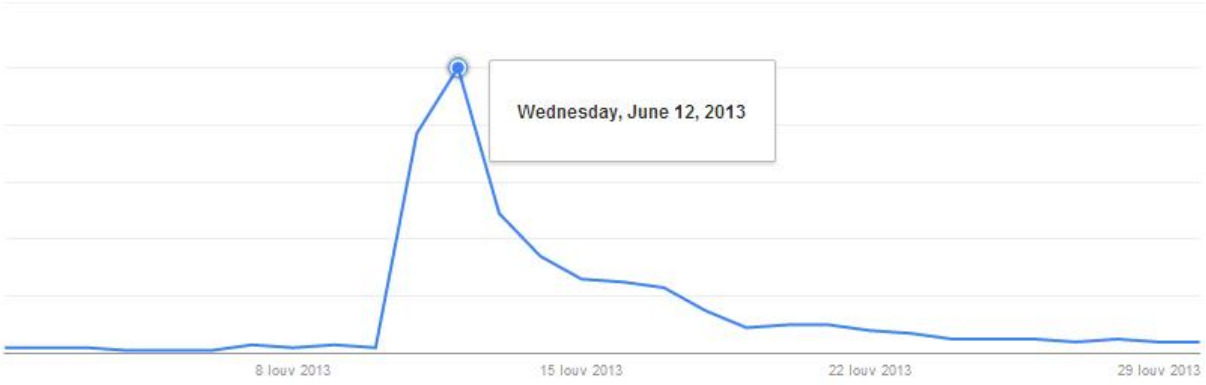
Α.4 Η ΕΡΤ στην αναζήτηση του Google

Όπως είναι γνωστό η Google μονοπωλεί το χώρο του διαδικτύου, μέσω των δεκάδων εφαρμογών της, οι οποίες όμως σαν πρώτο πυρήνα είχαν την μηχανή αναζήτησης, η οποία είναι αξιοσημείωτο ότι πολλαπλασιάζει καθημερινά τους επισκέπτες της ακόμα και σήμερα. Πιο συγκεκριμένα, σύμφωνα με τα επίσημα στοιχεία της εταιρείας (Statistic Brain, 2014), ο μέσος ημερήσιος αριθμός αναζητήσεων μέσα στο 2013 ξεπέρασε τα 5,9 δισεκατομμύρια, ενώ ο αντίστοιχος αριθμός που αφορά το 2012 ανήλθε στα 5,1 δισεκατομμύρια. Παράλληλα, προκειμένου να γίνει αντιληπτή η εκτόξευση των αριθμών της εταιρείας, αξίζει να τονιστεί ότι το 1998 όταν η Google γιόρταζε το πρώτο έτος ύπαρξής της, οι μέσες ημερήσιες αναζητήσεις δεν ξεπερνούσαν τις 9800.

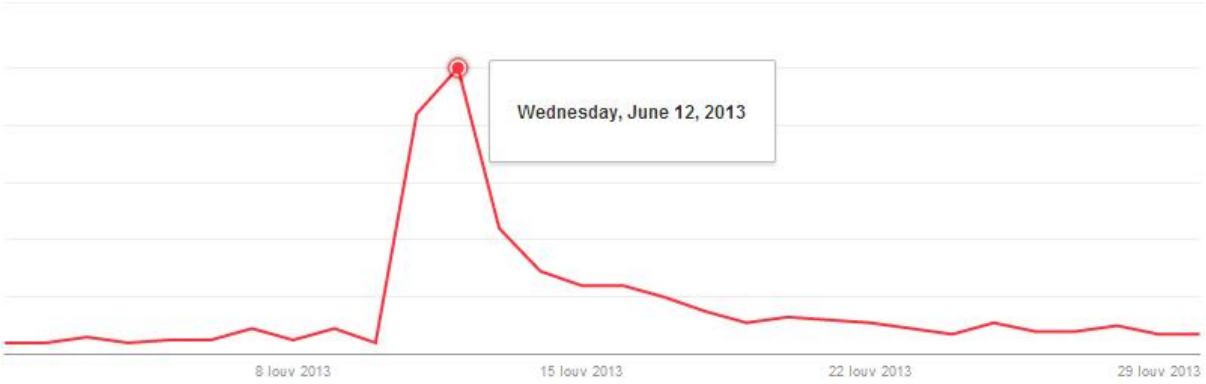
Έτσι ερωτηματικό που προκύπτει είναι αν η είδηση του κλεισίματος της ΕΡΤ, η οποία εκείνες τις μέρες συγκέντρωσε το ενδιαφέρον των πολιτών και των ΜΜΕ, τόσο εντός όσο και εκτός συνόρων, έχει αποτυπωθεί με όμοιο τρόπο και στα νούμερα που αφορούν τις αναζητήσεις του θέματος στο διαδίκτυο μέσω της δημοφιλέστερους μηχανής αναζήτησης που δεν είναι άλλη από αύτη της Google.

Προκειμένου να διαπιστωθεί η υπόθεση αυτή χρησιμοποιήθηκε η υπηρεσία Google Trends, η οποία αποτυπώνει συναρτήσει του χρόνου τη σχετική συχνότητα των αναζητήσεων μίας λέξης, ανά το συνολικό πλήθος αναζητήσεων. Το συγκεκριμένο εργαλείο της Google εμφανίζει αποτελέσματα από το 2004 μέχρι και σήμερα, ανανεώνοντας τα δεδομένα του ανά μία ώρα στα δημοφιλή θέματα και ημερησίως γενικότερα. Όπως έχει προαναφερθεί τα αποτελέσματα του Google Trends μπορούν να αξιοποιηθούν από ένα εξαιρετικά μεγάλο σύνολο φορέων και οργανισμών, όπως για παράδειγμα στην επιστήμη, στην υγεία, στην παιδεία, στην πολιτική και στις επιχειρήσεις.

Στα παρακάτω διαγράμματα ο οριζόντιος άξονας αναπαριστά τον χρόνο, ενώ ο κάθετος το πλήθος των αναζητήσεων μίας λέξης, προς το συνολικό πλήθος των αναζητήσεων που πραγματοποιήθηκαν. Η μελέτη αυτή αφορά τις ελληνικές ή τις παγκόσμιες αναζητήσεις, αναλόγως την περίπτωση.

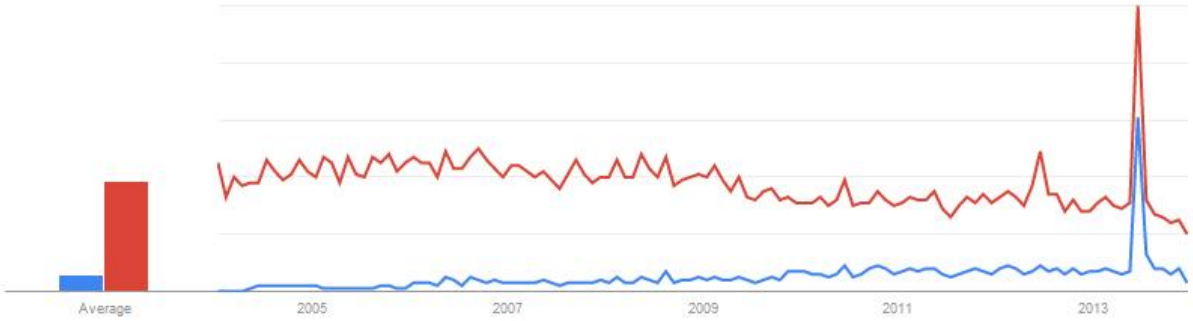


Διάγραμμα Α.4 - Αναζητήσεις της λέξης "ερτ" παγκοσμίως (Ιούνιος 2013)

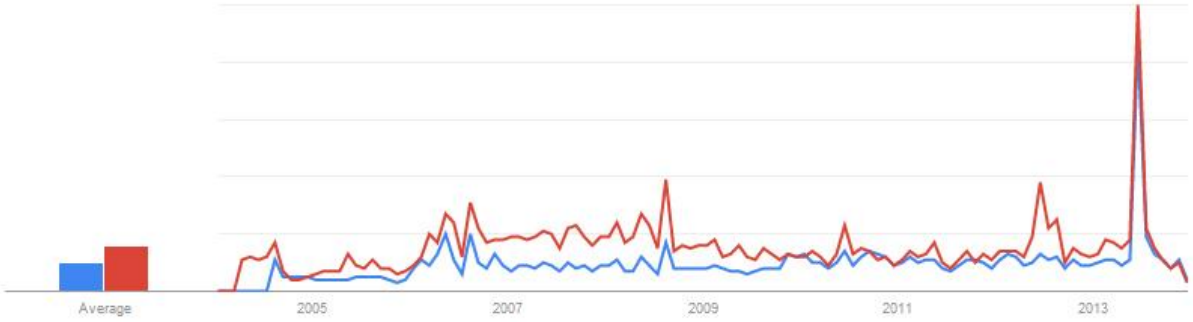


Διάγραμμα Α.5 - Αναζητήσεις της λέξης "ert" παγκοσμίως (Ιούνιος 2013)

Στα παραπάνω διαγράμματα εμφανίζονται οι παγκόσμιες αναζητήσεις των λέξεων "ερτ" και "ert" αντίστοιχα, κατά τη διάρκεια του Ιουνίου του 2013 (οι γραφικές παραστάσεις δεν απεικονίζουν το απόλυτο πλήθος των αναζητήσεων, αλλά το πλήθος τους σε σχέση με το σύνολο των παγκόσμιων αναζητήσεων). Αρχικά υπογραμμίζεται ότι τα γραφήματα είναι όμοια με αυτό που προέκυψε από τις μετρήσεις του Twitter (Βλ. προηγούμενη παράγραφο), εμφανίζοντας ίδια κατανομή και μέγιστα κατά τις ίδιες χρονικές περιόδους. Πιο συγκεκριμένα παρατηρείται μία αλματώδης αύξηση των αναζητήσεων, που αφορούν το ζήτημα της ΕΡΤ, την ημέρα που ανακοινώθηκε το κλείσιμό της (11η Ιουλίου). Εν συνεχεία εμφανίζουν και αυτές, έτσι ακριβώς όπως και οι αναφορές του θέματος στο Twitter, μέγιστο κατά την 12η του Ιουνίου. Τέλος αναφέρεται ότι η επικέντρωση της προσοχής της κοινωνίας, γύρω από το θέμα, αρχίζει να φθίνει περίπου μία εβδομάδα αργότερα, παραμένοντας ωστόσο υψηλότερη σε σύγκριση με το χαμηλό ενδιαφέρον πριν την ανακοίνωση και την εξάπλωση της είδησης.



Διάγραμμα Α.6 - Αναζητήσεις των λέξεων "ερτ" (μπλε) και "ert" (κόκκινο) παγκοσμίως (2004 – 2013)



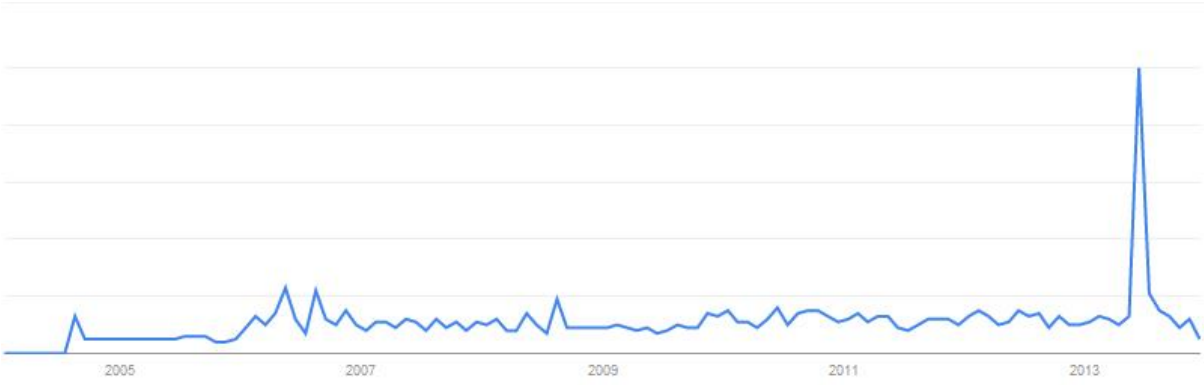
Διάγραμμα Α.7 - Αναζητήσεις των λέξεων "ερτ" (μπλε) και "ert" (κόκκινο) στην Ελλάδα (2004 – 2013)

Στα παραπάνω διαγράμματα παρουσιάζονται αντίστοιχα οι παγκόσμιες και οι ελληνικές αναζητήσεις των λέξεων "ερτ" (μπλε χρώμα) και "ert" (κόκκινο χρώμα), κατά τα έτη 2004 έως 2013 (δεν απεικονίζεται το απόλυτο πλήθος των αναζητήσεων, αλλά το πλήθος τους σε σχέση με το σύνολο των παγκόσμιων αναζητήσεων). Εύκολα παρατηρείται μία έντονη διαφοροποίηση μεταξύ των δύο διαγραμμάτων, καθώς στην παγκόσμια μελέτη η αναλογία της αναζήτησης της λέξης "ερτ" προς αυτή της λέξης "ert" είναι 1 προς 7,6 [a], ενώ στην ελληνική η αντίστοιχη αναλογία είναι 1 προς 1,7 [b]. Το γεγονός αυτό οφείλεται στο ότι πιθανότατα η λέξη "ert" χρησιμοποιείται σε άλλες γλώσσες[c], με αποτέλεσμα οι μετρήσεις αυτές να περιέχουν σφάλματα για τη συγκεκριμένη μελέτη τα οποία επιβάλλεται να αφαιρεθούν. Για το λόγο αυτό παρακάτω παρουσιάζονται και μελετώνται μονάχα οι ελληνικές αναζητήσεις που αφορούν τις δύο λέξεις.

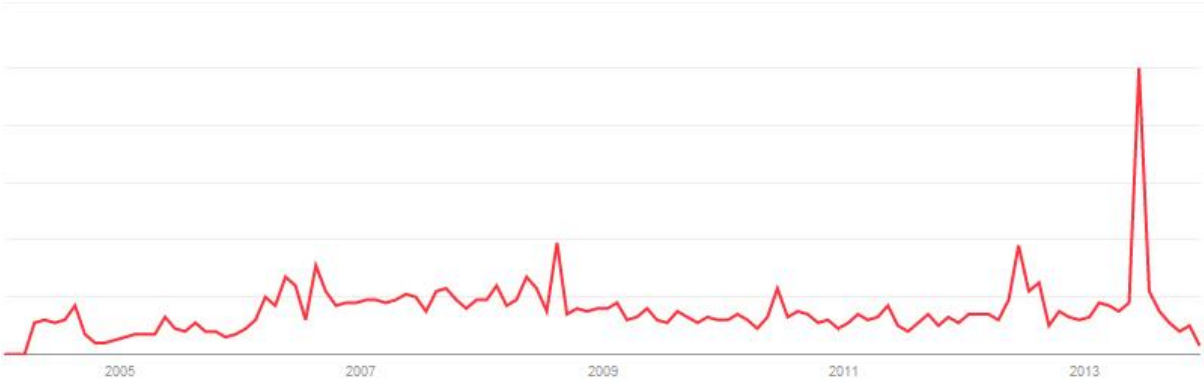
[a] Ανά μία αναζήτηση της λέξης "ερτ", έχουμε 7.6 αναζητήσεις της λέξης "ert".

[b] Ανά μία αναζήτηση της λέξης "ερτ", έχουμε 1.7 αναζητήσεις της λέξης "ert".

[c] Για παράδειγμα στα Ισλανδικά η λέξη "ert" είναι το δεύτερο ενικό του ρήματος "είμαι".



Διάγραμμα Α.8 - Αναζητήσεις της λέξης "ερτ" στην Ελλάδα (Ιανουάριος 2004 – Δεκέμβριος 2013)



Διάγραμμα Α.9 - Αναζητήσεις της λέξης "ert" στην Ελλάδα (Ιανουάριος 2004 – Δεκέμβριος 2013)

Στα παραπάνω διαγράμματα εμφανίζονται οι ελληνικές αναζητήσεις των λέξεων "ερτ" και "ert" αντίστοιχα, κατά τα έτη 2004 έως 2013 (οι γραφικές παραστάσεις δεν απεικονίζουν το απόλυτο πλήθος των αναζητήσεων, αλλά το πλήθος τους σε σχέση με το σύνολο των παγκόσμιων αναζητήσεων). Όπως ήταν αναμενόμενο παρατηρείται μία εκτόξευση των αναζητήσεων κατά τη διάρκεια του Ιουνίου του 2013, η οποία συνεπάγεται τη δραματική αύξηση του ενδιαφέροντος των Ελλήνων για το ζήτημα της ΕΡΤ. Επίσης σύμφωνα με τα στοιχεία της Google, σε σχέση με τον μήνα Μάιο, οι αναζητήσεις των λέξεων "ερτ" και "ert" στην Ελλάδα αυξήθηκαν κατά 700% και 500% αντίστοιχα κατά τη διάρκεια του Ιουνίου. Σε αυτό το σημείο αξίζει να σημειωθεί ότι οι δύο συγκεκριμένες μεταβολές βρέθηκαν στις δύο πρώτες θέσεις, ξεπερνώντας κατά πολύ την τρίτη πιο αυξανόμενη ελληνική αναζήτηση κατά τη διάρκεια του Ιουνίου, η οποία ήταν η "βάσεις 2012" με 150% (σημειώνεται ότι η συγκεκριμένη χρονική περίοδος βρίσκεται μετά την περίοδο των πανελλήνιων εξετάσεων).

Παράρτημα Β. Κώδικας για την αποτύπωση της οικονομικής κρίσης

Β.1 Κώδικας σε γλώσσα προγραμματισμού Java

```
import java.io.File;
import java.util.Scanner;
import java.io.FileNotFoundException;
import java.io.*;
import java.lang.*;
import java.util.*;
import java.io.BufferedReader;

public class twitter{

private static String path="C:/.../eutuxia.txt";
private int counter[];
private String hmeromhnies[];

private void hmeromhnies(){
hmeromhnies = new String[1652];
int mera=0;
while (mera<1652){
if (mera<31){
hmeromhnies[mera] = Integer.toString(mera+1)+" "+"Jan"+" "+"09";
}else if (mera<59){
hmeromhnies[mera] = Integer.toString(mera-31+1)+" "+"Feb"+" "+"09";
}else if (mera<90){
hmeromhnies[mera] = Integer.toString(mera-59+1)+" "+"Mar"+" "+"09";
}else if (mera<120){
hmeromhnies[mera] = Integer.toString(mera-90+1)+" "+"Apr"+" "+"09";
}else if (mera<151){
hmeromhnies[mera] = Integer.toString(mera-120+1)+" "+"May"+" "+"09";
}else if (mera<181){
hmeromhnies[mera] = Integer.toString(mera-151+1)+" "+"Jun"+" "+"09";
}else if (mera<212){
hmeromhnies[mera] = Integer.toString(mera-181+1)+" "+"Jul"+" "+"09";
}else if (mera<243){
hmeromhnies[mera] = Integer.toString(mera-212+1)+" "+"Aug"+" "+"09";
}else if (mera<273){
hmeromhnies[mera] = Integer.toString(mera-243+1)+" "+"Sep"+" "+"09";
}else if (mera<304){
hmeromhnies[mera] = Integer.toString(mera-273+1)+" "+"Oct"+" "+"09";
}else if (mera<334){
hmeromhnies[mera] = Integer.toString(mera-304+1)+" "+"Nov"+" "+"09";
}else if (mera<365){
hmeromhnies[mera] = Integer.toString(mera-334+1)+" "+"Dec"+" "+"09";
}else if (mera<396){
hmeromhnies[mera] = Integer.toString(mera-365+1)+" "+"Jan"+" "+"10";
}else if (mera<424){
hmeromhnies[mera] = Integer.toString(mera-396+1)+" "+"Feb"+" "+"10";
}else if (mera<455){
```

```
hmeromhnies[mera] = Integer.toString(mera-424+1)+" "+"Mar"+" "+"10";
}else if (mera<485){
hmeromhnies[mera] = Integer.toString(mera-455+1)+" "+"Apr"+" "+"10";
}else if (mera<516){
hmeromhnies[mera] = Integer.toString(mera-485+1)+" "+"May"+" "+"10";
}else if (mera<546){
hmeromhnies[mera] = Integer.toString(mera-516+1)+" "+"Jun"+" "+"10";
}else if (mera<577){
hmeromhnies[mera] = Integer.toString(mera-546+1)+" "+"Jul"+" "+"10";
}else if (mera<608){
hmeromhnies[mera] = Integer.toString(mera-577+1)+" "+"Aug"+" "+"10";
}else if (mera<638){
hmeromhnies[mera] = Integer.toString(mera-608+1)+" "+"Sep"+" "+"10";
}else if (mera<669){
hmeromhnies[mera] = Integer.toString(mera-638+1)+" "+"Oct"+" "+"10";
}else if (mera<699){
hmeromhnies[mera] = Integer.toString(mera-669+1)+" "+"Nov"+" "+"10";
}else if (mera<730){
hmeromhnies[mera] = Integer.toString(mera-699+1)+" "+"Dec"+" "+"10";
}else if (mera<761){
hmeromhnies[mera] = Integer.toString(mera-730+1)+" "+"Jan"+" "+"11";
}else if (mera<789){
hmeromhnies[mera] = Integer.toString(mera-761+1)+" "+"Feb"+" "+"11";
}else if (mera<820){
hmeromhnies[mera] = Integer.toString(mera-789+1)+" "+"Mar"+" "+"11";
}else if (mera<850){
hmeromhnies[mera] = Integer.toString(mera-820+1)+" "+"Apr"+" "+"11";
}else if (mera<881){
hmeromhnies[mera] = Integer.toString(mera-850+1)+" "+"May"+" "+"11";
}else if (mera<911){
hmeromhnies[mera] = Integer.toString(mera-881+1)+" "+"Jun"+" "+"11";
}else if (mera<942){
hmeromhnies[mera] = Integer.toString(mera-911+1)+" "+"Jul"+" "+"11";
}else if (mera<973){
hmeromhnies[mera] = Integer.toString(mera-942+1)+" "+"Aug"+" "+"11";
}else if (mera<1003){
hmeromhnies[mera] = Integer.toString(mera-973+1)+" "+"Sep"+" "+"11";
}else if (mera<1034){
hmeromhnies[mera] = Integer.toString(mera-1003+1)+" "+"Oct"+" "+"11";
}else if (mera<1064){
hmeromhnies[mera] = Integer.toString(mera-1034+1)+" "+"Nov"+" "+"11";
}else if (mera<1095){
hmeromhnies[mera] = Integer.toString(mera-1064+1)+" "+"Dec"+" "+"11";
}else if (mera<1126){
hmeromhnies[mera] = Integer.toString(mera-1095+1)+" "+"Jan"+" "+"12";
}else if (mera<1155){
hmeromhnies[mera] = Integer.toString(mera-1126+1)+" "+"Feb"+" "+"12";
}else if (mera<1186){
hmeromhnies[mera] = Integer.toString(mera-1155+1)+" "+"Mar"+" "+"12";
}else if (mera<1216){
hmeromhnies[mera] = Integer.toString(mera-1186+1)+" "+"Apr"+" "+"12";
}else if (mera<1247){
hmeromhnies[mera] = Integer.toString(mera-1216+1)+" "+"May"+" "+"12";
}else if (mera<1277){
hmeromhnies[mera] = Integer.toString(mera-1247+1)+" "+"Jun"+" "+"12";
}else if (mera<1308){
hmeromhnies[mera] = Integer.toString(mera-1277+1)+" "+"Jun"+" "+"12";
}else if (mera<1339){
hmeromhnies[mera] = Integer.toString(mera-1308+1)+" "+"Aug"+" "+"12";
}else if (mera<1369){
hmeromhnies[mera] = Integer.toString(mera-1339+1)+" "+"Sep"+" "+"12";
```

```
}else if (mera<1400){
hmeromhnies[mera] = Integer.toString(mera-1369+1)+" "+"Oct"+" "+"12";
}else if (mera<1430){
hmeromhnies[mera] = Integer.toString(mera-1400+1)+" "+"Nov"+" "+"12";
}else if (mera<1461){
hmeromhnies[mera] = Integer.toString(mera-1430+1)+" "+"Dec"+" "+"12";
}else if (mera<1492){
hmeromhnies[mera] = Integer.toString(mera-1461+1)+" "+"Jan"+" "+"13";
}else if (mera<1520){
hmeromhnies[mera] = Integer.toString(mera-1492+1)+" "+"Feb"+" "+"13";
}else if (mera<1551){
hmeromhnies[mera] = Integer.toString(mera-1520+1)+" "+"Mar"+" "+"13";
}else if (mera<1581){
hmeromhnies[mera] = Integer.toString(mera-1551+1)+" "+"Apr"+" "+"13";
}else if (mera<1612){
hmeromhnies[mera] = Integer.toString(mera-1581+1)+" "+"May"+" "+"13";
}else if (mera<1642){
hmeromhnies[mera] = Integer.toString(mera-1612+1)+" "+"Jun"+" "+"13";
}else if (mera<1652){
hmeromhnies[mera] = Integer.toString(mera-1642+1)+" "+"Jul"+" "+"13";
}mera++;}}

public void main() throws FileNotFoundException{
BufferedReader theFile = new BufferedReader(new FileReader(path));
Scanner input = new Scanner(theFile);
hmeromhnies();
counter = new int[1652];
for (int i=0;i<1652;i++){
counter[i]=0; }
String wrd="";
while (input.hasNextLine()==true){
wrd=input.nextLine();
for (int i=0; i<1652 ;i++){
if (wrd.contains(" "+hmeromhnies[i])){
counter[i]=counter[i]+1;
break;
}}}
for (int i=0; i<1652 ;i++){
System.out.println (counter[i]);
}}}
```

B.2 Δημιουργία διαγραμμάτων χρονοσειρών στην R

```
> #4.2
> data1<-read.csv(file.choose())
> par(mar=c(4.1,4.1,1.1,1.1))
> plot.ts (data1, col="green", axes=F, xlab="Date", ylab="Normalized
+ Emotion")
> axis(side=1, cex.axis=0.7,
+ at=c(1,182,366,547,731,912,1096,1278,1462,1643), labels=c("Jan 2009","Jul
+ 2009","Jan 2010","Jul 2010","Jan 2011","Jul 2011","Jan 2012","Jul
+ 2012","Jan 2013","Jul 2013"))
> axis(side=2, cex.axis=0.7, at=c(-2,0,2,4,6,8,10,12,14),
+ labels=c(-2,0,2,4,6,8,10,12,14))
> smooth<-filter(data1, filter = c(1/15, 1/15, 1/15, 1/15, 1/15, 1/15,
+ 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15), sides=2)
> lines (smooth)
> box()
> legend(500,15, c("Raw joy signal","15-day smoothed joy"), lty=c(1,1),
+ lwd=c(2.5,2.5), col=c("green","black"))
> text(1,6.5,"A",cex=0.7)
> text(45,2.5,"B",cex=0.7)
> text(108,3,"C",cex=0.7)
> text(227,1,"D",cex=0.7)
> text(366,15,"E",cex=0.7)
> text(410,2.5,"F",cex=0.7)
> text(459,1.5,"G",cex=0.7)
> text(592,1,"H",cex=0.7)
> text(731,9,"I",cex=0.7)
> text(775,2.3,"J",cex=0.7)
> text(843,2.5,"K",cex=0.7)
> text(957,2.2,"L",cex=0.7)
> text(1096,9,"M",cex=0.7)
> text(1140,4,"N",cex=0.7)
> text(1201,3.4,"O",cex=0.7)
> text(1323,2.8,"P",cex=0.7)
> text(1462,11.5,"Q",cex=0.7)
> text(1506,5.6,"R",cex=0.7)
> text(1586,3.5,"S",cex=0.7)
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
>
> #4.4
> data2<-read.csv(file.choose())
> par(mar=c(4.1,4.1,1.1,1.1))
> plot.ts (data2, col="green", axes=F, xlab="Date", ylab="Normalized
+ Emotion")
> axis(side=1, cex.axis=0.7,
+ at=c(1,182,366,547,731,912,1096,1278,1462,1643), labels=c("Jan 2009","Jul
+ 2009","Jan 2010","Jul 2010","Jan 2011","Jul 2011","Jan 2012","Jul
+ 2012","Jan 2013","Jul 2013"))
> axis(side=2, cex.axis=0.7, at=c(-2,0,2,4,6,8,10,12,14),
+ labels=c(-2,0,2,4,6,8,10,12,14))
> smooth<-filter(data1, filter = c(1/15, 1/15, 1/15, 1/15, 1/15, 1/15,
+ 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15), sides=2)
> lines (smooth)
> box()
> legend(-50,5.2, c("Raw disappointment signal","15-day smoothed
+ disappointment"), lty=c(1,1), lwd=c(2.5,2.5), col=c("green","black"))
> text(398,0.4,"A",cex=0.7)
```

```
> text(490,1.2,"B",cex=0.7)
> text(565,0.3,"C",cex=0.7)
> text(795,1.8,"D",cex=0.7)
> text(860,2.2,"E",cex=0.7)
> text(918,3.6,"F",cex=0.7)
> text(938,3.8,"G",cex=0.7)
> text(956,2.7,"H",cex=0.7)
> text(1033,3.2,"I",cex=0.7)
> text(1190,2.7,"J",cex=0.7)
> text(1444,2.6,"K",cex=0.7)
> text(1522,4,"L",cex=0.7)
> text(1545,2.6,"M",cex=0.7)
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
>
> #4.6
> data3<-read.csv(file.choose())
> par(mar=c(4.1,4.1,1.1,1.1))
> plot.ts (data3, col="green", axes=F, xlab="Date", ylab="Normalized
+ Emotion")
> axis(side=1, cex.axis=0.7,
+ at=c(1,182,366,547,731,912,1096,1278,1462,1643), labels=c("Jan 2009","Jul
+ 2009","Jan 2010","Jul 2010","Jan 2011","Jul 2011","Jan 2012","Jul
+ 2012","Jan 2013","Jul 2013"))
> axis(side=2, cex.axis=0.7, at=c(-2,0,2,4,6,8,10,12,14),
+ labels=c(-2,0,2,4,6,8,10,12,14))
> smooth<-filter(data1, filter = c(1/15, 1/15, 1/15, 1/15, 1/15, 1/15,
+ 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15), sides=2)
> lines (smooth)
> box()
> legend(-50,5.2, c("Raw anger signal","15-day smoothed anger"),
+ lty=c(1,1), lwd=c(2.5,2.5), col=c("green","black"))
> text(398,1.5,"A",cex=0.7)
> text(450,0.8,"B",cex=0.7)
> text(490,3,"C",cex=0.7)
> text(883,2.6,"D",cex=0.7)
> text(911,3.1,"E",cex=0.7)
> text(953,2.8,"F",cex=0.7)
> text(979,3.5,"G",cex=0.7)
> text(1033,4.2,"H",cex=0.7)
> text(1143,3.6,"I",cex=0.7)
> text(1223,2.6,"J",cex=0.7)
> text(1262,3.6,"K",cex=0.7)
> text(1304,1.9,"L",cex=0.7)
> text(1369,3.2,"M",cex=0.7)
> text(1414,2.8,"N",cex=0.7)
> text(1536,2.2,"O",cex=0.7)
> text(1569,1.8,"P",cex=0.7)
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
```

B.3 Συσχέτιση στην R

```
> install.packages('psych',.Library)
> library(psych)
> data1<-read.csv(file.choose())
> cor(data1$unemployment,data1$emotion,method="pearson")
[1] 0.8647381
> cor.test(data1$unemployment,data1$emotion,method="pearson")

        Pearson's product-moment correlation

data:  data1$unemployment and data1$emotion
t = 12.4162, df = 52, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7768519 0.9195772
sample estimates:
      cor
0.8647381
```

Παράρτημα Γ. Κώδικας απλού γραμμικού μοντέλου

Γ.1 Συσχέτιση στην R (για την A)

```
> install.packages('psych',.Library)
> library(psych)
> data1<-read.csv(file.choose())
> cor(data1$unemployment,data1$emotion,method="pearson")
[1] 0.5528085
> cor.test(data1$Xt,data1$Yt,method="pearson")

        Pearson's product-moment correlation

data:  data1$Xt and data1$Yt
t = 3.6936, df = 31, p-value = 0.0008492
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2585725 0.7531766
sample estimates:
      cor
0.5528085
```

Γ.2 Συσχέτιση στην R (για την H)

```
> install.packages('psych',.Library)
> library(psych)
> data1<-read.csv(file.choose())
> cor(data1$unemployment,data1$emotion,method="pearson")
[1] 0.6594438
> cor.test(data1$Xt,data1$Yt,method="pearson")

        Pearson's product-moment correlation

data:  data1$Xt and data1$Yt
t = 9.6891, df = 122, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5466915 0.7487075
sample estimates:
      cor
0.6594438
```

Γ.3 Εκτίμηση της συνάρτησης στην R

```
> data1<-read.csv(file.choose())
> mod1<-lm(formula = Yt ~ Xt, data = data1)
> par(mfrow=c(2,2))
> plot(mod1)
> summary(mod1)

Call:
lm(formula = Yt ~ Xt, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max 
-18170.9  -5734.4   -313.1   3868.2  23351.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3878.41    4233.26   0.916    0.361    
Xt            596.25      61.54   9.689   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9309 on 122 degrees of freedom
Multiple R-squared:  0.4349,	Adjusted R-squared:  0.4302 
F-statistic: 93.88 on 1 and 122 DF,  p-value: < 2.2e-16
```

Γ.4 Δημιουργία διαγραμμάτων στην R

```
> #6.13
> plot(scaled.res1, xlab="Time (Observation Order)", ylab="Standardized
+ Residual", main="Residuals vs. Order")
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
>
> #6.14
> plot(scaled.res1[-c(124)],scaled.res1[-c(1)], xlab="Standarized
+ Residual (t-1)", ylab="Standardized Residual (t)", main="Residuals(t) vs.
+ Residuals(t-1)")
> abline(h=0)
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
>
> #6.15
> plot(scaled.res1[-c(113,114,115,116,117,118,119,120,121,122,123,124)],
+ scaled.res1[-c(1,2,3,4,5,6,7,8,9,10,11,12)], xlab="Standarized
+ Residual (t-12)", ylab="Standardized Residual (t)", main="Residuals(t)
+ vs. Residuals(t-12)")
> abline(h=0)
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
>
> #6.17
> qqnorm(scaled.res1, main="Normal Probability Plot", xlab="Theoretical
+ Quantiles (Normal Scores)", ylab="Sample Quantiles (Standardized
+ Residuals)")
> qqline(scaled.res1)
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
```

Παράρτημα Δ. Κώδικας γενικού γραμμικού μοντέλου

Δ.1 Μέθοδος διαδοχικής αφαίρεσης στην R

```
> data2<-read.csv(file.choose())
> step(lm(Yt~X1+X2+X3+X12+X22+X32+X1X2+X1X3+X2X3,data=data2),
+ direction="backward", trace = FALSE)

Call:
lm(formula = Yt ~ X1 + X2 + X3 + X12 + X22 + X2X3, data = data2)

Coefficients:
(Intercept)           X1           X2           X3          X22
    44124.2       2643.1       2805.0       7363.9      -1918.8
       X2X3
     1827.4
```

Δ.2 Εκτίμηση της συνάρτησης στην R

```
> mod2<-lm(Yt~X1+X2+X3+X22+X2X3,data=data2)
> par(mfrow=c(2,2))
> plot(mod2)
> summary(mod2)

Call:
lm(formula = Yt ~ X1 + X2 + X3 + X22 + X2X3, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-13149.2  -2481.4   -165.3   2607.2  11727.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44595.7      557.9  79.939  < 2e-16 ***
X1            3059.1      512.5   5.969 2.57e-08 ***
X2            2651.9      640.5   4.141 6.54e-05 ***
X3            7208.1      604.1  11.931  < 2e-16 ***
X22          -1828.9      527.1  -3.470 0.000728 ***
X2X3          1851.5      587.5   3.152 0.002058 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4522 on 118 degrees of freedom
Multiple R-squared:  0.871,     Adjusted R-squared:  0.8655
F-statistic: 159.4 on 5 and 118 DF,  p-value: < 2.2e-16
```

Δ.3 Δημιουργία διαγραμμάτων στην R

```
> #6.18
> res2<-mod2$residuals
> scaled.res2<-scale(res2)
> fv2<-fitted.values(mod2)
> plot(fv2,scaled.res2, xlab="Fitted Value", ylab="Standardized Residual",
+ main="Residuals vs. Fitted")
> abline(h=0)
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
>
> #6.19
> plot(scaled.res2, xlab="Time (Observation Order)", ylab="Standardized
+ Residual", main="Residuals vs. Order")
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
>
> #6.20
> plot(scaled.res2[-c(124)],scaled.res2[-c(1)], xlab="Standarized
+ Residual (t-1)", ylab="Standardized Residual (t)", main="Residuals(t) vs.
+ Residuals(t-1)")
> abline(h=0)
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
>
> #6.21
> qqnorm(scaled.res2, main="Normal Probability Plot", xlab="Theoretical
+ Quantiles (Normal Scores)", ylab="Sample Quantiles (Standardized
+ Residuals)")
> qqline(scaled.res2)
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
>
> #6.22
> par(mfrow=c(2,3))
> plot(data2$X1, scaled.res2, xlab="Standardized X(t)", ylab="Standardized
+ Residual", main="Residuals vs. X(t)")
> abline(h=0)
> plot(data2$X2, scaled.res2, xlab="Standardized Y(t-1)",
+ ylab="Standardized Residual", main="Residuals vs. Y(t-1)")
> abline(h=0)
> plot(data2$X3, scaled.res2, xlab="Standardized Y(t-12)",
+ ylab="Standardized Residual", main="Residuals vs. Y(t-12)")
> abline(h=0)
> plot(data2$X22, scaled.res2, xlab="Standardized Y^2(t-1)",
+ ylab="Standardized Residual", main="Residuals vs. Y^2(t-1)")
> abline(h=0)
> plot(data2$X2X3, scaled.res2, xlab="Standardized Y(t-1)*Y(t-12)",
+ ylab="Standardized Residual", main="Residuals vs. Y(t-1)*Y(t-12)")
> abline(h=0)
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
>
> #6.24
> plot(cooks.distance(mod2),type="h", xlab="Observation Number",
+ ylab="Cook's Distance", main="Cook's Distance")
> savePlot(filename="C:/...", type=c("png"), res=144, device=dev.cur())
```

Δ.4 Έλεγχος για ύπαρξη πολυσυγγραμμικότητας

```
> library(car)
> vif(mod2)
      X1       X2       X3      X22     X2X3
1.592540 2.487086 2.213030 2.205525 2.406208
```

Δ.5 Πρόβλεψη πωλήσεων στην R

```
> new<-data.frame(X1=c(0.041387527), X2=c(1.801838178), X3=c(2.31884952),
X22=c(3.246620818), X2X3=c(4.178191593))
> predict(mod2, new, level=0.95, interval=c("confidence"))
       fit      lwr      upr
1 68013.72 64994.21 71033.24
> predict(mod2, new, level=0.95, interval=c("prediction"))
       fit      lwr      upr
1 68013.72 58562.98 77464.47
```

Βιβλιογραφία

24wro. (2012, December 15). Ανάκτηση από 24wro.eu: http://www.24wro.eu/2012/12/google_15.html

Acquia. (2014). Ανάκτηση από acquia.com: http://www.acquia.com/examples-big-data-projects

Ahlqvist, T., Asta, B., Halonen, M., & Heinonen, S. (2008). *Social media road maps exploring the futures triggered by social media.*

AIDSVu. (2010). Ανάκτηση από http://aidsvu.org/map/

Akiyama, T. (2003). The Continued Growth of Text Information: From an Analysis of Information Flow Censuses Taken furing the Past Twenty Years.

BBC News. (2011, August 15). *England's week of riots*.

Becker, H. (1986). *Can users really absorb data at today's rates? Tomorrow's?*

Bohn, R. E., & Short, J. E. (2009). *How Much Information? 2009 Report on American Consumers.*

Bounie, D., & Gille, L. (2012). *International Production and Dissemination of Information: Results, Methodological Issues, and Statistical Perspectives.*

Childress, D. (2008). *Johannes Gutenberg and the Printing Press.*

Cisco. (2008). *Cisco Visual Networking Index – Forecast and Methodology, 2007–2012.*

Coffman, K., & Odlyzko, A. (1998). *The Size and Growth Rate of the Internet.*

Cortada, J. W. (1993). *Before The Computer.*

Cox, M., & Ellsworth, D. (1997). *Application-controlled demand paging for out-of-core visualization.*

Cukier, K. (2010, Feb 25). *Data, data everywhere*. Ανάκτηση από Economist: http://www.economist.com/node/15557443

Draper, N., & Smith, H. (1997). *Εφαρμοσμένη Ανάλυση Παλινδρόμησης.* Εκδόσεις Παπαζήση.

Economist. (2011, August 15). *The BlackBerry riots*. Ανάκτηση από economist.com: http://www.economist.com/node/21525976

Frank, B. H. (2013, Aug 9). Ανάκτηση από http://www.geekwire.com/2013/windermere-partners-inrix-provide-drive-time-recommendations-homebuyers/

Gantz, J. F., Reinsel, D., Chute, C., Schlichting, W., McArthur, J., Minton, S., . . . Alex, M. (2007). *The Expanding Digital Universe: A Forecast of Worldwide Information Growth through 2010.*

Hearst, M. (2003). *What Is Text Mining?*

Hemsoth, N. (2013, Mar 16). Ανάκτηση από http://www.datanami.com/datanami/2013-03-16/how_ford_is_putting_hadoop_pedal_to_the_metal.html

Hilbert. (2012). How to Measure "How Much Information"? Theoretical, Methodological, and Statistical Challenges for the Social Sciences.

Hilbert, M., & Lopez, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*.

HIV Surveillance Report. (2011). Diagnoses of HIV Infection in the United States and Dependent Areas. *23*.

Hyundai News. (2014). Ανάκτηση από hyundainews.com: http://www.hyundainews.com/us/en-us/Corporate/SalesReleases/PressReleases.aspx

In. (2013, Jun 11). Ανάκτηση από in.gr: http://news.in.gr/greece/article/?aid=1231252896

James, I. (2009). Claude Elwood Shannon 30 April 1916 - 24 February 2001. *Biographical Memoirs of Fellows of the Royal Society, 55*.

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety.*

Larose, D. T. (2004). *Discovering Knowledge in Data: An Introduction to Data Mining.*

Lesk, M. (1997). *How much information is there in the world?*

Lyman, P., & Varian, H. R. (2000). *How Much Information?*

McCue, T. (2013, Jan 1). Ανάκτηση από http://www.forbes.com/sites/tjmccue/2013/01/01/108-mpg-with-ford-fusion-energi-plus-25-gigabytes-of-data/

Messina, C. (2008, Jan 2). Ανάκτηση από http://twitter.pbworks.com/w/page/1779812/Hashtags

Moe, W. W., & Fader, P. S. (2004). *Capturing Evolving Visit Behavior in Clickstream Data.*

Morris, R., & Truskowski, B. (2003). The Evolution of Storage Systems. *IBM Systems Journal, 42*(2).

Murphy, C. (2012, Aug 8). Ανάκτηση από http://www.informationweek.com/global-cio/interviews/union-pacific-delivers-internet-of-thing/240004930

Naisbitt, J. (1982). *Megatrends.*

Neuman, W. R., Park, Y. J., & Panek, E. (2012). *Tracking the Flow of Information into the Home: An Empirical Assessment of the Digital Revolution in the United States, 1960-2005.*

News247. (2011, May 10). Ανάκτηση από news247.gr: http://news247.gr/eidiseis/politiki/fonos_kentro_athhnas.858371.html

News247. (2011, September 6). Ανάκτηση από news247.gr: http://news247.gr/eidiseis/politiki/dramatikes_apofaseis_ypo_thn_piesh_ths_troikas.1357730.html

Newsbomb. (2012, October 3). Ανάκτηση από newsbomb.gr: http://www.newsbomb.gr/politikh/story/239865/lista-lagkarnt-i-periergi-diadromi-toy-amartoloy-cd

Newsit. (2013, Jun 12). Ανάκτηση από newsit.gr: http://www.newsit.gr/default.php?pname=Article&art_id=215027&catid=12

O'Reilly, T. (2012). *Big Data Now.*

Pool, I. d. (1983). Tracking the Flow of Information. *Science, 221*(4611).

Price, D. (1961). *Science Since Babylon.*

Public Issue. (2012, May 13). Ανάκτηση από publicissue.gr: http://www.publicissue.gr/2005/ekloges-2012-telos-metapoliteftikou-dikommatismou/

Radio Netherlands Worldwide. (2013, March 6). Ανάκτηση από rnw.nl: http://www.rnw.nl/english/article/ch%C3%A1vezs-revolution-lives-social-media

Rider, F. (1944). *The Scholar and the Future of the Research Library. A Problem and Its Solution.*

Rivero, E. (2014, February 26). *UCLA Newsroom*. Ανάκτηση από http://newsroom.ucla.edu/releases/twitter-big-data-can-be-used-to-250162

RTE News. (2013, March 17). Ανάκτηση από rte.ie: http://www.rte.ie/news/2013/0316/376965-cyprus-bailout/

SchweserNotes. (2013). *Ethical and Professional Standars, Quantitative Methods, and Economics.* Kaplan Schweser.

Smith, C. (2013, Aug 25). By the-numbers: 17 Amazing Facebook User & Demographic Statistics. Ανάκτηση από http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/

Statistic Brain. (2014, Jan 1). *Google Annual Search Statistics*. Ανάκτηση από http://www.statisticbrain.com/google-searches/

Stross, R. (2010, May 1). When History Is Compiled 140 Characters at a Time. *The New York Times*. Ανάκτηση από

http://www.nytimes.com/2010/05/02/business/02digi.html?scp=1&sq=twitter%20+%20history&st=cse&_r=1&

Swanson, B., & Gilder, G. (2008). *Estimating the Exaflood.*

The Best News. (2010, February 2). Ανάκτηση από thebest.gr: http://www.thebest.gr/news/index/viewStory/8812

The Telegraph. (2011, July 22). Ανάκτηση από telegraph.co.uk: http://www.telegraph.co.uk/news/worldnews/europe/norway/9495025/Norway-massacre-A-timeline-of-the-attacks-that-horrified-a-nation.html

Tonkin, E., Pfeiffer, H. D., & Tourte, G. (2012). *Twitter, information sharing and the London riots?* doi:10.1002/bult.2012.1720380212

Tvxs. (2013, March 21). Ανάκτηση από tvxs.gr: http://tvxs.gr/news/%CE%B5%CE%BB%CE%BB%CE%AC%CE%B4%CE%B1/%CE%B4%CF%85%CE%BF-%CF%86%CE%BF%CF%81%CE%AD%CF%82-%CE%B9%CF%83%CF%8C%CE%B2%CE%B9%CE%B1-%CE%B3%CE%B9%CE%B1-%CF%84%CE%BF-%C2%AB%CE%BC%CE%B1%CE%BA%CE%B5%CE%BB%CE%B5%CE%B9%CF%8C-%CF%83%CF%84%CE%BF%CF%85

Tvxs. (2013, March 6). Ανάκτηση από tvxs.gr: http://tvxs.gr/news/%CE%BA%CF%8C%CF%83%CE%BC%CE%BF%CF%82/%C2%AB%CE%BA%CF%81%CE%AF%CF%83%CE%B9%CE%BC%CE%B5%CF%82-%CF%8E%CF%81%CE%B5%CF%82%C2%BB-%CE%B3%CE%B9%CE%B1-%CF%84%CF%83%CE%AC%CE%B2%CE%B5%CF%82-%CE%BA%CE%B1%CE%B9-%CE%B2%CE%B5%CE%BD%CE%B5%CE%B6%CE%BF%

Tvxs. (2013, April 18). Ανάκτηση από tvxs.gr: http://tvxs.gr/news/%CE%B5%CE%BB%CE%BB%CE%AC%CE%B4%CE%B1/%CE%BC%CE%B1%CE%BD%CF%89%CE%BB%CE%AC%CE%B4%CE%B1-%CE%B7%CE%BB%CE%B5%CE%AF%CE%B1%CF%82-%CE%B5%CF%80%CE%B9%CF%83%CF%84%CE%AC%CF%84%CE%B7%CF%82-%CE%AC%CE%BD%CE%BF%CE%B9%CE%BE%CE%B5-%CF%80%CF%85%CF%81-%

Tvxs. (2013, April 18). Ανάκτηση από tvxs.gr: http://tvxs.gr/news/%CE%B5%CE%BB%CE%BB%CE%AC%CE%B4%CE%B1/%CE%BC%CF%80%CE%BF%CF%8A%CE%BA%CE%BF%CF%84%CE%AC%CE%B6-%CF%83%CF%84%CE%B9%CF%82-%C2%AB%CE%BC%CE%B1%CF%84%CF%89%CE%BC%CE%AD%CE%BD%CE%B5%CF%82-%CF%86%CF%81%CE%AC%CE%BF%CF%85%CE%BB%CE%B5%CF%82%C2%BB-%C

Tvxs. (2014, March 1). Ανάκτηση από tvxs.gr: http://tvxs.gr/news/%CE%B5%CE%BB%CE%BB%CE%AC%CE%B4%CE%B1/%CF%83%CE%B5-%CE%BA%CF%81%CE%AF%CF%83%CE%B9%CE%BC%CE%B7-%CE%BA%CE%B1%CF%84%CE%AC%CF%83%CF%84%CE%B1%CF%83%CE%B7-%CE%BF%CE%B9-%CF%84%CF%81%CE%B5%CE%B9%CF%82-%CF%86%CE%BF%CE%B9%CF%84%CE%B7%CF%84%CE%AD%

Visible Intelligence. (2012). Ανάκτηση από www.visibletechnologies.com

Volkswagen Group. (2014). Ανάκτηση από volkswagenag.com: http://www.volkswagenag.com/content/vwcorp/content/de/misc/search.html?_charset_=UTF-8&q=report

W. P. Carey School of Business. (2012, May 8). Ανάκτηση από knowwpcarey: http://knowwpcarey.com/article.cfm?cid=25&aid=1171

Έθνος. (2010, April 28).

Έθνος. (2010, January 21).

Έθνος. (2012, April 4). Ανάκτηση από ethnos.gr: http://www.ethnos.gr/article.asp?catid=22768&subid=2&pubid=63640008

Ελευθεροτυπία. (2010, March 26). Ανάκτηση από enet.gr: http://www.enet.gr/?i=news.el.article&id=145379

Ελευθεροτυπία. (2010, February 9).

Ελευθεροτυπία. (2011, July 11). Ανάκτηση από enet.gr: http://www.enet.gr/?i=news.el.kosmos&id=292175

Ελευθεροτυπία. (2011, June 29). Ανάκτηση από enet.gr: http://www.enet.gr/?i=news.el.politikh&id=288805

Ελευθεροτυπία. (2011, October 28). Ανάκτηση από enet.gr: http://www.enet.gr/?i=news.el.article&id=321809

Ελληνική Στατιστική Αρχή. (2013). Ανάκτηση από statistics.gr: http://www.statistics.gr/

Έρευνα της HP. (2013, Jun 18). Αποτυγχάνουμε να αντλήσουμε γνώση από τα big data, συνεχίζουμε την εξόρυξη. Ανάκτηση από tech.in.gr: http://tech.in.gr/analysis/article/?aid=1231253504

Ζώτου, Έ. (2010, July 28). Ανάκτηση από Η Αυγή: http://archive.avgi.gr/ArticleActionshow.action?articleID=558543

Η Καθημερινή. (2012, May 19). Ανάκτηση από kathimerini.gr: http://www.kathimerini.gr/1573/article/epikairothta/politikh/ekloges-stis-17-ioynioy

Οικονόμου, Π., & Καρώνη, Χ. (2010). *Στατιστικά Μοντέλα Παλινδρόμησης.* Αθήνα: Εκδόσεις Συμεών.

Πηλίδης, Κ. (2012, Dec 28). Ανάκτηση από http://www.autoblog.gr/2012/12/28/h-ford-pisteyei-oti-oi-aisthhthres-sta-aytokinhta-exoyn-mellon-mhnythhke-logw-twn-yperbolika-aisiodokswn-noymerwn-katanalwshs-sta-ybridika-ths/

Πρώτο Θέμα. (2012, November 1). Ανάκτηση από protothema.gr: http://www.protothema.gr/economy/article/233469/sta-18_9-dis-eyro-o-logariasmos-ths-epimhkynshs/

Πρώτο Θέμα. (2013, Jun 11). Ανάκτηση από protothema.gr: http://www.protothema.gr/greece/article/285571/to-kleisimo-ths-ert-sto-twitter/

ΣΚΑΪ. (2012, December 15). Ανάκτηση από skai.gr: http://www.skai.gr/news/world/article/219628/-to-hroniko-tou-makeleiou-sto-konektikat/

Τα Νέα. (2010, July 27).

Το Βήμα. (2011, October 27). Ανάκτηση από tovima.gr: http://www.tovima.gr/finance/article/?aid=427166

Το Βήμα. (2012, February 12). Ανάκτηση από tovima.gr: http://www.tovima.gr/politics/article/?aid=443281&h1=true