ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

ΕΡΓΑΣΙΑ 2: ΟΜΑΔΟΠΟΙΗΣΗ

ΖΑΜΠΟΚΑΣ ΓΙΩΡΓΟΣ	**ΑΕΜ: 7173**	**zampokas@auth.gr**
ΜΟΥΡΟΥΖΗ ΧΡΙΣΤΟΣ	**ΑΕΜ: 7571**	**chrimour@auth.gr**
ΜΠΟΣΔΕΛΕΚΙΔΗΣ ΒΑΣΙΛΗΣ	**ΑΕΜ: 7488**	**vmposde@auth.gr**

Table of Contents

Περιγραφή του προβλήματος

Στα πλαίσια της εργασίας αυτής κληθήκαμε να ομαδοποιήσουμε τις διάφορες εγκαταστάσεις μιας οικίας βάσει της ενεργειακής τους κατανάλωσης ανάλογα με τις διάφορες χρονικές περιόδους κάποιων ετών. Πιο συγκεκριμένα, από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption μας παρέχεται ένα dataset με στοιχεία για την ενεργειακή κατανάλωση μια οικίας για περίπου 4 χρόνια (2007 - 2010, οι εγγραφές για τις 15 μέρες Δεκεμβρίου του 2006 απορρίφθηκαν ως μη αντιπροσωπευτικές). Ως χαρακτηριστικά σε κάθε εγγραφή διατίθενται η πλήρης ημερομηνία, η ώρα και το λεπτό, η γενική ενεργή ισχύς της οικίας (kilowatt), η γενική άεργος ισχύς της οικιας( ), η τάση (V), η γενική ένταση του ρεύματος(A), και η ενεργειακή κατανάλωση σε watt-hour ενεργής ισχύος για τρεις διαφορετικές εγκαταστάσεις ως εξής: η πρώτη εγκατάσταση είναι η κουζίνα (αποτελούμενη από ένα πλυντήριο πιάτων, έναν φούρνο και έναν φούρνο μικροκυμάτων), η δεύτερη είναι το πλυσταριό (αποτελούμενο από ένα πλυντήριο, ένα στεγνωτήριο, ένα ψυγείο και μια λάμπα) και η τρίτη αποτελείται από ένα ηλεκτρικό θερμοσίφωνα και ένα κλιματιστικό.

Δομή της εργασίας

Στην επόμενη ενότητα ξεκινάμε με την περιγραφή της προεπεξεργασίας που κάναμε στο dataset, έτσι ώστε να είναι και πιο εύκολα διαχειρίσιμο αλλά και να εξαχθούν κάποια νέα πιο χρήσιμα χαρακτηριστικά για την ομαδοποίηση. Στην συνέχεια για κάθε εγκατάσταση χρησιμοποιώντας διάφορους αλγορίθμους, ομαδοποιούμε τα δεδομένα που αφορούν συγκεκριμένες χρονικές περιόδους. Μετά, θέλοντας να εξάγουμε πληροφορία και για την συνολική κατανάλωση της οικίας σε βάθος χρόνου, ομαδοποιούμε και τα δεδομένα για την γενική ενεργή ισχύ και για πολλές διαφορετικές χρονικές περιόδους. Μετά την ολοκλήρωση των πειραμάτων για τις διαφορετικές ομαδοποιήσεις, συγκρίνονται οι διάφοροι αλγόριθμοι ως προς την ποιότητα της ομαδοποίησης και εξάγονται χρήσιμα συμπεράσματα. Ένα γενικό συμπέρασμα ακολουθεί στην τελευταία ενότητα.

Προεπεξεργασία των δεδομένων

Το αρχικό dataset διέρχεται αρχικά από το στάδιο της προεπεξεργασίας. Πρώτα, για πιο εύκολη διαχείριση το αρχικό dataset χωρίζεται σε πολλά μικρά αρχεία (με ορισμένο αριθμό γραμμών το καθένα) χρησιμοποιώντας μια εντολή split στο τερματικό.

Έχοντας αυτά τα αρχεία σε ένα φάκελο “household”, μετά για αυτά (που ουσιαστικά είναι ακριβώς το αρχικό dataset) δημιουργούνται δώδεκα νέα αρχεία ανάλογα με την εποχή του χρόνου και την χρονική ζώνη της ημέρας. Σημειώνεται πως κάθε έτος έχει χωριστεί σε τέσσερις εποχές (χειμώνας: μήνες Δεκέμβριος-Φεβρουάριος, άνοιξη: μήνες Μάρτιος-Μάιος, καλοκαίρι: μήνες Ιούνιος-Αύγουστος, φθινόπωρο: μήνες Σεπτέμβριος-Νοέμβριος) και κάθε μέρα σε τρεις χρονικές ζώνες(πρωί: 8.00-15.00. απόγευμα: 15.00: 23.00, βράδυ: 23.00-8.00). Αυτό γίνεται τρέχοντας το script “split_epochs_timezones” από τον φάκελο “Preprocessing_Scripts” .

Συνεχίζοντας, ανακαλύπτονται αρνητικές τιμές, τιμές που λείπουν και μη λογικές μηδενικές τιμές και αφαιρούνται οι εγγραφές που έχουν χαρακτηριστικά με τέτοιες τιμές. Αυτό επιτυγχάνεται με το script “clean_from_zeros_negatives” το οποίο δέχεται τα δώδεκα αρχεία από έναν φάκελο “epochs_timezones” και κρατά μόνο τις εγγραφές που έχουν μη αρνητική και μη μηδενική ενεργό ισχύ, μη αρνητική άεργο ισχύ, μη αρνητική και μη μηδενική τάση, μη αρνητική και μη μηδενική ένταση ρεύματος και μη αρνητικές μετρήσεις για τις τρεις εγκαταστάσεις

Σε επόμενο στάδιο εξομαλύνονται ακραίες τιμές στα δώδεκα αρχεία που προέκυψαν. Συγκεκριμένα, τρέχοντας το script “remove_extreme” εντοπίζονται, ανάλογα με κάποια thresholds (βλ. script) , τεράστιες αποκλίσεις σε κάποια από τις μετρήσεις κάποιων εγγραφών από τον μέσο όρο των μετρήσεων για το συγκεκριμένο attribute και για το ανάλογο αρχείο, και εξομαλύνονται αναλόγως. Στα παρόντα thresholds καταλήξαμε μετά από εκτενή πειραματισμό

Στο μικρότερο και εξομαλυμένο dataset που έχει προκύψει, εφαρμόζουμε το script “merge_data”, το οποίο συγχωνεύει εγγραφές (βρίσκοντας μέσους όρους) ανά συγκεκριμένο αριθμό λεπτών (στην εκτέλεση της εργασίας ανά 5 λεπτά) αν και εφόσον οι εγγραφές που συγχωνεύονται δεν απέχουν χρονικά μεταξύ τους πάνω από 15 λεπτά.

Στο τελικό στάδιο της προεπεξεργασίας και από τα δώδεκα αρχεία που έχουν προκύψει από τις προηγούμενες διεργασίες, υπολογίζονται χρήσιμα μεγέθη για τα χαρακτηριστικά Συγκεκριμένα, στο script “calculate_avg_sum_max”, υπολογίζονται μέση όροι, αθροίσματα και μέγιστα στα 7 χαρακτηριστικά, ενώ επιλέγεται ως προς πιο χρονικό μέγεθος κατηγοριοποιούνται Δηλαδή, στα αρχεία που προκύπτουν για το κάθε μέγεθος οι γραμμές θα αναφέρονται είτε σε συνδυασμό εποχής και χρονικής ζώνης (απόγευμα φθινοπώρου, πρωί φθινοπώρου κλπ), είτε σε συνδυασμό έτους, εποχής και χρονικής ζώνης (απόγευμα φθινοπώρου 2007, απόγευμα φθινοπώρου 2008 κλπ), είτε σε συνδυασμό έτους και εποχής (χειμώνας 2007, άνοιξη 2007 κλπ). Τα δεδομένα που χρησιμοποιήθηκαν για τις ομαδοποιήσεις βασίζονται στο διαχωρισμό σε συνδυασμό έτους, εποχής και χρονικής ζώνης.

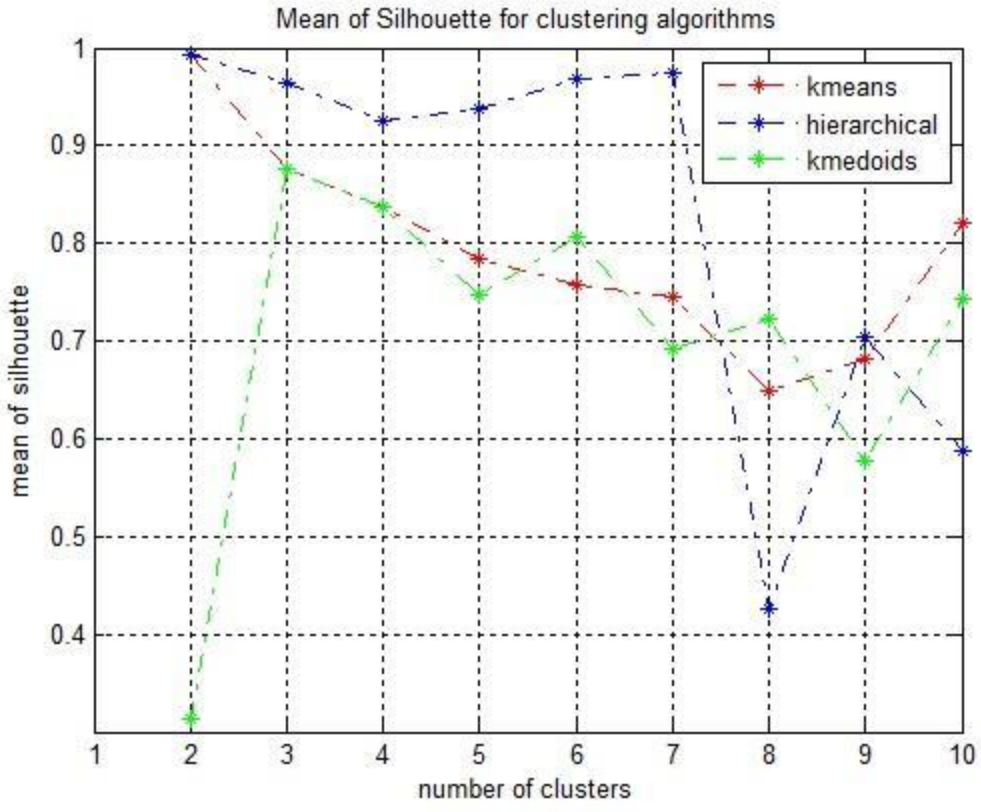
Εκτέλεση αλγορίθμων ομαδοποίησης

Αρχικά, η ομαδοποίηση γίνεται βάσει χρονικής περιόδου για κάθε εγκατάσταση , λαμβάνοντας υπόψη μ.ο. κατανάλωσης, μέγιστη κατανάλωση και άθροισμα κατανάλωσης για κάθε χρονική περίοδο.

ΚΟΥΖΙΝΑ

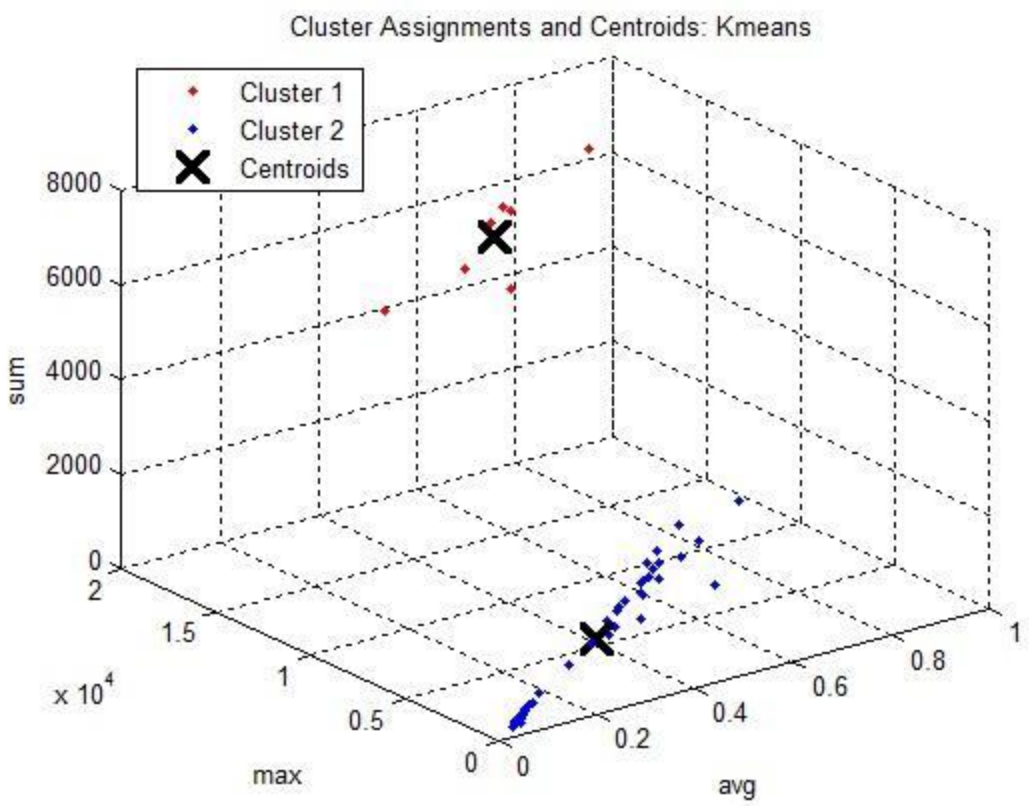
Παρακάτω φαίνεται η ομαδοποίηση για την κατανάλωση ηλεκτρικού ρεύματος από συσκευές που βρίσκονται στην **ΚΟΥΖΙΝΑ.**

Χρησιμοποιήθηκαν οι αλγόριθμοι KMeans, Hierarchical, KMedoids με διάφορες παραμέτρους. Τα καλύτερα αποτέλεσμα όπως φαίνεται και από το διάγραμμα μέσου όρου της μετρικής sihlouette είναι για **clusters=2 (KMeans, Hierarchical)** ή για **clusters=3(KMedoids)**



KMeans

- **Clusters=2**



Sihlouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.9881	715591249.531623	13836715.1102082	8	Cluster 1
0.9933	3577956247.65811	47077687.9340664	40	Cluster 2

Mean Sihlouette = 0.9924

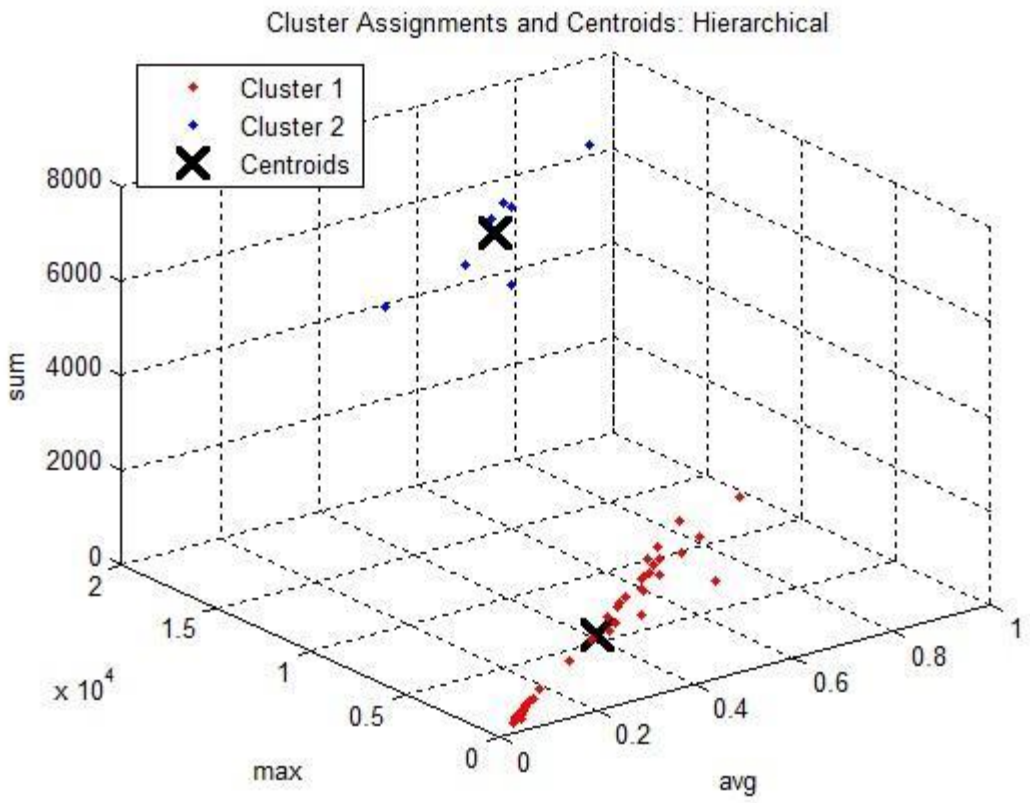
Mean Cohesion = 3.0457e+007

Mean Separation = 2.1468e+009

SSE = 6.0914e+007

Hierarchical

- **Clusters=2**



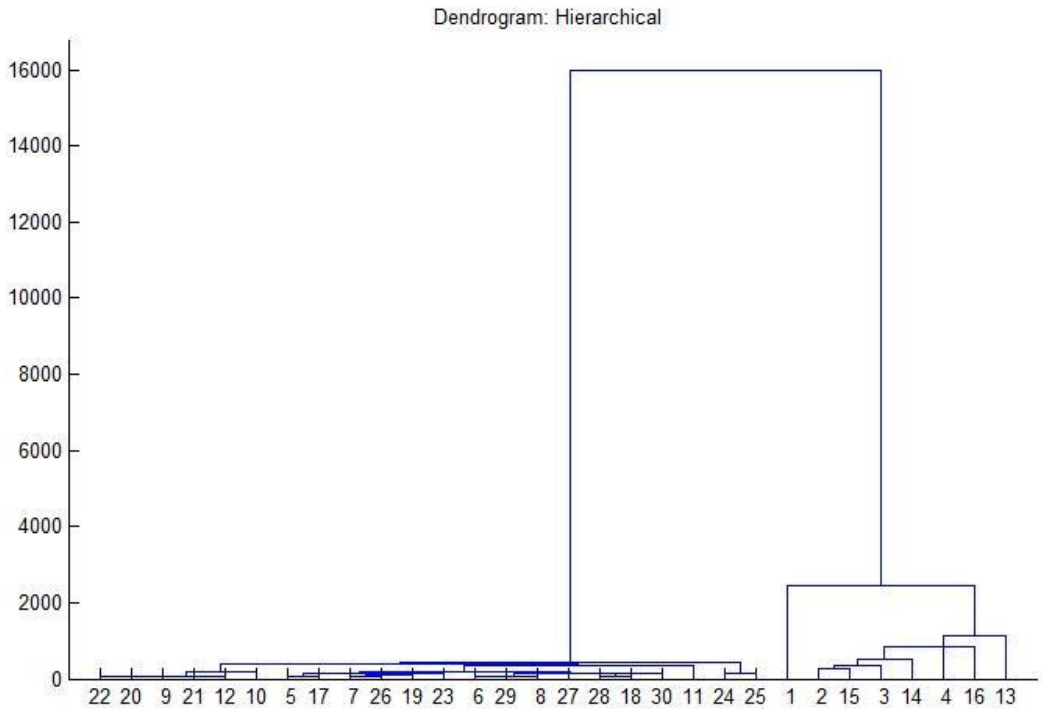
Sihlouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.9933	3577956247.65811	47077687.9340664	8	Cluster 1
0.9881	715591249.531623	13836715.1102082	40	Cluster 2

Mean Sihlouette = 0.9924

Mean Cohesion = 3.0457e+007

Mean Separation = 2.1468e+009

SSE = 6.0914e+007



Δενδόγραμμα Ιεραρχικού Αλγόριθμου

Μέθοδος Ward:

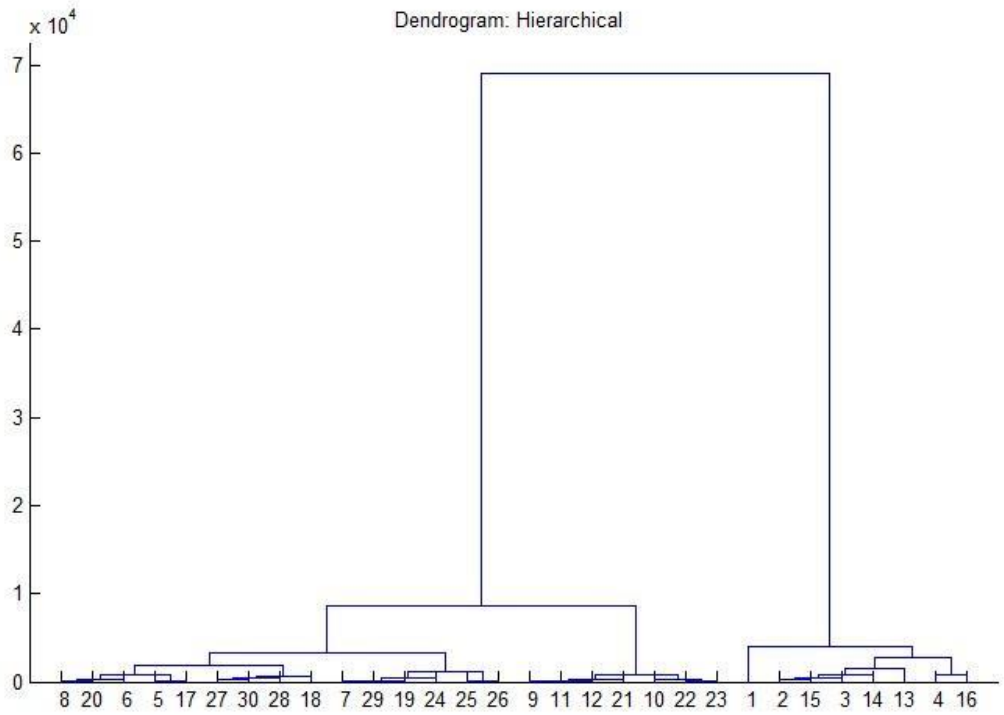
Sihlouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.9881	715591249.531623	13836715.1102082	8	Cluster 1
0.9933	3577956247.65811	47077687.9340664	40	Cluster 2

Mean Sihlouette = 0.9924

Mean Cohesion = 3.0457e+007

Mean Separation = 2.1468e+009

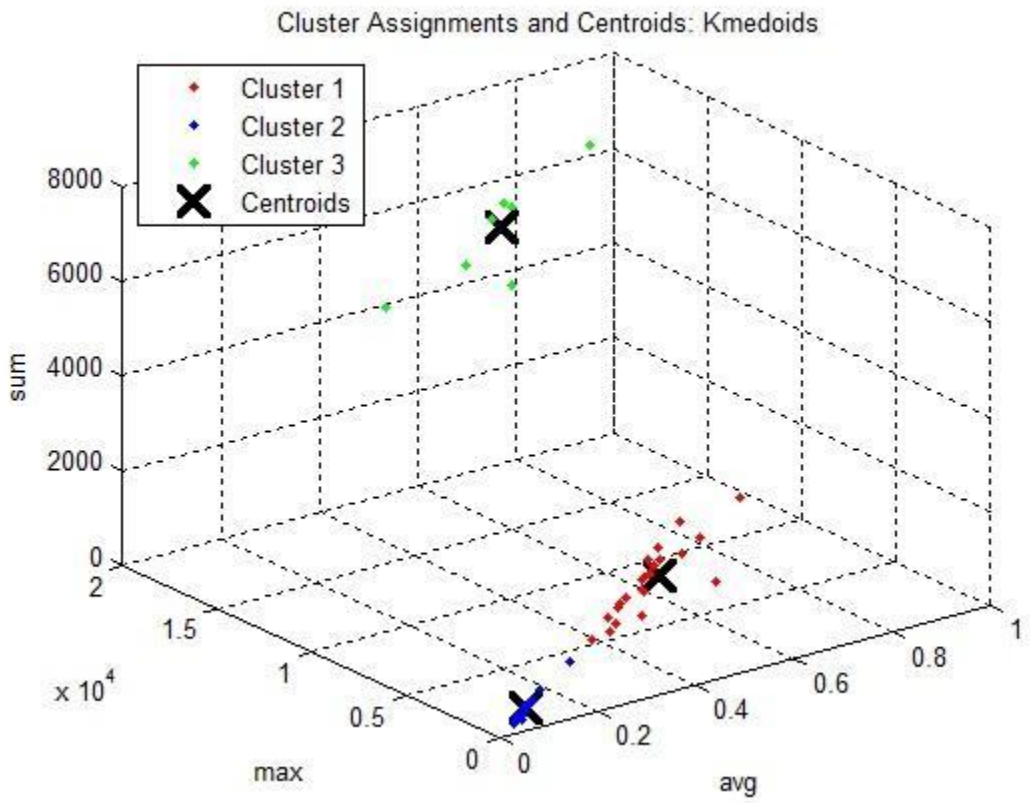
SSE = 6.0914e+007



Δενδόγραμμα Ιεραρχικού Αλγόριθμου με μέθοδο Ward

KMedoids

- **Clusters=3**



Sihlouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.7729	860837652.838354	7455588.43523182	23	Cluster 1
0.9624	729454848.342619	939106.112342152	17	Cluster 2
0.9879	1252305295.06808	14361489.3554026	8	Cluster 3

Mean Sihlouette = 0.8758

Mean Cohesion = 7.5854e+006

Mean Separation = 9.4753e+008

SSE = 2.2756e+007

Παρατηρούμε ότι οι καλύτεροι αλγόριθμοι βάσει των μετρικών είναι ο Kmeans και ο Hierarchical για αριθμό clusters=2.

Από την ομαδοποίηση στον ιεραρχικό αλγόριθμο βλέπουμε ότι η χρήση της κουζίνας χωρίζεται σε 2 τύπους. Έτσι έχουμε αυξημένη χρήση (μπλε), και λίγη έως καθόλου (κόκκινο). Κάνοντας την αντιστοίχηση με τις ζώνες ώρας ανά εποχή και έτος συμπεραίνουμε τα εξής:

Το απόγευμα (15:00-23:00) του φθινοπώρου και της άνοιξης για κάθε έτος έχουμε την μεγαλύτερη κατανάλωση ρεύματος από τις συσκευές στην κουζίνα. Όλες οι εγγραφές ομαδοποιούνται στο μπλε cluster.

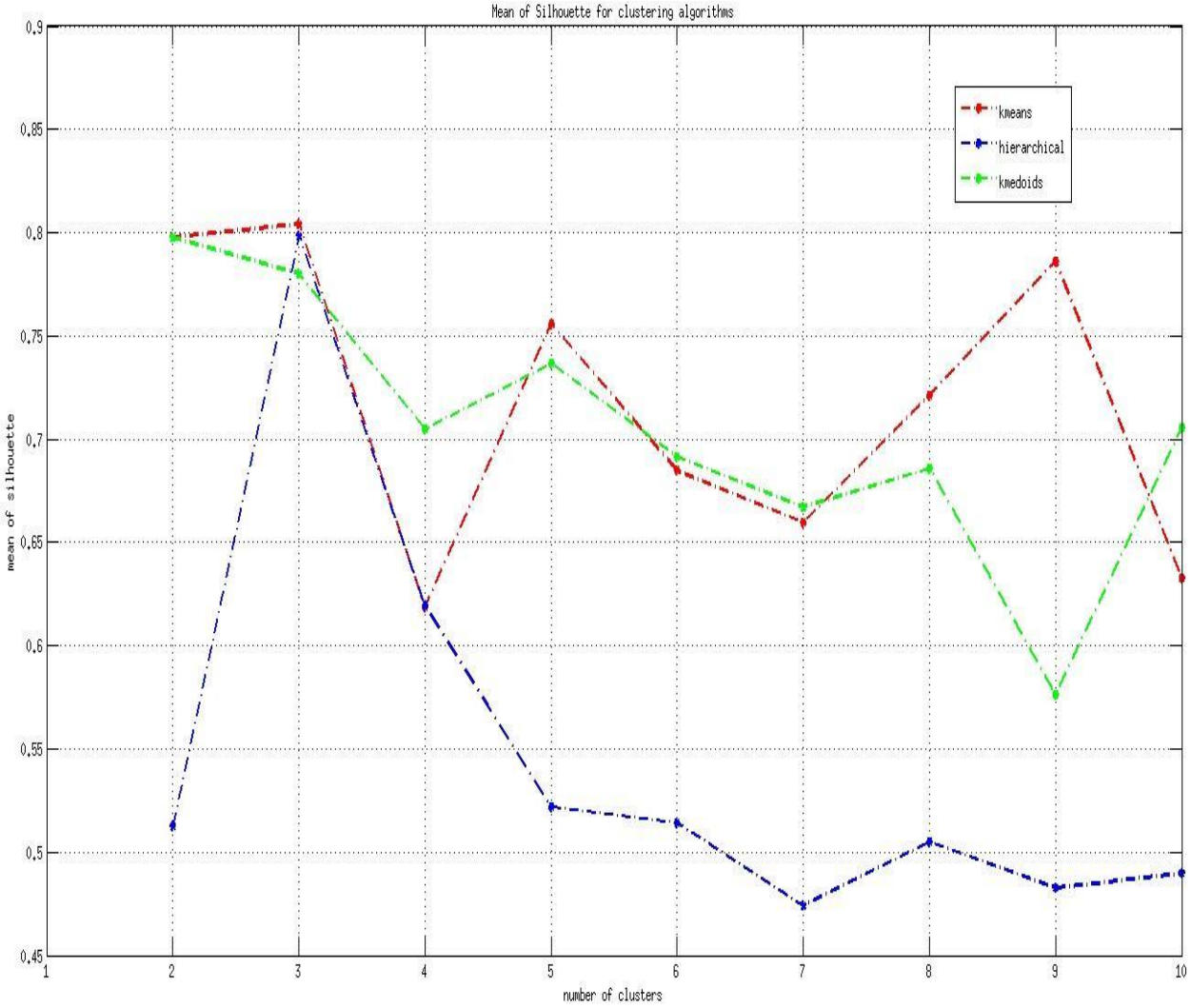
Οι υπόλοιπες χρονικές ζώνες για όλες τις εποχές έχουν γενικά χαμηλή κατανάλωση και ομαδοποιούνται στο κόκκινο cluster.

Δεν παρατηρείται διαφορά για διαφορετικά έτη. Δηλαδή, οι εγγραφές που αναφέρονται σε συγκεκριμένη εποχή και ζώνη ώρας ομαδοποιούνται στο ίδιο cluster για κάθε χρονιά.

ΤΟ ΠΛΥΣΤΑΡΙΟ

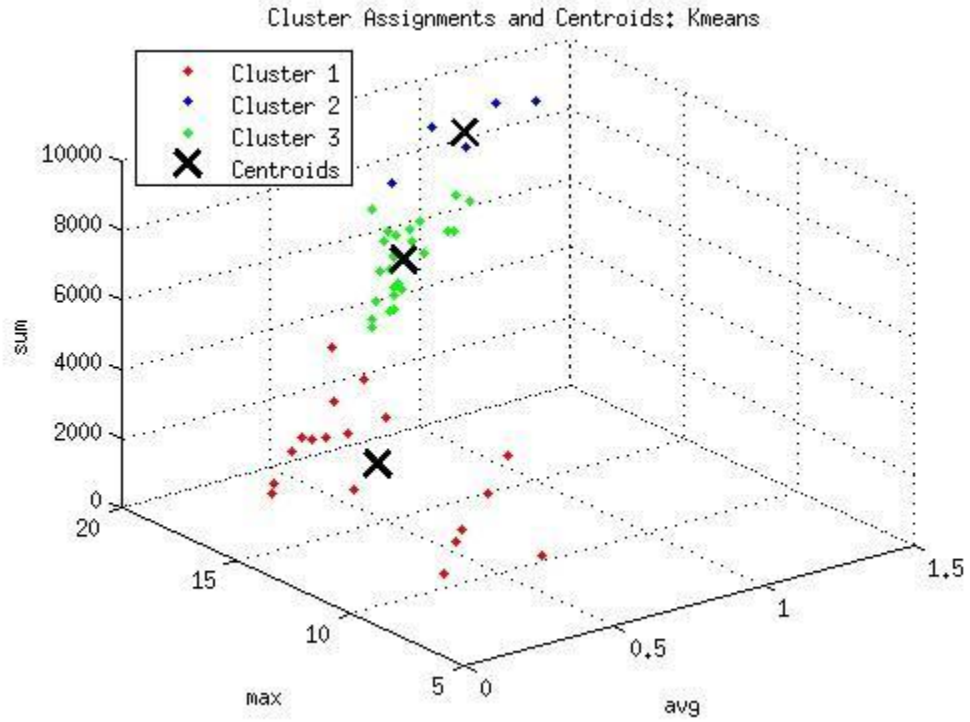
Παρακάτω φαίνεται η ομαδοποίηση για την κατανάλωση ηλεκτρικού ρεύματος από συσκευές που βρίσκονται στο **ΠΛΥΣΤΑΡΙΟ.**

Χρησιμοποιήθηκαν οι αλγόριθμοι KMeans, Hierarchical, KMedoids με διάφορες παραμέτρους. Τα καλύτερα αποτέλεσμα όπως φαίνεται και από το διάγραμμα μέσου όρου της μετρικής sihlouette είναι για **clusters=2 (αλγόριθμος Kmedoids)** ή για **clusters=3 (αλγόριθμοι KMeans και Hierarchical)**.



KMeans

- **Clusters=3**



Sihlouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.8476	1.8643e+08	1.2222e+07	18	Cluster 1
0.8409	4.6626e+07	1.9664e+06	5	Cluster 2
0.7653	6.7721e+05	1.4579e+07	25	Cluster 3

Mean Sihlouette = 0.804

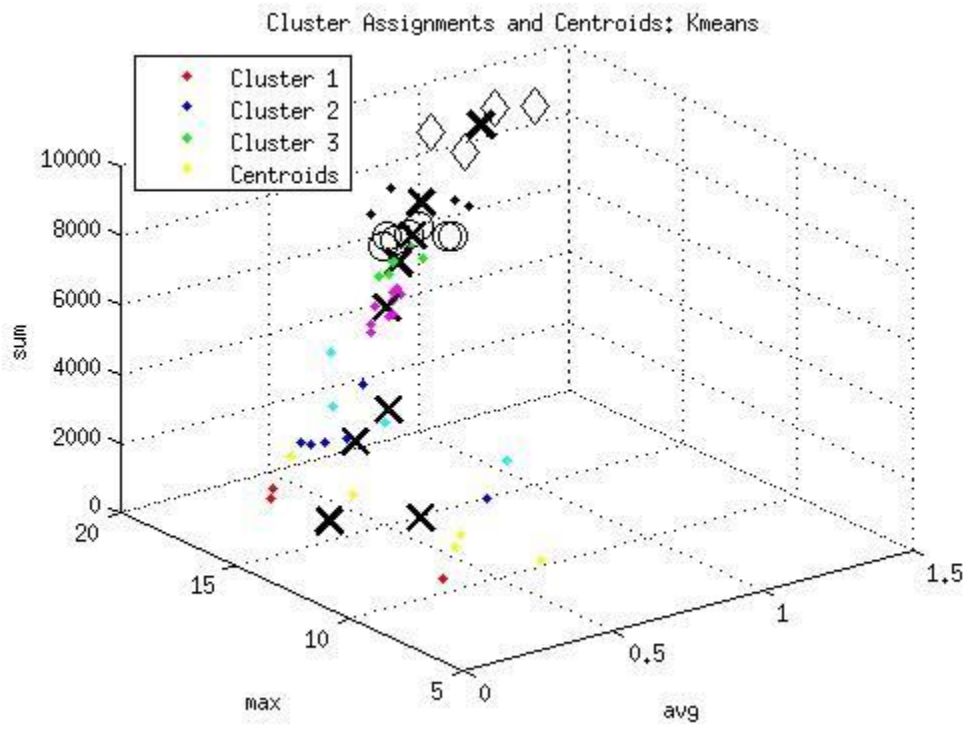
Mean Cohesion = 9.5892e+006

Mean Separation = 7.7912e+007

SSE = 2.8768e+007

Για τον συγκεκριμένο αλγόριθμο αξίζει να γίνει σύγκριση και με τα αποτελέσματα αν χρησιμοποιηθούν εννιά ομάδες:

- **Clusters=9**



Mean Sihlouette = 0.7819

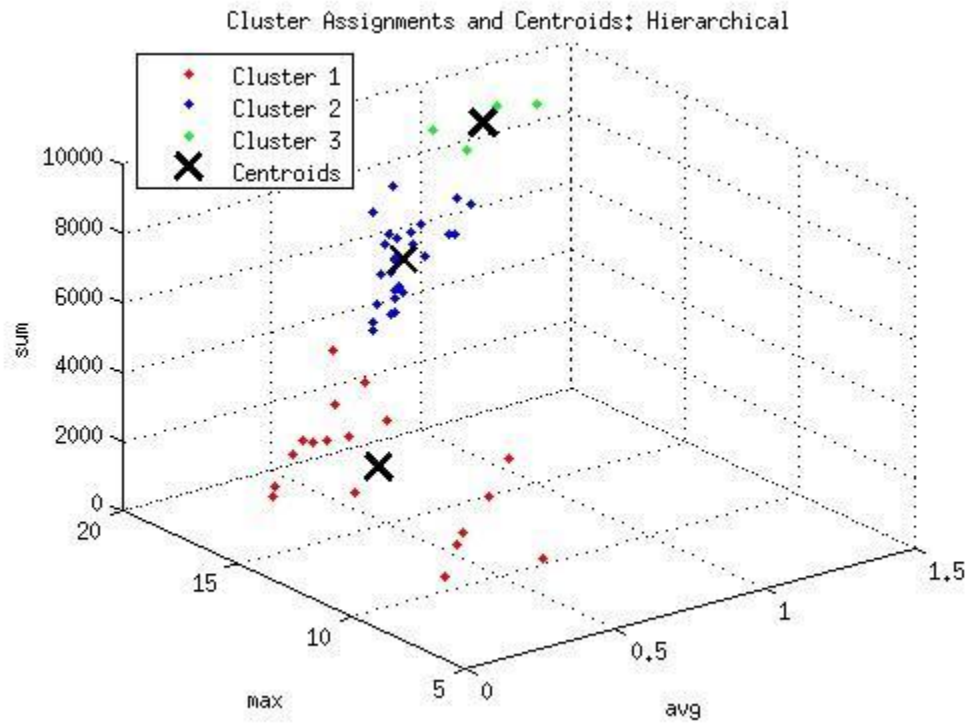
Mean Cohesion = 2.2711e+005

Mean Separation = 2.5559e+007

SSE = 2.044e+006

Hierarchical

- **Clusters=3**



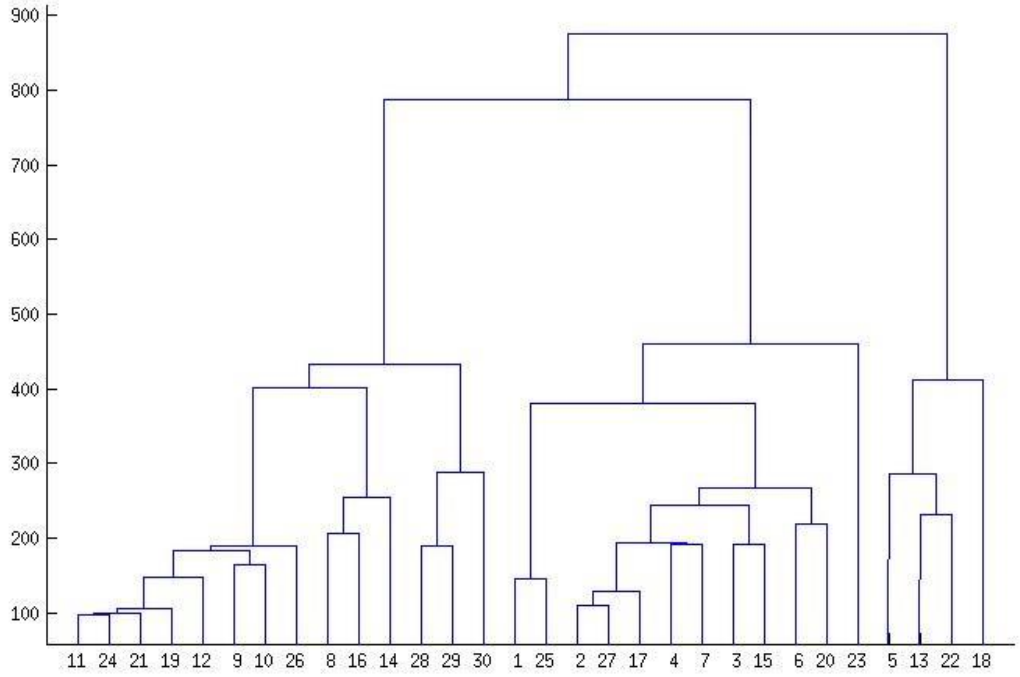
Sihlouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.8566	1.9992e+08	1.2222e+07	18	Cluster 1
0.7321	3.689e+05	1.7667e+07	26	Cluster 2
0.9678	4.1308e+07	4.6193e+05	4	Cluster 3

Mean Sihlouette = 0.7984

Mean Cohesion = 1.0117e+007

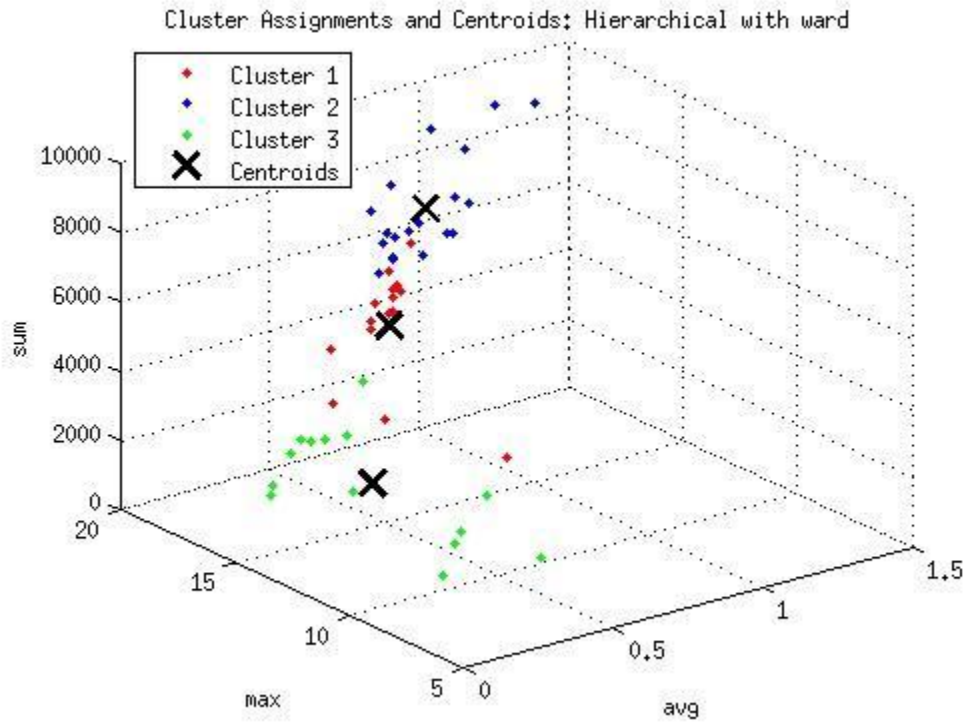
Mean Separation = 8.0533e+007

SSE = 3.0351e+007



Δενδρόγραμμα Ιεραρχικού Αλγόριθμου

Μέθοδος Ward:



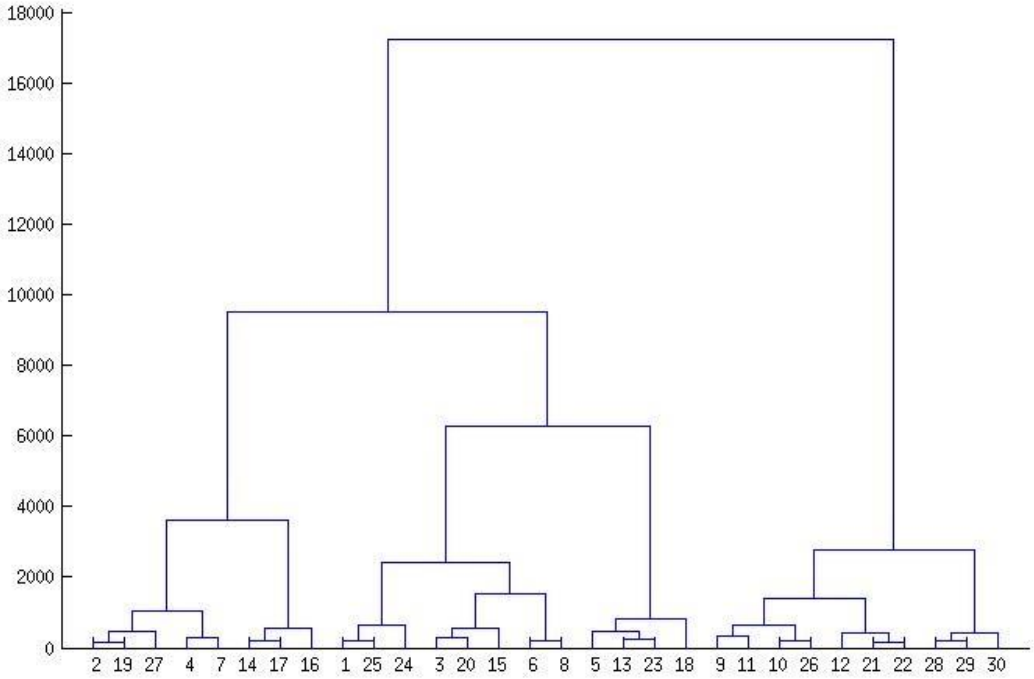
Sihlouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.6675	1.0109e+05	7.5113e+06	15	Cluster 1
0.4216	1.1015e+08	2.475e+07	19	Cluster 2
0.8789	8.6795e+07	5.2985e+06	14	Cluster 3

Mean Sihlouette = 0.6318

Mean Cohesion = 1.252e+007

Mean Separation = 6.5684e+007

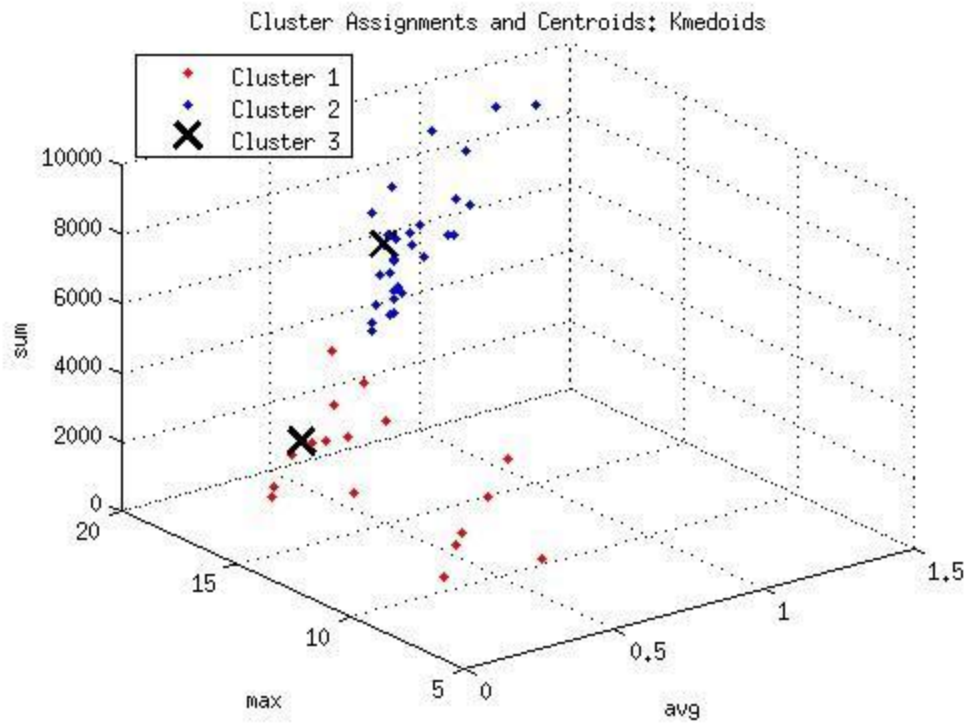
SSE = 3.756e+007



Δενδόγραμμα Ιεραρχικού Αλγόριθμου με μέθοδο Ward

KMedoids

- **Clusters=2**



Sihlouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.8994	6.5189e+07	1.2240e+07	18	Cluster 1
0.7367	1.0865e+08	5.1563e+07	30	Cluster 2

Mean Sihlouette = 0.7977

Mean Cohesion = 3.1901e+007

Mean Separation = 8.6918e+007

SSE = 6.3803e+007

Παρατηρούμε ότι οι καλύτεροι αλγόριθμοι βάσει των μετρικών είναι ο Hierarchical για αριθμό clusters=3 και ο KMedoid για αριθμό clusters=2.

Από παρατήρηση στα αποτελέσματα του καλύτερου αλγορίθμου δηλαδή του ιεραρχικού βλέπουμε πως:

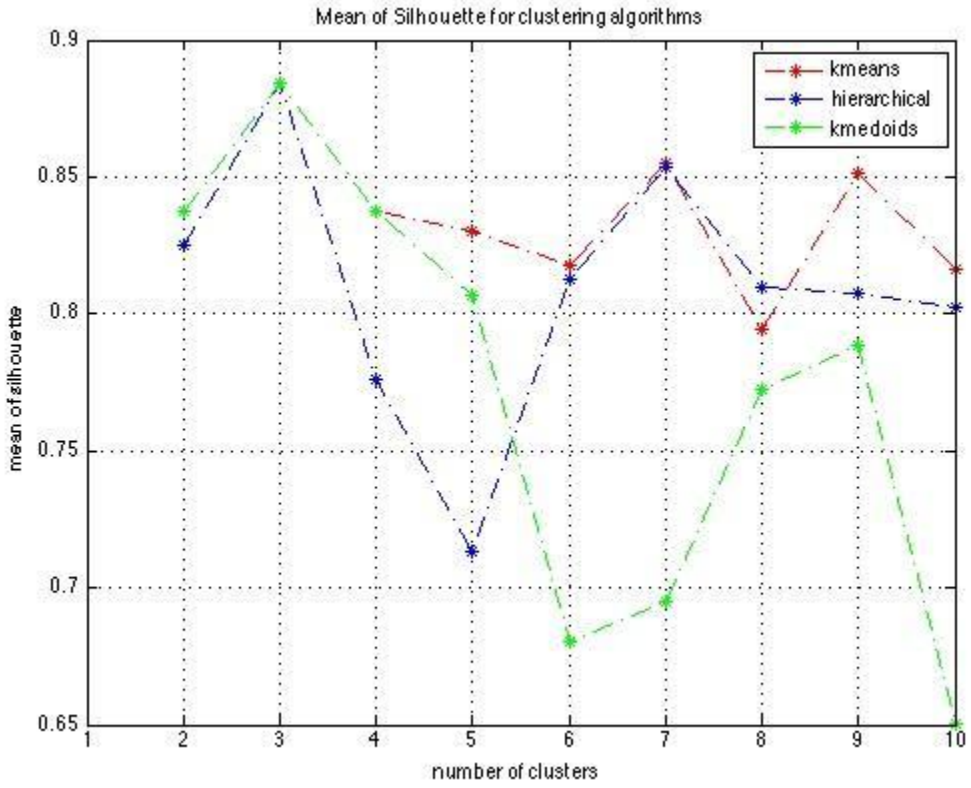
Τις βραδινές ώρες (23.00 - 8.00) η κατανάλωση σε αυτή την εγκατάσταση είναι πάντα η ελάχιστη και φαίνεται στην κόκκινη ομάδα.

Από εκεί και πέρα δεν μπορούν να εξαχθούν περαιτέρω ασφαλή συμπεράσματα αφού άλλα έτη και άλλες εποχές η κατανάλωση είναι η υψηλότερη είτε για πρωί είτε για απόγευμα. Αυτό είναι απόλυτα λογικό καθώς οι συσκευές που περιέχει το πλυσταριό (πλυντήριο, ψυγείο, στεγνωτήριο και λάμπα) χρησιμοποιούνται καθ' όλη την διάρκεια του χρόνου.

Θερμοσίφωνας και Air-condition

Εδώ κάνουμε ομαδοποίηση της κατανάλωσης που αναφέρεται στην θέρμανση/ψύξη του χώρου από την χρήση του καλοριφέρ και του air-condition.

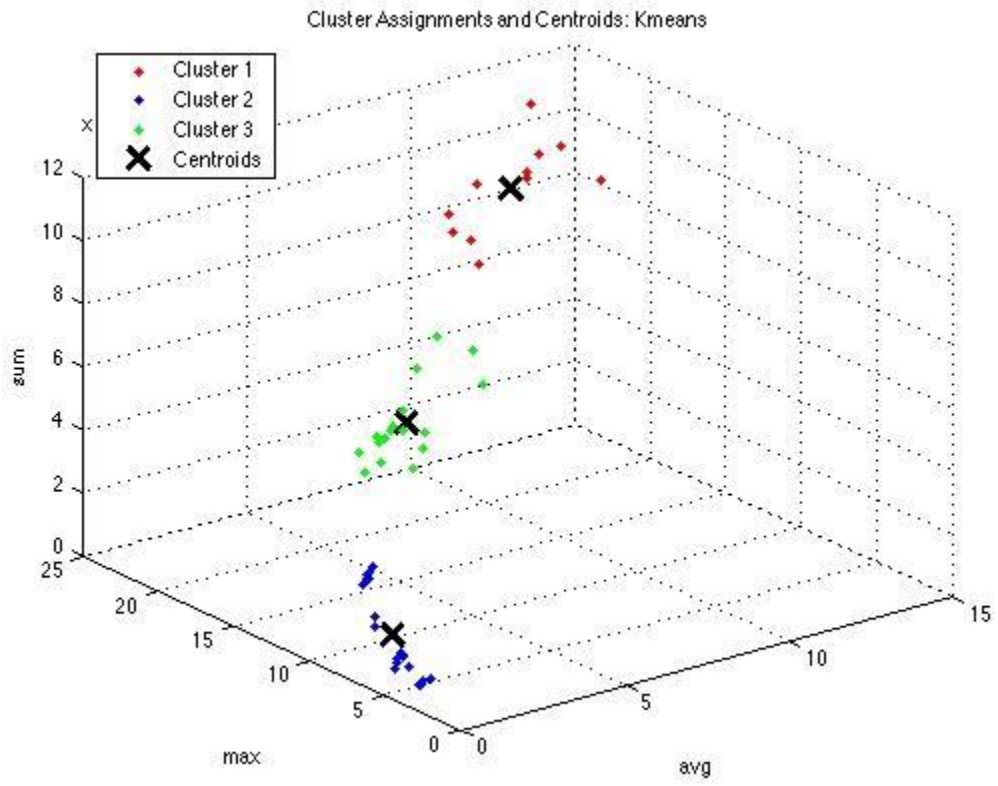
Αρχικά, τρέχουμε το script : silhouetteeval.m με σκοπό να δούμε τον αριθμό των clusters για τον οποίο εμφανίζεται ο υψηλότερος μέσος όρος silhouette. Τα αποτελέσματα φαίνονται στο παρακάτω σχήμα:



Όπως είναι προφανές το βέλτιστο clustering επιτυγχάνεται με χρήση 3 clusters. Ακόμη, αρκετά καλό αποτέλεσμα έχουμε και για 7 clusters για τους αλγόριθμους kmeans και hierarchical.

Kmeans

- **Clusters=3**



Παρατηρούμε ότι η χρήση θέρμανσης χωρίζεται σε 3 τύπους. Έτσι έχουμε αυξημένη χρήση (κόκκινο), μέτρια (πράσινο), λίγη έως καθόλου (μπλε). Κάνοντας την αντιστοίχηση με τις ζώνες ώρας ανά εποχή και έτος συμπεραίνουμε τα εξής:

Τα βράδια (23:00 - 7:00) έχουμε πάντοτε μικρή χρήση της θέρμανσης για όλες τις εποχές και για κάθε έτος. Όλες οι εγγραφές ομαδοποιούνται στο κόκκινο cluster.

Η μεγαλύτερη χρήση του θερμοσίφωνα και του air-condition εμφανίζεται στην πρωινή ζώνη (7:00-15:00) την άνοιξη, το φθινόπωρο και το χειμώνα. Οι εγγραφές με την μεγάλη κατανάλωση ομαδοποιούνται στο κόκκινο cluster.

Οι υπόλοιπες εγγραφές που αναφέρονται κυρίως στην απογευματινή ζώνη (15:00-23:00) εμφανίζουν μέτρια κατανάλωση και ομαδοποιούνται με πράσινο χρώμα.

Η εποχή κατά την οποία εμφανίζεται η ελάχιστη κατανάλωση είναι το καλοκαίρι κατά την οποία οι εγγραφές τις ανήκουν σχέδόν αποκλειστικά στο μπλε cluster.

Δεν παρατηρείται διαφορά για διαφορετικά έτη. Δηλαδή, οι εγγραφές που αναφέρονται σε συγκεκριμένη εποχή και ζώνη ώρας ομαδοποιούνται στο ίδιο cluster για κάθε χρονιά.

Silhouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.6906	2.234272816277065e+10	1.618905801175412e+09	11	Cluster 1
0.7754	3.108072406409027e+10	7.411611644374399e+08	20	Cluster 2
0.7083	5.421246116762352e+08	1.103964868824251e+09	17	Cluster 3

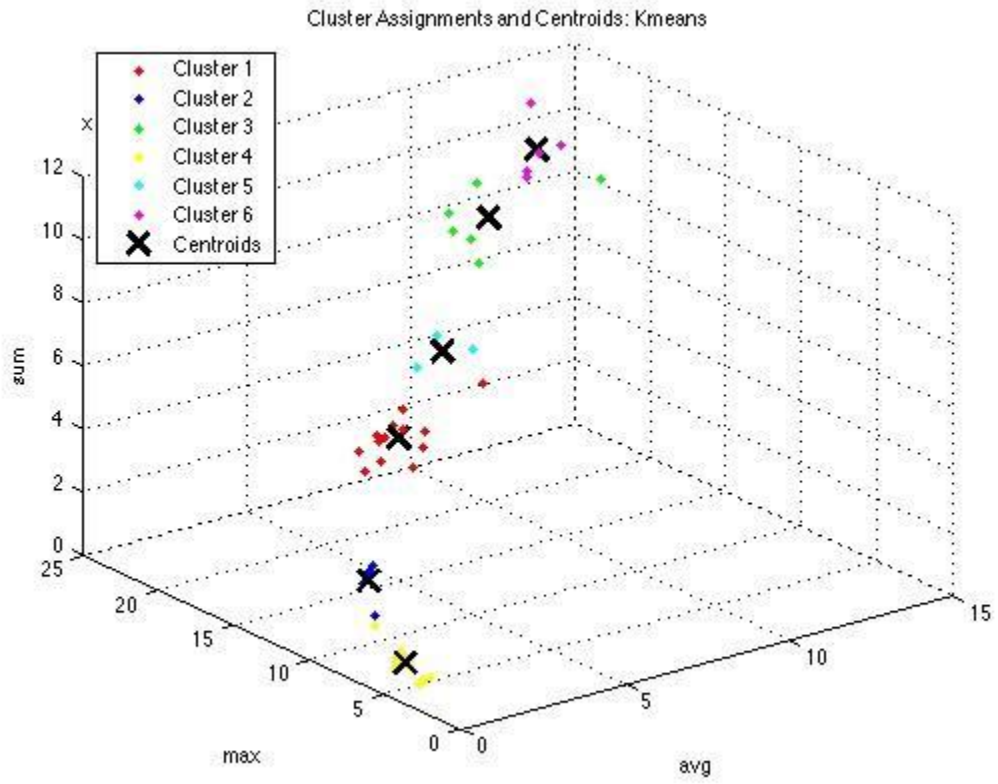
Mean Cohesion: 1.125348435595495e+08

Mean Separation: 9.901741105212824e+09

Mean Silhouette: 0.825

SSE: 6.752090613572972e+08

Για να πάρουμε καλύτερη ιδέα των τιμών που ξεχωρίζουν εφαρμόζουμε ***kmeans clustering με 6 clusters.***



Εδώ έχουμε 6 επίπεδα χρήσης του θερμοσίφωνα και air-condition. Αυτά που μας ενδιαφέρουν είναι τα πολύ υψηλής κατανάλωσης (μωβ) και αυτά πολύ χαμηλής κατανάλωσης (κίτρινο).

Βλέπουμε ότι την μεγαλύτερη κατανάλωση την έχουμε για τα πρωινά της άνοιξης για τα έτη 2008 και 2009, και για τα πρωινά του χειμώνα για τα έτη 2007, 2008 και 2009.

Από την άλλη μεριά η μικρότερη κατανάλωση εμφανίζεται κατά κύριο λόγο το καλοκαίρι όπως επίσης και στην νυχτερινή ζώνη για την άνοιξη και το χειμώνα. Το αποτέλεσμα αυτό ήταν αναμενόμενο αφού τα βράδια ο κόσμος κοιμάται και δεν αφήνει ανοικτές τις παραπάνω συσκευές γι' αυτό και τα βράδια έχουμε μικρή κατανάλωση.

Silhouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.6029	1.937403149683540e	2.262153199630214e+08	14	Cluster 1
0.8063	8.349823369999915e+09	4.547102706627893e+07	7	Cluster 2
0.6908	6.217078337232144e	1.263990256659541e+08	6	Cluster 3
0.5703	2.812103765547372e	4.350598185270663e+07	12	Cluster 4
0.6714	1.257540335929891e	5.595598072200325e+07	3	Cluster 5
0.6962	1.465935008529464e+10	1.776617260873328e	6	Cluster 6

Mean Cohesion: 1.125348435595495e+08

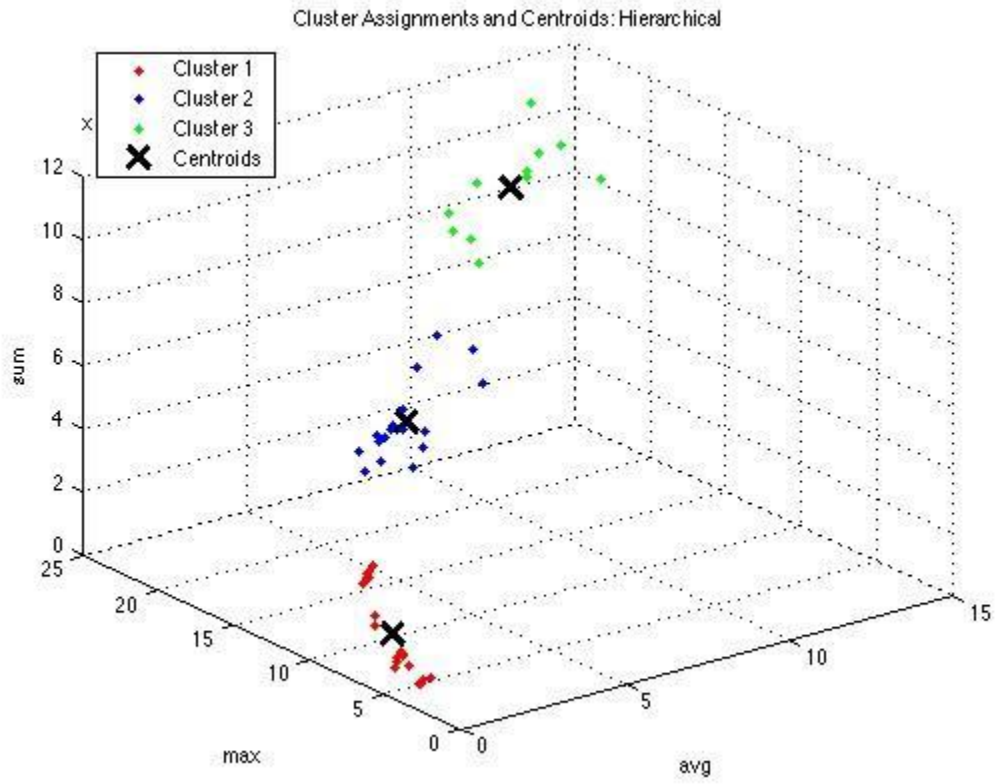
Mean Separation: 9.901741105212824e+09

Mean Silhouette: 0.7038

SSE: 6.752090613572972e+08

Hierarchical

- **Clusters=3**



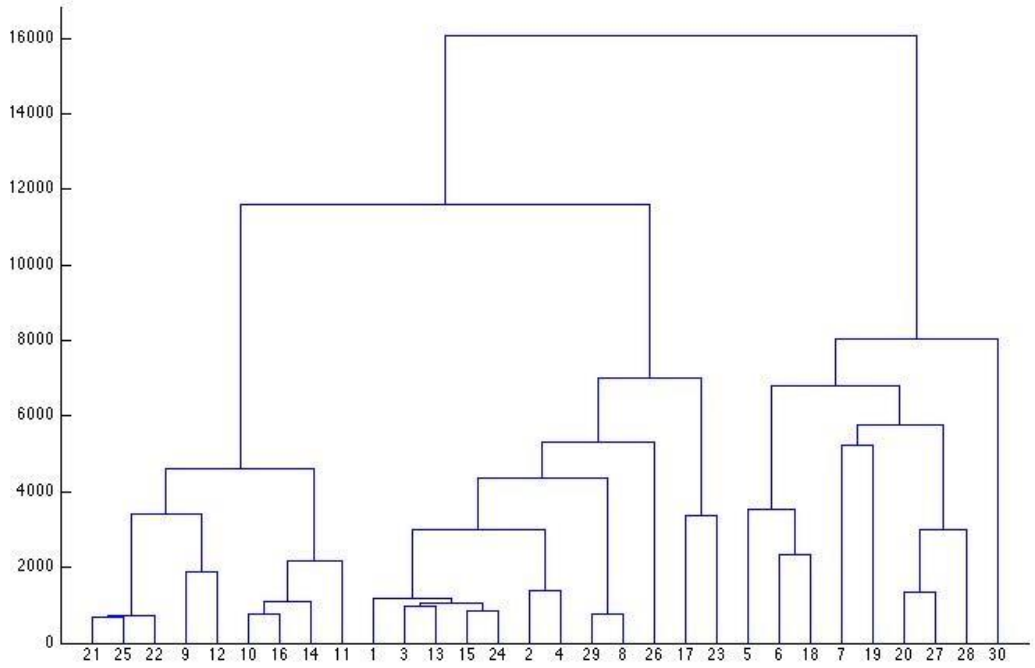
Silhouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.9184	3.108072406409027e+10	7.411611644374399e	20	Cluster 1
0.8593	5.421246116762352e+08	1.103964868824251e+09	17	Cluster 2
0.8595	2.234272816277065e+10	1.618905801175412e+09	11	Cluster 3

Mean Cohesion: 1.154677278145701e+09

Mean Separation: 1.798852561284572e+10

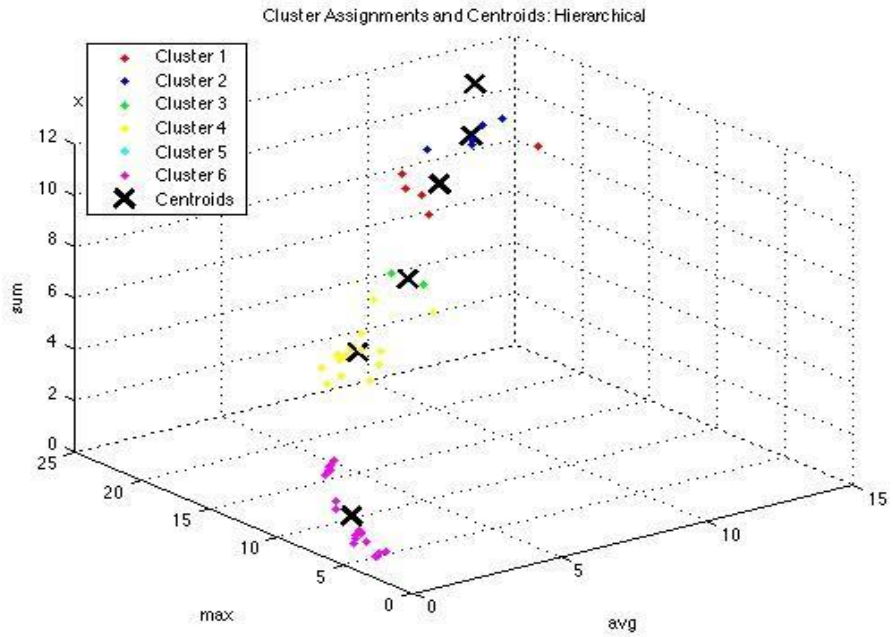
Mean Silhouette: 0.8440

SSE: 3.464031834437104e+09



Δενδρόγραμμα Ιεραρχικού Αλγόριθμου

- **Clusters=6**



Silhouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.9347	9.523251507587576e+08	4.014724353564537e+07	5	Cluster 1
0.3189	5.369333822860399e+09	1.508809038002176e+08	5	Cluster 2
0.9701	1.072671256644787e+08	5.658248898883620e+06	2	Cluster 3
0.7890	1.144890011025037e+10	3.708262332200335e+08	20	Cluster 4
1.0000	2.246049670472953e+09	0	1	Cluster 5
0.8978	6.964989297163420e+10	7.411611644374399e+08	15	Cluster 6

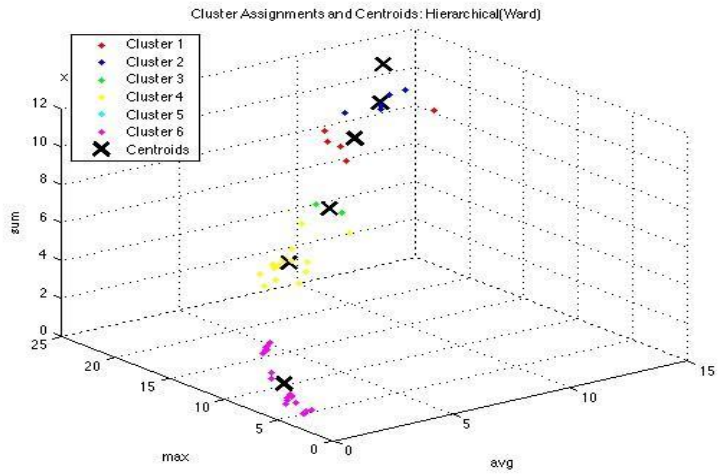
Mean Cohesion: 2.181122989820367e+08

Mean Separation: 1.496229480860686e+10

Mean Silhouette: 0.812508662977534

SSE: 1.308673793892220e+09

Μέθοδος Ward



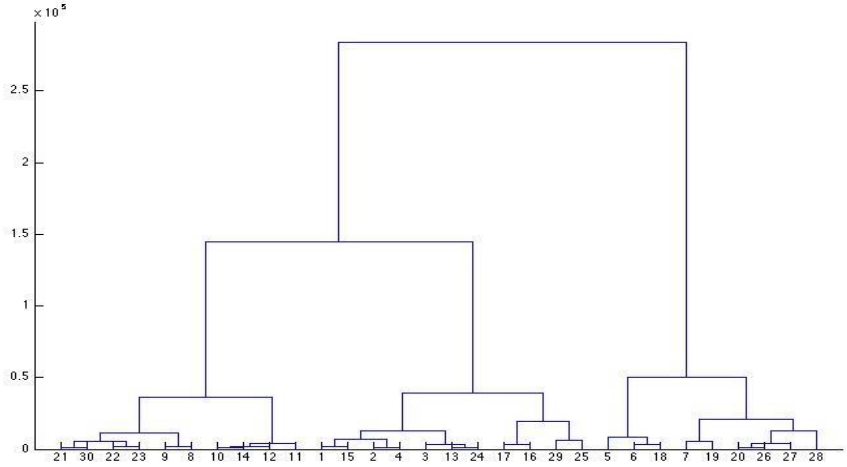
Silhouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.9742	6.238114395513231e+09	1.053959913713077e+07	6	Cluster 1
0.8560	2.792842517228238e+10	8.274407761778615e+07	14	Cluster 2
0.8980	1.588768995506694e+09	1.198216988974223e+08	12	Cluster 3
0.2725	5.306661653650491e+07	2.147945969197967e+08	5	Cluster 4
0.9595	5.080208807311661e+09	4.014724353564537e+07	5	Cluster 5
0.5749	6.964989297163420e+10	7.411611644374399e+08	6	Cluster 6

Mean Cohesion: 1.328519691181717e+08

Mean Separation: 9.653683266974825e+09

Mean Silhouette: 0.7961

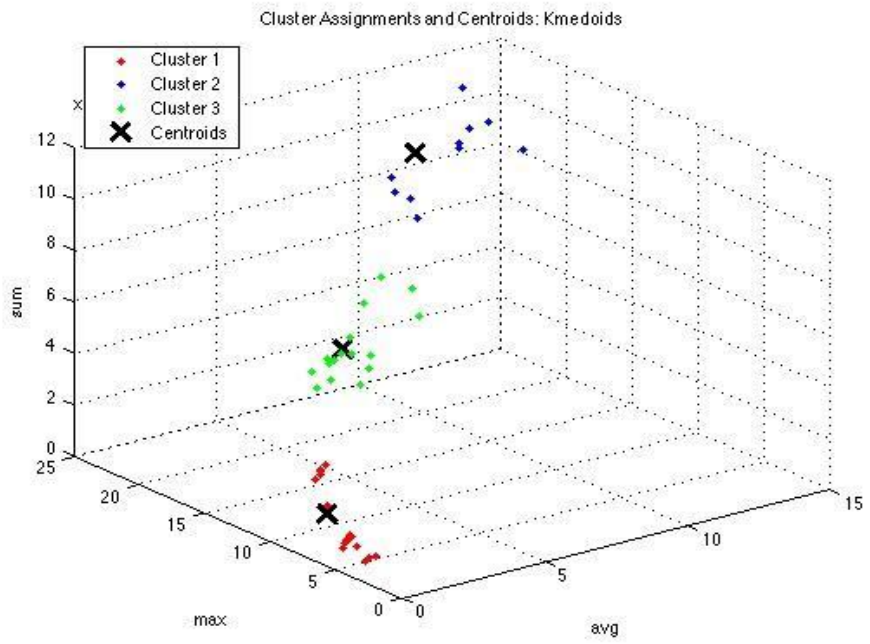
SSE: 7.971118147090302e+08



Δενδρόγραμμα Ιεραρχικού Αλγόριθμου με μέθοδο Ward

KMedoids

- **Clusters=3**



Silhouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.9184	2.922556249097687e+10	7.450714714658632e+08	20	Cluster 1
0.8595	2.160526209437743e+10	1.643736897488327e+09	11	Cluster 2
0.8593	6.308428677685955e+08	1.128322625828041e+09	17	Cluster 3

Mean Cohesion: 1.328519691181717e+08

Mean Separation: 9.653683266974825e+09

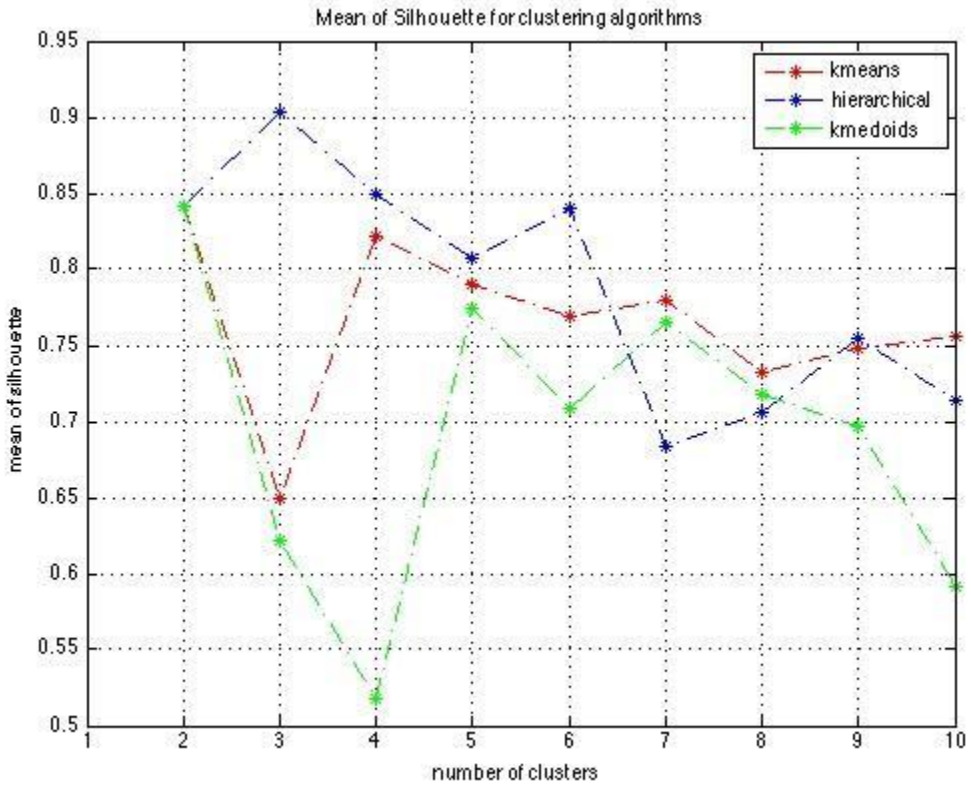
Mean Silhouette: 0.7961

SSE: 7.971118147090302e+08

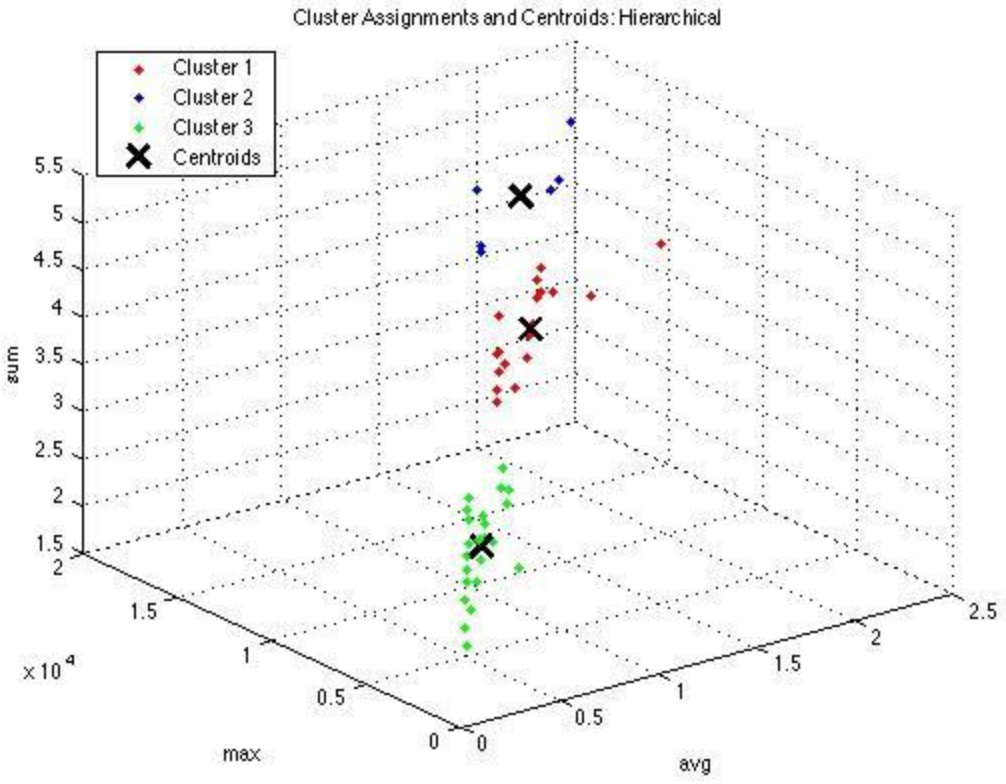
Συνολική Ισχύς

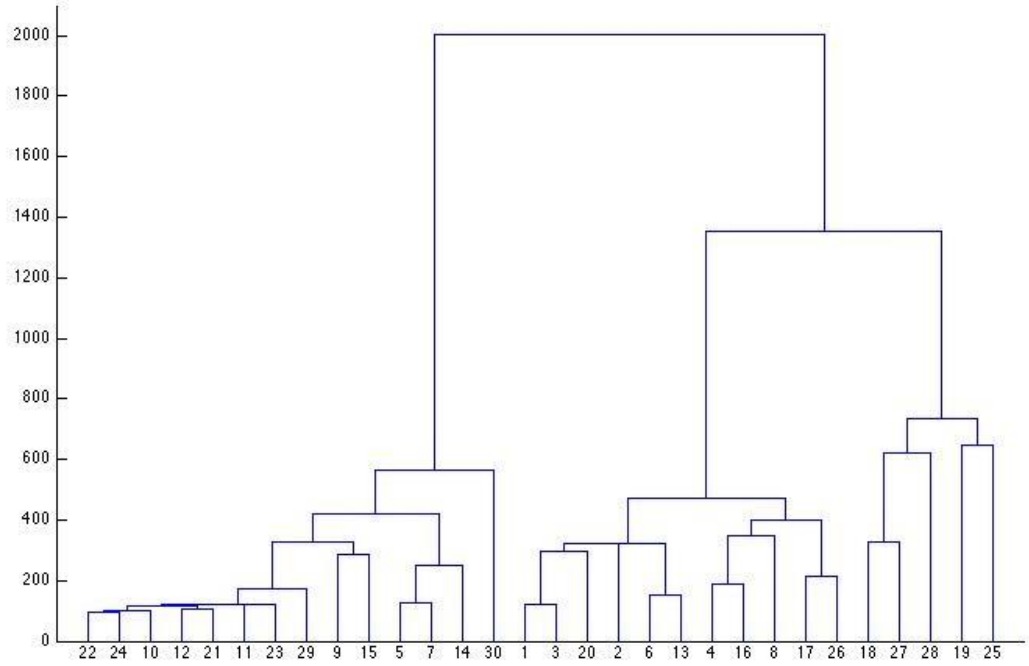
Εδώ ομαδοποιούμε τις εγγραφές που αναφέρονται στην συνολική ισχύ για κάθε ζώνη ώρας για κάθε εποχή και για κάθε έτος (σύνολο 48 εγγραφές). Η συνολική ισχύς αντιπροσωπεύει την κατανάλωση ηλεκτρικού ρεύματος της οικίας.

Χρησιμοποιούμε το script: silhouetteevaluation.m ώστε να αποφανθούμε για τον βέλτιστο αλγόριθμο ομαδοποίησης και αριθμό clusters. Το αποτέλεσμα φαίνεται παρακάτω:



Γίνεται φανερό ότι η καλύτερη ομαδοποίηση επιτυγχάνεται με τον ιεραρχικό αλγόριθμο για 3 clusters. Εφαρμόζοντάς τον λοιπόν έχουμε:





Δενδρόγραμμα ιεραρχικού αλγόριθμου

Silhouette	Separation	Cohesion	Στοιχεία	Cluster ID
0.8640	1.106230008955441e+06	1.349768523972303e+07	18	Cluster 1
0.8917	1.110641692811527e+08	3.720214416390215e+06	6	Cluster 2
0.9365	4.969279951300576e+08	1.766359023259908e+07	24	Cluster 3

Mean Cohesion: 1.162716329623745e+07

Mean Separation: 2.030327981400552e+08

Mean Silhouette: 0.9037

SSE: 3.488148988871233e+07

Προκύπτουν λοιπόν 3 ομάδες που αντιπροσωπεύουν την κατανάλωση ως εξής: Υψηλή κατανάλωση (μπλε), μέση κατανάλωση (κόκκινο), μικρή κατανάλωση (πράσινο). Οι εγγραφές που ομαδοποιούνται στην υψηλή κατανάλωση αναφέρονται αποκλειστικά στο χειμώνα και συγκεκριμένα στην πρωινή και απογευματινή ζώνη. Στη συνέχεια, μέση κατανάλωση έχουμε την άνοιξη και το φθινόπωρο πάλι στην πρωινή και απογευματινή ζώνη. Τέλος, οι υπόλοιπες εγγραφές αντιπροσωπεύουν μικρή κατανάλωση και αφορούν τις νυχτερινή ζώνη καθώς επίσης και όλη την διάρκεια του καλοκαιριού.

Σύνοψη και τελικά σχόλια

Όπως φάνηκε ο καλύτερος αλγόριθμος ομαδοποίησης ήταν ο ιεραρχικός για 2 ή 3 clusters.

Για όλες τις εγκαταστάσεις οι νυχτερινές ώρες καθ' όλη την διάρκεια του χρόνου ήταν οι λιγότερο ενεργειοβόρες. Οι συσκευές της κουζίνας είχαν την μεγαλύτερη κατανάλωση τα απογεύματα άνοιξης και φθινοπώρου. Οι συσκευές του πλυσταριού είχαν μια αρκετά ομοιόμορφη κατανάλωση, τα απογεύματα και τα πρωινά. Ο θερμοσίφωνας και το κλιματιστικό έιχαν την μεγαλύτερη κατανάλωση τα πρωινά της άνοιξης και του χειμώνα.

Με αυτά τα αποτελέσματα μπορούμε να εξάγουμε συμπεράσματα για τις συνήθειες του συγκεκριμένου καταναλωτή.