ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΕΚΤΙΜΗΣΗ ΤΙΜΩΝ ΑΚΙΝΗΤΩΝ ΜΕ ΜΕΘΟΔΟΥΣ ΧΩΡΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Αθηνά Γ. Παπαγεωργίου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών στο πλαίσιο του

Προπτυχιακού Προγράμματος Σπουδών

Αθήνα

Ιούλιος 2021

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω τις θερμές ευχαριστίες μου στον επιβλέποντα καθηγητή μου κύριο Καρλή Δημήτριο, Καθηγητή και Αναπληρωτή Πρόεδρο του τμήματος Στατιστικής του Οικονομικού Πανεπιστημίου Αθηνών, για την εμπιστοσύνη που μου έδειξε. Η καθοδήγηση και η βοήθειά του κατά τη διάρκεια εκπόνησης της διπλωματικής εργασίας ήταν συνεχής και πολύτιμη.

Παράλληλα, θα ήθελα να ευχαριστήσω τον κύριο Δεμίρη Νικόλαο, Επίκουρο Καθηγητή του τμήματος Στατιστικής του Οικονομικού Πανεπιστημίου Αθηνών και την κυρία Πεντελή Ξανθή, Επίκουρη Καθηγήτρια του ίδιου τμήματος, για τη συμμετοχή τους στη τριμελή επιτροπή.

Στη συνέχεια, θα επιθυμούσα να ευχαριστήσω από καρδιάς την οικογένειά μου για την άρρηκτη στήριξή τους όλα αυτά τα χρόνια. Ακόμη, θα ήθελα να εκφράσω την ευγνωμοσύνη μου προς τους φίλους μου για την αμέριστη συμπαράσταση και ενθάρρυνσή τους σε κάθε μου εγχείρημα. Τέλος, θα ήθελα να ευχαριστήσω όλους όσους πιστεύουν σ ’εμένα και με βοηθούν συνεχώς να εξελίσσομαι.

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

Γεννήθηκα στην Αθήνα και είμαι τελειόφοιτη στο τμήμα Στατιστικής του Οικονομικού Πανεπιστημίου Αθηνών. Μου αρέσουν πολύ τα μαθηματικά, να επεξεργάζομαι πληροφορίες και να αποκτώ νέες εμπειρίες. Με το πέρας των σπουδών μου, επιθυμώ να ασχοληθώ με την επιστήμη της Στατιστικής και να εφαρμόσω τις γνώσεις μου στην πραγματική αγορά.

ΠΕΡΙΛΗΨΗ

Αθηνά Παπαγεωργίου

ΕΚΤΙΜΗΣΗ ΤΙΜΩΝ ΑΚΙΝΗΤΩΝ ΜΕ ΜΕΘΟΔΟΥΣ ΧΩΡΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Ιούλιος 2021

Η πρόβλεψη της αξίας των ακινήτων αποτελεί σημαντικό ζήτημα, τόσο για τα χρηματοπιστωτικά συστήματα και τους διαχειριστές κινδύνου όσο και για τους κυβερνητικούς φορείς. Στην παρούσα εργασία, παρουσιάζουμε δύο μεθοδολογίες με σκοπό την ακριβή εκτίμηση της τιμής των ιδιοκτησιών. Πιο συγκεκριμένα, εφαρμόζουμε ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης και στη συνέχεια, προκειμένου να ληφθούν υπόψη τα χωρικά χαρακτηριστικά, ένα βαθμονομημένο μοντέλο γεωγραφικής παλινδρόμησης. Η υπόθεση είναι πως ακίνητα τα οποία βρίσκονται σε κοντινά σημεία τείνουν, λογικά, να έχουν σχετικές τιμές όταν διορθώσουμε για τα γνωρίσματά τους. Τα αποτελέσματα προκύπτουν από την ανάλυση ενός διαθέσιμου δείγματος 3978 ακινήτων. Από τη σύγκριση των δύο διαφορετικών προσεγγίσεων βάσει ενός μέτρου προβλεπτικής ικανότητας, καταλήγουμε πως η γεωγραφική παλινδρόμηση με βάρη, συμπεριφέρεται καλύτερα συγκριτικά με την πρώτη, η οποία αγνοεί τις χωρικές σχέσεις. Η ένταξη της χωρικής πληροφορίας στην παλινδρόμηση, μέσω ενός σχεδίου βαθμονόμησης ανάλογα με την τοποθεσία των ακινήτων, παρέχει βελτιωμένες προβλέψεις ως προς την ακρίβεια, όταν η απόσταση του καθορισμένου εύρους ζώνης μιας περιοχής οριστεί κατάλληλα.

ABSTRACT

ESTIMATION OF REAL ESTATE VALUES WITH SPATIAL REGRESSION METHODS

July 2021

The prediction of real estate values is an issue of great importance both for financial systems and risk managers as well as for government agencies. In this project, two methods are presented in order to accurately estimate the price of the properties. More specifically, we implement a multiple linear regression model and then, with the aim of taking into account the spatial characteristics, a geographically weighted regression model. The assumption is that properties that are located nearby tend, logically, to have relative prices when we correct for their features. The results come from the analysis of an available sample of 3978 properties. From the comparison of the two different approaches based on measures of predictive ability, we come to the conclusion that the geographically weighted regression behaves better compared to the first which ignores the spatial relations. The inclusion of spatial information in the regression through a calibration plan depending on the location of the properties provides improved predictions as to the accuracy, when the distance of the fixed bandwidth is set appropriately.

ΚΑΤΑΛΟΓΟΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

1. ΕΙΣΑΓΩΓΗ

Ο κλάδος του Real estate, αποτελεί έναν από τους σημαντικότερους οικονομικούς παράγοντες, τόσο σε πανελλήνιο όσο και σε παγκόσμιο επίπεδο. Μετά την οικονομική κρίση, η οποία εκδηλώθηκε στην Ελλάδα στα τέλη του 2008, ο συγκεκριμένος τομέας υπέστη σοβαρό πλήγμα καθώς η κατασκευαστική δραστηριότητα περιορίστηκε σε μεγάλο βαθμό και οι τιμές των ακινήτων παρουσίασαν μείωση της τάξης του 40%. Λόγω των προηγούμενων γεγονότων, τα χρηματοπιστωτικά ιδρύματα, τα οποία αλληλοεπιδρούν με τις αγοραπωλησίες ιδιοκτησιών, αναζητούν αξιόπιστες προβλέψεις των τιμών των ακινήτων. Η ανάγκη αυτή, θέτει μία ενδιαφέρουσα πρόκληση καθώς τα δεδομένα που είναι διαθέσιμα και συνυφασμένα με το ζητούμενο είναι περιορισμένα. Ακόμη, η χώρα μας διακρίνεται από ανομοιογένεια χαρακτηριστικών που έχουν άμεση επίδραση στην τιμή των ακινήτων και καθιστά πιο περίπλοκη τη διαδικασία εύρεσης του κατάλληλου εργαλείου που θα λαμβάνει υπόψη τα ιδιαίτερα χαρακτηριστικά.

Κρίνεται, λοιπόν, αναγκαίο οι προβλέψεις που παρέχουν οι επιστήμονες και οι διαχειριστές κινδύνου αναφορικά με την τιμή των ακινήτων, να έχουν προκύψει μέσω της βέλτιστης μεθόδου με στόχο τη μεγαλύτερη ακρίβεια. Οι αποτελεσματικές προβλέψεις αποτελούν χρήσιμο εργαλείο για τους εργαζόμενους στον τομέα της διαχείρισης κινδύνου των τραπεζών, οι οποίοι είναι υπεύθυνοι να αξιολογήσουν και να προβλέψουν αν, και κατά πόσο, ένας υποψήφιος δανειολήπτης θα έχει τη δυνατότητα να ανταπεξέλθει στην αποπληρωμή ενός στεγαστικού δανείου. Η έγκυρη και έγκαιρη πρόβλεψη των τιμών των ακινήτων είναι, επίσης, σημαντική για τους κυβερνητικούς φορείς που επιδίδονται στη θέσπιση των φόρων. Πιο συγκεκριμένα, η Ελλάδα κατατάσσεται στην τρίτη θέση μεταξύ των χωρών της ΕΕ με τον υψηλότερο φόρο κατοχής ακινήτων, έχοντας λάβει υπόψη όλους τους επαναλαμβανόμενους φόρους που σχετίζονται με την ιδιοκτησία. Ως αποτέλεσμα, είναι απαραίτητο να γίνουν διαθέσιμες σωστές αξιολογήσεις και προβλέψεις στον χώρο των ακινήτων, προκειμένου να αναδιαμορφωθεί ο συντελεστής φορολόγησής τους και να λειτουργήσουν ως ένα επενδυτικό και αναπτυξιακό εργαλείο (enikonomia, 2020).

Κύριος σκοπός της παρούσας εργασίας είναι η διερεύνηση ενός μοντέλου το οποίο θα λαμβάνει υπόψη τη χωρική πληροφορία με στόχο τη βέλτιστη πρόβλεψη των τιμών των ακινήτων. Για την επίλυση του συγκεκριμένου ζητήματος έχουν προταθεί κατά

καιρούς διάφορες τεχνικές οι οποίες θα μπορούσαν να διακριθούν σε παραδοσιακές και πιο εξελιγμένες. Οι παραδοσιακές μέθοδοι αναφέρονται κυρίως στη γραμμική παλινδρόμηση σε διάφορες μορφές όπως η πολλαπλή, stepwise, quantile, additive, τοπική κ.ά. Από την άλλη μεριά, πιο αναπτυγμένες, υπολογιστικά, τεχνικές που έχουν προταθεί και εφαρμοστεί στο συγκεκριμένο πρόβλημα είναι εκείνες της τεχνικής μάθησης, όπως τα neural networks, random forests, tree-based models, χωρική ανάλυση, fuzzy logics κ.ά.

Στόχος είναι να παρουσιάσουμε ένα μοντέλο με τη χρήση του οποίου θα καταφέρουμε να αποκτήσουμε προβλέψεις των τιμών και όχι αξιολόγηση των χαρακτηριστικών των ακινήτων και σε τι βαθμό εκείνα τις επηρεάζουν. Το πρόβλημα αυτό, παρουσιάζει ενδιαφέρον καθώς η Ελλάδα, είναι χώρα με ανομοιογενή χαρακτηριστικά, όπως μεγάλες αστικές περιοχές, δημοφιλείς τουριστικούς προορισμούς, αγροτικές περιοχές, τα οποία ενδέχεται να δρουν με διαφορετικό τρόπο στη διαμόρφωση των αντίστοιχων τιμών. Για το λόγο αυτό, επιθυμούμε να εξετάσουμε αν η συμπερίληψη της χωρικής πληροφορίας ενός ακινήτου στο μοντέλο πρόβλεψης, μπορεί να βελτιώσει και να προσδώσει μεγαλύτερη ακρίβεια συγκριτικά με εκείνο που δεν τη λαμβάνει υπόψη. Για παράδειγμα, η αξία ενός ακινήτου το οποίο βρίσκεται στην Αθήνα, είναι πιθανό να επηρεάζεται σε μικρότερο ή μεγαλύτερο βαθμό από ορισμένα χαρακτηριστικά σε σχέση με κάποιο ακίνητο το οποίο εντοπίζεται σε αγροτική περιοχή. Προκειμένου να εξεταστεί αν ισχύει η παραπάνω υπόθεση, γίνεται χρήση ενός γεωγραφικού μοντέλου παλινδρόμησης με βαθμονόμηση (Brunsdon, Fotheringham and Charlton, 1998). Το συγκεκριμένο μοντέλο, λαμβάνει υπόψη τη χωρική ετερογένεια μεταξύ των παρατηρήσεων, και η επιρροή του κάθε χαρακτηριστικού σε συγκεκριμένο ακίνητο (παράμετροι β) μεταβάλλεται ανάλογα με την τοποθεσία του. Ειδικότερα, οι παρατηρήσεις που εμπίπτουν σε μία περιοχή, η οποία μπορεί να οριστεί τόσο από τον ερευνητή όσο και αυτόματα, λαμβάνουν διαφορετική βαθμονόμηση, ανάλογα με τη θέση τους, για κάθε τους μεταβλητή. Είναι λογικό να αναμένουμε πως ακίνητα τα οποία βρίσκονται σε κοντινά σημεία θα παίρνουν και παρεμφερείς τιμές όταν η συγκεκριμένη ιδιότητα διορθώσει τα χαρακτηριστικά τους. Ωστόσο, όταν τα βάρη του γεωγραφικού μοντέλου προσεγγίζουν τη μονάδα, τότε κι εκείνο συγκλίνει στο απλούστερο μοντέλο παλινδρόμησης. Συνεπώς, ενδιαφερόμαστε να διερευνήσουμε αν η προσθήκη της χωρικής πληροφορίας στο μοντέλο και η χρήση μίας συνάρτησης βαθμονόμησης,

συνεισφέρουν σημαντικά και παρέχουν βελτιωμένα αποτελέσματα ως προς την πρόβλεψη της αξίας των ιδιοκτησιών.

Πιο συγκεκριμένα, στην παρούσα εργασία, θα παρουσιαστούν, αρχικά, ορισμένες διαφορετικές προσεγγίσεις που υπάρχουν και έχουν προταθεί στη βιβλιογραφία σχετικά με το παρόν ζήτημα και θα συζητηθούν τα πλεονεκτήματα και τα μειονεκτήματα των εκάστοτε μεθόδων. Γίνεται νύξη της πολλαπλής γραμμικής παλινδρόμησης, της stepwise, η μέθοδος χωρικής επέκτασης καθώς και η γεωγραφική βαθμονομημένη παλινδρόμηση, την οποία προτείνουμε. Ακόμη, αναφέρονται εξελιγμένες τεχνικές όπως τα neural networks, η χωρική ανάλυση, random forests και άλλες μέθοδοι machine learning.

Δεύτερον, θα αναπτύξουμε αναλυτικά τα μοντέλα πολλαπλής και γεωγραφικής παλινδρόμησης με βαθμονόμηση. Θα γίνει παρουσίαση του μαθηματικού υποβάθρου και σχολιασμός για το καθένα, με ιδιαίτερη έμφαση στο δεύτερο εξ αυτών που λαμβάνει υπόψη τη χωρική συσχέτιση μεταξύ των ακινήτων.

Στη συνέχεια, παρουσιάζονται τα δεδομένα στα οποία θα εφαρμοστούν τα προαναφερθέντα μοντέλα και θα αξιολογηθούν με βάση την προβλεπτική τους ικανότητα. Πιο συγκεκριμένα, έχουμε διαθέσιμο ένα σετ παρατηρήσεων από τον τομέα του real estate που αφορά 3978 ακίνητα και η καταγραφή της αξίας τους έχει γίνει κατά τη διάρκεια των ετών 2010-2019. Αναφέρονται, δηλαδή σε χρονική περίοδο μετά την οικονομική κρίση, και για το λόγο αυτό, δεν θα μας απασχολήσει οποιαδήποτε σύγκριση και συμπερασματολογία στις τιμές των ακινήτων αναφορικά με το συγκεκριμένο γεγονός. Για την εκτίμηση της αξίας τους, γίνεται χρήση τόσο ποσοτικών όσο και ποιοτικών χαρακτηριστικών όπως είναι η έκταση τους σε τετραγωνικά μέτρα, ο τύπος του ακινήτου, ο αριθμός των δωματίων κ.ά.

Στο πέμπτο κεφάλαιο της εργασίας, θα εφαρμόσουμε, αρχικά, ένα μοντέλο πολλαπλής παλινδρόμησης με μεταβλητή απόκρισης την τιμή σε λογαριθμική κλίμακα και εξαρτημένες μεταβλητές, χαρακτηριστικά των ακινήτων. Όπως έχει γίνει κατανοητό, κύριο μέλημα είναι η πρόβλεψη της αξίας των ακινήτων, και για το λόγο αυτό, η αξιολόγηση των υποψήφιων μοντέλων θα γίνει με την εισαγωγή μέτρων καλής πρόβλεψης και όχι προσαρμοστικότητας όπως συνηθίζεται. Ακόμη, δεν θα μας απασχολήσει η ικανοποίηση των υποθέσεων του μοντέλου της πολλαπλής γραμμικής παλινδρόμησης. Στη συνέχεια, έχοντας και πάλι ως μεταβλητή απόκρισης τον λογάριθμο της τιμής των ακινήτων και χρησιμοποιώντας τα ίδια χαρακτηριστικά που

συμπεριλήφθηκαν στο μοντέλο γραμμικής παλινδρόμησης, θα εφαρμόσουμε στα δεδομένα ένα γεωγραφικό μοντέλο που θα δίνει διαφορετική βαθμονόμηση σε κάθε παρατήρηση ανάλογα με την τοποθεσία όπου συναντάται. Θα εξεταστεί, ακόμη, η συμπεριφορά των μοντέλων από τα οποία έχουμε αφαιρέσει διαδοχικά μεταβλητές. Θα παρουσιαστούν τα αποτελέσματα της κάθε ανάλυσης και θα γίνει αξιολόγηση και σύγκρισή τους με κύριο γνώμονα την προβλεπτική τους ικανότητα αναφορικά με την αξία.

Τα τελικά συμπεράσματα καθώς και περαιτέρω σχόλια και παρατηρήσεις που αφορούν το ζητούμενο της πρόβλεψης των ακινήτων με μεθόδους χωρικής παλινδρόμησης θα συζητηθούν στο τελευταίο κεφάλαιο.

Συνοπτικά, το υπόλοιπο της παρούσας εργασίας οργανώνεται ως εξής: Το Κεφάλαιο 2 αναφέρεται στη βιβλιογραφία και κοινοποιούνται ορισμένες μέθοδοι που έχουν προταθεί για την επίλυση του προβλήματος. Στο Κεφάλαιο 3 παρουσιάζονται εκτενώς, τα μοντέλα πολλαπλής και γεωγραφικής παλινδρόμησης με βάρη. Τα δεδομένα των ακινήτων που είναι διαθέσιμα παρουσιάζονται στο Κεφάλαιο 4, ενώ η ανάλυση και τα αποτελέσματα των μοντέλων βρίσκονται στο Κεφάλαιο 5. Τέλος, στο τελευταίο Κεφάλαιο 6 της εργασίας, γίνεται σχολιασμός των συμπερασμάτων της παρούσας έρευνας και προτροπή για περαιτέρω διερεύνηση.

2. ΒΙΒΛΙΟΓΡΑΦΙΑ

Η ανάγκη εύρεσης ενός κατάλληλου εργαλείου για έγκυρες προβλέψεις των τιμών των ακινήτων, έχει οδηγήσει πλήθος ερευνητών να προτείνει πολλές και διαφορετικές μεθόδους για την αντιμετώπισή της. Το σύνολο αυτών των προσεγγίσεων μπορεί να διακριθεί σε δύο μεγάλες κατηγορίες, τα παραδοσιακά εργαλεία και τα πιο εξελιγμένα (Pagourtzi, et al., 2003). Στην πρώτη κατηγορία, εντάσσονται μέθοδοι όπως η απλή ή πολλαπλή γραμμική παλινδρόμηση, η stepwise παλινδρόμηση, η μέθοδος χωρικής επέκτασης και η χωρική παλινδρόμηση. Στη δεύτερη κατηγορία συναντώνται πιο περίπλοκες, υπολογιστικά, τεχνικές όπως είναι τα neural networks, η χωρική ανάλυση, τεχνικές machine learning, random forests και decision trees.

Η απλή ή πολλαπλή γραμμική παλινδρόμηση, εξετάζει με ποιον τρόπο, και σε τι βαθμό, ορισμένα χαρακτηριστικά επηρεάζουν τη μεταβλητή απόκρισης για την οποία ενδιαφερόμαστε. Ακόμη, με τη χρήση ενός τέτοιου μοντέλου, δεδομένου ότι έχουν χρησιμοποιηθεί κατάλληλα κριτήρια αξιολόγησης, είμαστε σε θέση να προβούμε και σε προβλέψεις. Αποτελεί ένα απλό και κατανοητό εργαλείο, στο οποίο έχουμε τη δυνατότητα να επιλέξουμε τις μεταβλητές που θα συμμετέχουν, ενώ είναι αρκετά εύκολη η γενίκευση και η ερμηνεία του. Ωστόσο, η απλοϊκότητα της συγκεκριμένης μεθόδου και το γεγονός ότι η τιμή των ακινήτων δεσμεύεται να διαμορφώνεται από συγκεκριμένα χαρακτηριστικά τους (Alexandridis, et al., 2018), αποτελούν και από τα κύρια μειονεκτήματά της. Πιο συγκεκριμένα, η γραμμική παλινδρόμηση προσδίδει σε κάθε χαρακτηριστικό που έχουμε εντάξει, και θεωρείται ότι επηρεάζει στατιστικά σημαντικά την εξαρτημένη μεταβλητή, τον ίδιο συντελεστή. Το γεγονός αυτό, διευκολύνει σε μεγάλο βαθμό την κατανόηση και ερμηνεία του, ωστόσο δεν είναι ρεαλιστικό, καθώς δε λαμβάνει υπόψη τη χωρική πληροφορία που ενδέχεται να παίζει σημαντικό ρόλο στο αποτέλεσμα. Ειδικότερα, όταν αντιμετωπίζουμε το πρόβλημα πρόβλεψης τιμών των ακινήτων, είναι λογικό να αναμένουμε πως υπάρχει χωρική ετερογένεια, δηλαδή κάποια χαρακτηριστικά επηρεάζουν σε μεγαλύτερο βαθμό την αξία, ανάλογα με την τοποθεσία τους. Συνεπώς, η ιδιαίτερη αυτή συνθήκη, που ενδέχεται να μεταβάλλει σημαντικά τις προβλέψεις των ακινήτων, αγνοείται. Ένα ακόμη μειονέκτημα της παλινδρόμησης αποτελεί το πλήθος των προϋποθέσεων, οι οποίες πρέπει να ικανοποιούνται προκειμένου τα συμπεράσματα που θα προκύψουν από το μοντέλο να είναι αξιόπιστα. Επίσης, είναι πιθανόν να εμφανιστούν προβλήματα όπως η υπόθεση λανθασμένης μορφής του μοντέλου, μη γραμμικότητα,

η πολυσυγγραμικότητα, ετεροσκεδαστικότητα, ορισμένα από τα οποία μπορούν να επιλυθούν ή να παραλειφθούν εφόσον δεν επηρεάζουν την προβλεπτική ικανότητα του μοντέλου. Σύμφωνα με τους Nghiep Nguyen and Al Cripps (2001), οι οποίοι συγκρίνουν πέντε διαφορετικά μοντέλα παλινδρόμησης με τη μέθοδο neural networks, προκύπτει πως τα πρώτα είναι πιο αποδοτικά στις προβλέψεις όταν το μέγεθος δείγματος είναι μικρό.

Αναφορικά με τη stepwise παλινδρόμηση, αποτελεί μία μέθοδο σύμφωνα με την οποία, επιλέγονται ποιες από τις υποψήφιες ανεξάρτητες μεταβλητές είναι κατάλληλες να πλαισιώσουν το μοντέλο (Pagourtzi, et al., 2003). Εξετάζοντας αν η προσθήκη ενός χαρακτηριστικού βελτιώνει σημαντικά ή όχι την προσαρμοστικότητα του μοντέλου, αποφασίζει αν θα συμπεριληφθεί σε αυτό. Πρόκειται, ουσιαστικά, για μία κλιμακωτή διαδικασία προσθαφαίρεσης μεταβλητών όπου σε κάθε βήμα ελέγχεται αν μία μεταβλητή πρέπει να προστεθεί ή να αφαιρεθεί από το μοντέλο βάσει ενός μέτρου. Σύνηθες μέτρο, σύμφωνα με το οποίο η διαδικασία αυτή μπορεί να επιλέξει σε κάθε εφαρμογή την κατάλληλη μεταβλητή, είναι το κριτήριο πληροφορίας Akaike (AIC), το οποίο ορίζεται ως εξής:

$$AIC = -2\log L + 2d$$

όπου $\log L$ είναι ο νεπέριος λογάριθμος της μέγιστης πιθανοφάνειας του μοντέλου και d, ο αριθμός παραμέτρων του μοντέλου.

Η μέθοδος, με βάση το παραπάνω μέτρο, επιλέγει σε κάθε βήμα να παραμείνουν στο μοντέλο ο μεταβλητές που δίνουν την ελάχιστη τιμή του. Η διαδικασία προσθαφαίρεσης μεταβλητών είναι ιδιαίτερα χρήσιμη όταν ο αριθμός των υποψήφιων χαρακτηριστικών στο μοντέλο είναι αρκετά μεγάλος ώστε να εφαρμοστεί και να αξιολογηθεί από τον ερευνητή κάθε πιθανό μοντέλο. Ωστόσο, αν και προκύπτει ένα αξιόλογο μοντέλο χωρίς ιδιαίτερο πρόβλημα πολυσυγγραμικότητας, δε συλλαμβάνει τη χωρική αυτοσυσχέτιση μεταξύ των καταλοίπων.

Το κύριο μειονέκτημα των παραπάνω παραδοσιακών εργαλείων για το πρόβλημα που εξετάζουμε, αντιμετωπίζεται με την ένταξη της χωρικής πληροφορίας. Είναι σαφές, πως, εκτός των γνωρισμάτων που διαθέτει κάθε ακίνητο, παίζουν ρόλο και άλλα χαρακτηριστικά όπως είναι η τοποθεσία στην οποία βρίσκεται, κοινωνικοοικονομικά

κριτήρια κ.ά. Συνεπώς, δημιουργούνται χωρικά μοτίβα, μεταξύ των οποίων η αξία των ακινήτων διαφοροποιείται σημαντικά και καθιστά επιτακτικό να ληφθούν υπόψη στο εργαλείο πρόβλεψής της. Η υπόθεση αυτή βασίζεται στον πρώτο νόμο της γεωγραφίας, σύμφωνα με τον οποίο: Όλα σχετίζονται με οτιδήποτε άλλο, αλλά κοντινά πράγματα σχετίζονται περισσότερο από τα μακρινά πράγματα (Waldo R. Tobler).

Μία προσπάθεια να συμπεριληφθεί η χωρική ετερογένεια στο μοντέλο πρόβλεψης είναι με την μέθοδο χωρικής επέκτασης (Casetti, 1971). Η συγκεκριμένη προσέγγιση διαφοροποιείται από την απλή ή πολλαπλή παλινδρόμηση που αναφέρθηκε παραπάνω, καθώς οι παράμετροι του γραμμικού ή μη μοντέλου είναι συναρτήσεις πολυώνυμα. Για παράδειγμα, έστω ότι εφαρμόζουμε, σε μία συγκεκριμένη περιοχή, ένα μοντέλο της μορφής:

$$y(t) = \frac{1}{1+ e^{a-bt}}$$

όπου $y(t)$ η εξαρτημένη μεταβλητή που εκφράζει ποσοστό τη χρονική στιγμή t, a και b οι παράμετροι του μοντέλου.

Σύμφωνα με την παραπάνω μέθοδο, οι παράμετροι αυτές μπορούν να οριστούν ως συναρτήσεις πολυώνυμα χωρικής πληροφορίας, όπως είναι η πυκνότητα του πληθυσμού (D) που χρησιμοποιεί ο Casetti (1971) στο συγκεκριμένο παράδειγμα:

$$a = a_0 + a_1 D$$

$$b = b_0 + b_1 D$$

τα οποία όταν αντικαταστήσουμε στο μοντέλο και το φέρουμε σε γραμμική μορφή προκύπτει:

$$\ln\left[\left(\frac{1}{y}\right) - 1\right] = a_0 + a_1 D + b_0 t + b_1 t D$$

όπου η εκτίμηση των παραμέτρων γίνεται μέσω της πολλαπλής παλινδρόμησης.

Με αυτόν τον τρόπο, η μέθοδος χωρικής επέκτασης μπορεί να συλλάβει επιπλέον πληροφορία με τη βοήθεια κάποιων ποσοτήτων όπως η πυκνότητα του πληθυσμού, οι συντεταγμένες των ακινήτων κ.ά. Παρόλα αυτά, φαίνεται πως με τη χρήση πολυωνύμων υψηλού βαθμού, παρουσιάζεται το πρόβλημα της πολυσυγγραμικότητας (Pace, et al., 1998), ενώ είναι πιθανό να μη γίνεται εύκολα ξεκάθαρο στον ερευνητή ότι η χρήση τους θα ωφελήσει. Ελλοχεύει, λοιπόν, ο κίνδυνος επιλογής λανθασμένης μορφής επέκτασης που ενδέχεται να οδηγήσει σε απόκρυψη σημαντικής χωρικής πληροφορίας (Charlton and Fotheringham, 2009). Επιπλέον, υστερούν της αξιοπιστίας και έχουν την τάση να εξομαλύνουν σε μεγάλο βαθμό την χωρική ετερογένεια. Ως αποτέλεσμα, η συγκεκριμένη μέθοδος αδυνατεί να συμπεριλάβει περίπλοκες χωρικές εξαρτήσεις που τυχόν παρουσιάζονται στο πρόβλημα (Pavlov, 2000), γεγονός που την καθιστά αρκετά αδύναμη παρά την ευρεία χρήση της. Ωστόσο, αξίζει να σημειωθεί πως βελτιώνει σημαντικά το global μοντέλο, το οποίο αγνοεί πλήρως τις σχέσεις στον χώρο.

Η μέθοδος, η οποία προτείνεται και εξετάζεται στην παρούσα εργασία ανήκει στις τεχνικές χωρικής παλινδρόμησης και είναι η γεωγραφική παλινδρόμηση με χρήση μοτίβου βαθμονόμησης. Αποτελεί τη βελτίωση της προηγούμενης μεθόδου και προτάθηκε από τους Brunsdon, Fotheringham and Charlton (1998). Η γεωγραφική παλινδρόμηση με βάρη, η οποία παρουσιάζεται εκτενώς στο Κεφάλαιο 3, προτείνει να εφαρμόζεται ένα μοντέλο παλινδρόμησης σε κάθε τοποθεσία, ενώ κάθε παρατήρηση λαμβάνει και διαφορετική βαθμονόμηση με βάση την επιρροή της σε αυτή. Ειδικότερα, παρατηρήσεις που βρίσκονται πιο κοντά στην περιοχή που εξετάζουμε κάθε φορά, περιμένουμε πως θα παίρνουν μεγαλύτερο βάρος σε σχέση με πιο μακρινές. Η ιδέα πίσω από τη συγκεκριμένη μεθοδολογία είναι πως ακίνητα τα οποία βρίσκονται σε κοντινές αποστάσεις, θα επηρεάζονται στον ίδιο βαθμό από ορισμένα χαρακτηριστικά και άρα θα τείνουν λογικά να έχουν σχετικές τιμές.

Ένα βασικό πλεονέκτημα του συγκεκριμένου εργαλείου είναι το γεγονός ότι έχει τη δυνατότητα να παραλλαχθεί και να προσαρμοστεί στις παραμέτρους του. Για παράδειγμα, αναφορικά με την παράμετρο που ορίζει την απόσταση μίας παρατήρησης από το κεντρικό σημείο μιας περιοχής, δύναται να χρησιμοποιηθούν, εκτός από ευκλείδειες αποστάσεις, διαφορετικά μέτρα, τα οποία εξαρτώνται συχνά από φυσικούς ή κοινωνικοοικονομικούς παράγοντες. Πιο συγκεκριμένα, σύμφωνα με

τους Lu, et al. (2014), προτείνεται η ένταξη μη ευκλείδειων αποστάσεων οι οποίες προκύπτουν από τουριστικά αξιοθέατα, δίκτυα μετακίνησης, ποταμών ή πιο περίπλοκων γεωγραφικών συνθηκών, καθώς είναι πιο ρεαλιστικά και αντικατοπτρίζουν πραγματικές χωρικές εξαρτήσεις. Συγκρίνοντας, μάλιστα, δύο γεωγραφικά μοντέλα με βάρη και μη ευκλείδειες αποστάσεις με ένα αντίστοιχο που χρησιμοποιεί ευκλείδειες, κατέληξαν στο συμπέρασμα πως, και στις δύο περιπτώσεις, η χρήση πιο ρεαλιστικών μετρήσεων για την απόσταση βελτιώνει σημαντικά το αποτέλεσμα συγκριτικά με το τελευταίο. Ακόμη, με τη χρήση μη ευκλείδειας απόστασης σε γεωγραφικά βαθμονομημένο μοντέλο, εκτός από καλύτερη επίδοση, προέκυψαν και επιπλέον χρήσιμες πληροφορίες για τις σχέσεις ετερογένειας.

Μία ακόμη επιλογή στη συγκεκριμένη μέθοδο, είναι το εύρος ζώνης που θέτει τα όρια των περιοχών και είναι αναγκαία για την εφαρμογή των ξεχωριστών παλινδρομήσεων. Ειδικότερα, υπάρχει η δυνατότητα ο ερευνητής να ορίσει με ακρίβεια το μέγεθος κάθε γειτονιάς (fixed) ανεξαρτήτως των παρατηρήσεων που περιέχει ή αυτό να αυξομειώνεται ανάλογα με το πόσο αραιά ή πυκνά κατοικημένη είναι αντίστοιχα (adaptive). Οι Lu, et al. (2014), με βάση την ανάλυση που πραγματοποίησαν, κατέληξαν ότι το προκαθορισμένο μέγεθος μίας περιοχής προσαρμόζει καλύτερα τα δεδομένα και υπερτερεί του προσαρμοσμένου στο σύνολο.

Αναφορικά με τη στάθμιση που κάνει η συγκεκριμένη γεωγραφική παλινδρόμηση, επικρατούν κυρίως δύο τρόποι που μπορούν να υιοθετηθούν. Ο πρώτος τρόπος, προσδίδει, ουσιαστικά, βάρος μονάδα στις παρατηρήσεις μίας γειτονιάς και μηδενικό βάρος σε όσες βρίσκονται εκτός αυτής. Με αυτόν τον τρόπο, τα τελευταία σημεία, δε συμμετέχουν και ούτε επηρεάζουν την τοπική παλινδρόμηση. Ως αποτέλεσμα, αυξάνεται η πιθανότητα, σε κάποια περιοχή να παραμείνουν λίγες παρατηρήσεις και άρα να μην έχουμε αξιόπιστη ανάλυση, ενώ οι παράμετροι που εκτιμώνται θα αλλάζουν δραστικά ανάλογα με το αν μία παρατήρηση συμπεριληφθεί ή όχι στην τοπική παλινδρόμηση.

Το πρόβλημα αυτό επιλύεται με τη χρήση μίας συνεχούς συνάρτησης βαθμονόμησης, η οποία προσδίδει σε όλες τις παρατηρήσεις βάρος, το οποίο μειώνεται σταδιακά και ομαλά σε σχέση με την απόσταση. Με αυτόν τον τρόπο, η επιρροή των σημείων μειώνεται σύμφωνα με Gaussian καμπύλη όσο αυξάνεται η απόσταση από την κεντρική παρατήρηση, χωρίς, ωστόσο, να μηδενίζεται. Η συγκεκριμένη συνάρτηση

αποσύνθεσης για τα βάρη χρησιμοποιείται και αναφέρεται στις περισσότερες περιπτώσεις (Lu, et al., 2014, Sulekan and S. Jamaludin, 2020).

Συνοψίζοντας και συγκρίνοντας τις παραπάνω μεθόδους, από πλήθος ερευνών προκύπτει το συμπέρασμα πως όσες εμπεριέχουν και αξιοποιούν τη χωρική πληροφορία, αποδίδουν καλύτερα από εκείνες που δε τη λαμβάνουν υπόψη. Πιο συγκεκριμένα, οι Doumpos, et al. (2020), συμπέραναν ότι οι χωρικές παλινδρομήσεις γραμμικής μορφής δίνουν καλύτερα αποτελέσματα σε σχέση με τις αντίστοιχες global. Οι Lu, et al. (2014) κατέληξαν ότι και τα τρία μοντέλα γεωγραφικής παλινδρόμησης με χρήση βαθμονόμησης (GWR) υπερτερούν της μεθόδου ελαχίστων τετραγώνων (OLS). Ακόμη, σύμφωνα με τους Helbich and Griffith (2016), αναφορικά με την προβλεπτική ικανότητα, τα μοντέλα GWR και χωρικής επέκτασης εμφανίζονται παρόμοια. Ωστόσο, το τελευταίο δεν είναι τόσο ικανό να μοντελοποιήσει κατάλληλα την αξία συναρτήσει του χώρου καθώς η ετερογένεια που υπάρχει είναι αρκετά περίπλοκη. Η ανάλυση των Bitter, Mulligan and Dall'erba (2006) σε 11732 παρατηρήσεις ακινήτων στο Tucson της Αριζόνα, οδήγησε στο πόρισμα ότι τόσο η μέθοδος της χωρικής επέκτασης όσο και η GWR έδωσαν καλύτερα αποτελέσματα από το global μοντέλο. Μάλιστα, η μέθοδος GWR παρουσιάζει τη βέλτιστη επεξηγηματική και προβλεπτική ικανότητα συγκριτικά με τα δύο άλλα μοντέλα. Οι Szymanowski and Kryza (2011), τοποθετήθηκαν ότι, συγκριτικά με ένα global μοντέλο παλινδρόμησης, το GWR κρίνεται καταλληλότερο για τη χωρική μοντελοποίηση καθώς συλλαμβάνει τη μη στασιμότητα. Πρότειναν, ωστόσο, πως ο συνδυασμός αυτών θα μπορούσε να δώσει πιο βελτιωμένα αποτελέσματα. Ακόμη, αναφορικά με την επιλογή ενός σχεδίου βαθμονόμησης στο μοντέλο που θα χρησιμοποιηθεί, οι Doumpos, et al. (2020) οδηγήθηκαν στο πόρισμα ότι η χρήση ενός σχεδίου που προσδίδει διαφορετικά βάρη στις παρατηρήσεις, φέρνει καλύτερα αποτελέσματα, με μεγαλύτερη βελτίωση στα γραμμικά μοντέλα συγκριτικά με τα μη γραμμικά.

Συνεχίζοντας στη διερεύνηση της βιβλιογραφίας και των μεθόδων που έχουν αναπτυχθεί και αξιοποιηθεί για την πρόβλεψη της αξίας των ακινήτων, συναντάμε πιο εξελιγμένες, υπολογιστικά, μεθόδους σε σχέση με τις προαναφερθείσες. Πιο συγκεκριμένα, στην κατηγορία αυτή, εντάσσονται τεχνικές από machine learning, neural networks, μέθοδοι χωρικής ανάλυσης, random forest και decision trees.

Σε πολλές περιπτώσεις που αφορούν τον κλάδο του real estate και την πρόβλεψη της τιμής των ακινήτων, έχει γίνει χρήση των λεγόμενων neural networks. Η συγκεκριμένη

τεχνική αναπτύχθηκε με βασική ιδέα την παραγωγή ενός εργαλείου που μιμείται τη διαδικασία εκμάθησης του ανθρώπινου εγκεφάλου (Pagourtzi, et al., 2003). Για το λόγο αυτό, ένα βασικό χαρακτηριστικό της μεθόδου είναι πως δεν κρίνεται αναγκαίο να υπάρξει κάποια συνθήκη a priori σχετικά με την φύση των παρατηρήσεων (Curry, Morgan and Silver, 2002). Στο τεχνικό της κομμάτι, αποτελείται από τρία κύρια στάδια, εκ των οποίων στο δεύτερο, το οποίο συχνά αναφέρεται ως «black box», περιλαμβάνονται συναρτήσεις οι οποίες συνδέουν τα χαρακτηριστικά ενός ακινήτου με τη ζητούμενη μεταβλητή απόκρισης (τιμή). Ειδικότερα, χρησιμοποιείται μία συνάρτηση βαθμονόμησης των προσδιοριστικών μεταβλητών για τον προσδιορισμό της τιμής. Στη συνέχεια, τα αποτελέσματα αυτά συνδέονται μέσω μίας άλλης συνάρτησης με τις τιμές της εξαρτημένης μεταβλητής. Η τελευταία συνάρτηση, έχει τη δυνατότητα να φέρει πλήθος μοτίβων, ωστόσο, συνήθως είναι μη γραμμικής μορφής.

Σύμφωνα με έρευνες οι οποίες έχουν εντρυφήσει στη χρήση των neural networks για τον καθορισμό της αξίας των ακινήτων, υπάρχουν αρκετά θετικά αποτελέσματα. Ειδικότερα, η ανάλυση που πραγματοποίησαν οι Mimis, Rovolis and Stamou (2013) σε ακίνητα τα οποία εντοπίζονται στην Αθήνα, οδηγείται στο συμπέρασμα ότι neural networks μοντέλα που έχουν εμπλουτιστεί με χωρικούς όρους, δίνουν πιο ακριβείς προβλέψεις και καλύτερα αποτελέσματα συγκριτικά με πιο παραδοσιακά χωρικά μοντέλα. Οι Nguyen and Cripps (2001), κατέληξαν πως η χρήση των neural networks μοντέλων είναι αποτελεσματικότερη όταν εφαρμόζεται σε μεγάλο πλήθος παρατηρήσεων. Οι Alexandridis, et al. (2018), ακόμη, επισήμαναν πως σε αντίθεση με άλλες μελέτες, η μοντελοποίηση της τιμής των ακινήτων μέσω του neural network μοντέλου έχει συνεχώς καλύτερη απόδοση σε σχέση με πιο παραδοσιακές μεθόδους. Σε αυτό το συμπέρασμα κατέληξαν, χρησιμοποιώντας δείγμα 36527 ακινήτων από όλη την Ελλάδα, ενώ αξίζει να σημειωθεί ότι, δόθηκε ιδιαίτερη φροντίδα στην εφαρμογή του μοντέλου συνδυαστικά με επιπλέον στατιστικά εργαλεία. Λιγότερο αισιόδοξα αποτελέσματα παρουσιάζονται, ωστόσο, σε μελέτες όπως είναι των McCluskey, et al. (2013). Σύμφωνα με τα αποτελέσματα που προκύπτουν, το μοντέλο neural networks δίνει τα χειρότερα αποτελέσματα μεταξύ των μεθόδων OLS, χωρικής επέκτασης και γεωγραφικής παλινδρόμησης με βαθμονόμηση, με την τελευταία να έχει την καλύτερη απόδοση.

Επιπλέον εργαλεία τα οποία βελτιώνουν την επεξεργασία της χωρικής πληροφορίας για την πρόβλεψη της αξίας των ακινήτων, αποτελούν οι μέθοδοι χωρικής ανάλυσης.

Σε αυτές, συμπεριλαμβάνονται τεχνικές όπως spatial pattern analysis και autocorrelation analysis, variography και kriging τεχνικές που παρέχουν περισσότερες πληροφορίες αναφορικά με τον κλάδο της αγοράς. Πιο συγκεκριμένα, οι παραπάνω τεχνικές βασίζονται στην ιδέα της χωρικής παρεμβολής. Επιλέγοντας διακριτά δεδομένα από υποπεριοχές, αναπτύσσεται μία συνάρτηση που αντιπροσωπεύει με το βέλτιστο τρόπο όλες τις παρατηρήσεις και δύναται να χρησιμοποιηθεί για προβλέψεις της αξίας των ακινήτων στις υπόλοιπες υποπεριοχές (Pagourtzi, et al., 2003).

Ειδικότερα, όπως αναφέρεται στους Tiefelsdorf and Boots (1997), ο Anselin (1995), πρότεινε ένα στατιστικό εργαλείο με όνομα local Moran's I test, με τη βοήθεια του οποίου θα ελέγχεται η αυτοσυσχέτιση στον χώρο. Με τη βοήθεια αυτού, είναι εφικτό να προσδιοριστούν χωρικοί παράγοντες και μοτίβα που επηρεάζουν τον έλεγχο σε global επίπεδο. Για κάθε παρατήρηση i, το μέτρο ορίζεται ως:

$$I_i = z_i \sum_j w_{ij} z_j$$

όπου οι παρατηρήσεις z_i, z_j είναι οι αποκλίσεις από το μέσο, w_{ij} τα βάρη των παρατηρήσεων και j είναι οι παρατηρήσεις από τις οποίες απαρτίζεται η «γειτονιά» J της παρατήρησης i.

Για τον δείκτη αυτό, που θα εντοπίσει οποιαδήποτε τοπική αυτοσυσχέτιση, είναι αναγκαίο να πληρούνται δύο βασικές προϋποθέσεις. Καταρχάς, ο δείκτης πρέπει να υποδείξει ομαδοποίηση αρνητικής ή θετικής φύσεως στα κατάλοιπα (θετική χωρική αυτοσυσχέτιση) καθώς και εκείνα τα οποία δε συνδέονται με τις συστάδες (αρνητική χωρική αυτοσυσχέτιση). Δεύτερη υπόθεση, είναι πως όλοι οι τοπικοί δείκτες θα μπορούν να συνθέτουν με κατάλληλο τρόπο τον global και θα είναι εφικτό να διερευνηθεί η επιρροή τους σε αυτόν. Δηλαδή:

$$\sum_i I_i = \sum_i z_i \sum_j w_{ij} z_j$$

ενώ το global Moran's I είναι:

$$I = \sum_i I_i \; / \; [\, \sum_i \sum_j w_{ij} \, (\frac{\sum_i {z_i}^2}{n}) \,]$$

Εφόσον εντοπιστεί χωρική αυτοσυσχέτιση στα κατάλοιπα, δηλαδή τα κατάλοιπα μίας περιοχής να έχουν σημαντικό αντίκτυπο στα κατάλοιπα μίας άλλης, προτείνεται να την εντάξουμε στα μοντέλα παλινδρόμησης. Προκειμένου να συμβεί αυτό, εφαρμόζουμε στα κατάλοιπα αυτοπαλίνδρομο μοντέλο με βαθμονόμηση, η οποία εξαρτάται από την τοποθεσία τους (John Odland, 2020). Ειδικότερα, τα κατάλοιπα είναι της μορφής:

$$\varepsilon_i = \sum \rho w_{ij} \varepsilon_j + v_i$$

όπου ε_i τα κατάλοιπα, ρ παράμετρος, w_{ij} τα βάρη που ορίζονται με βάση την απόσταση ανάμεσα στην τοποθεσία i και j και v_i όρος για το σφάλμα που κατανέμεται ταυτόσημα και ανεξάρτητα.

Με τον τρόπο αυτό, γίνεται διαθέσιμο ένα πιο γενικό μοντέλο από εκείνο της απλής ή πολλαπλής γραμμικής παλινδρόμησης το οποίο συλλαμβάνει επιτυχώς τη χωρική πληροφορία. Ωστόσο, το συγκεκριμένο πρόβλημα είναι δυνατό να επιλυθεί μέσω δύο ακόμη τρόπων. Η αυτοσυσχέτιση των καταλοίπων ενδέχεται να οφείλεται στην ιδιοσυγκρασία του μοντέλου, το οποίο δεν αντικατοπτρίζει την αληθινή φύση των δεδομένων. Για παράδειγμα, είναι πιθανό να υποθέσουμε γραμμική μορφή των μεταβλητών, χωρίς αυτό να ισχύει στην πραγματικότητα. Συνεπώς, η αλλαγή της συνάρτησης σε άλλη μορφή, όπως μη γραμμική, που θα αντιπροσωπεύει καλύτερα τη σχέση την οποία ενδιαφερόμαστε να μελετήσουμε, ενδέχεται να διορθώσει και την αυτοσυσχέτιση που παρουσιάζεται στα κατάλοιπα.

Εναλλακτικά, η χωρική διασπορά που παρατηρείται στα κατάλοιπα, μπορεί να ενσωματωθεί στο μοντέλο με τη χρήση ανεξάρτητων μεταβλητών, οι οποίες προσδιορίζουν τις αλληλεπιδράσεις μεταξύ των περιοχών. Ακολουθώντας την τελευταία επιλογή, θα πρέπει κανείς να αναρωτηθεί ποιες είναι εκείνες οι μεταβλητές οι οποίες εμπεριέχουν με βέλτιστο τρόπο την επιρροή της γειτονιάς. Το συγκεκριμένο πρόβλημα, όπως αναφέρει ο Dubin (1992), έγκειται στο γεγονός πως ο ορισμός μιας «γειτονιάς» έχει διαφορετικές προσεγγίσεις ανάλογα από που προέρχονται τα δεδομένα. Εκτός από ορισμένα γεωγραφικά στοιχεία τα οποία είναι ευρέως γνωστά και αποδεκτά ως σύνορα περιοχών, είναι δύσκολο να οριστούν αντικειμενικά όρια. Συνεπώς, δημιουργούνται πιο ευέλικτοι διαχωρισμοί των περιοχών, ενώ η

λανθασμένη επιλογή αυτών ενδέχεται να οδηγήσει σε μεγάλα σφάλματα των ανεξάρτητων μεταβλητών.

Λόγω του παραπάνω μη επιθυμητού αποτελέσματος, το οποίο θα προκαλέσει ασταθείς και μεροληπτικές μετρήσεις, ο Dubin (1992), χρησιμοποιεί την τεχνική kriging. Σύμφωνα με αυτή, οποιαδήποτε επεξηγηματική μεταβλητή η οποία περιέχει χωρική πληροφορία παραλείπεται από το μοντέλο, και η ετερογένεια του χώρου λαμβάνεται υπόψη από τα κατάλοιπα. Πιο συγκεκριμένα, ο πίνακας συσχέτισης των καταλοίπων, συνδυάζεται με μία συνάρτηση που συνδέεται άμεσα με τις αποστάσεις των σημείων ανά δύο. Βασικός στόχος, είναι να πάρουμε προβλέψεις για σημεία εκτός των παρατηρήσεων που διαθέτουμε, οι οποίες προκύπτουν ως ένας βαθμονομημένος μέσος των δεδομένων. Το τελευταίο αποτέλεσμα για τα βάρη έχει υπολογιστεί έτσι ώστε, παρατηρήσεις που βρίσκονται κοντά στην τιμή που επιθυμούμε να προβλέψουμε να έχουν μεγαλύτερη επιρροή. Ακόμη, απομονωμένες παρατηρήσεις λαμβάνουν περισσότερο βάρος συγκριτικά με εκείνες που ανήκουν σε κάποια συστάδα. Αναλύοντας ένα σετ δεδομένων που αφορούν ακίνητα στη Βαλτιμόρη, Maryland, η τεχνική kriging κατέληξε σε εύλογα αποτελέσματα, καθώς αξιοποιήθηκε η χωρική αυτοσυσχέτιση και προέκυψαν σημαντικά συμπεράσματα για τις περιοχές (Dubin, 1992).

Σε συνέχεια των πιο εξελιγμένων τεχνικών για την πρόβλεψη της αξίας των ακινήτων, περνάμε σε κάποιες άλλες τεχνικές του machine learning. Αποτελούν μεθόδους, οι οποίες είναι ευρέως εφαρμοσμένες σε πολλούς τομείς όπως η διοίκηση, η μηχανική, η ιατρική κ.ά. για την εκτίμηση μελλοντικών γεγονότων, ωστόσο η χρήση τους για προβλέψεις στον τομέα του real estate είναι κάπως περιορισμένη. Στην έρευνα των Park and Bae (2015), συγκρίνονται τέσσερις διαφορετικοί αλγόριθμοι της επιστήμης του machine learning, οι οποίοι είναι οι C4.5, RIPPER, Naïve Bayesian και AdaBoost. Το μέγεθος των παρατηρήσεων που χρησιμοποιήθηκαν είναι 15135 που προέρχονται από τρεις διαφορετικές πηγές και αφορούν κυρίως σπίτια στη Virginia τα οποία έχουν αγοραστεί. Η εφαρμογή δύο ελέγχων απόδοσης αναφορικά με πόση ακρίβεια, οι παραπάνω τεχνικές, μπορούσαν να προβλέψουν αν η τιμή στην οποία πωλήθηκε μία κατοικία ήταν μικρότερη ή μεγαλύτερη εκείνης που είχε προταθεί, κατέληξαν στο ίδιο συμπέρασμα. Ο αλγόριθμος RIPPER (Repeated Incremental Pruning to Produce Error Reduction) αποδίδει καλύτερα από τους υπόλοιπους καθώς έχει το μικρότερο ποσοστό σφάλματος, ενώ ο αλγόριθμος Naïve Bayesian έχει το μεγαλύτερο.

Μία ακόμη δημοφιλής τεχνική του machine learning αποτελεί η μέθοδος random forests. Βασίζεται στην tree-based μεθοδολογία και στον συνδυασμό πολλών ανεξάρτητων μοντέλων δένδρων-παλινδρόμησης, σε ένα συνολικό σύστημα πρόβλεψης (Doumpos, et al., 2020). Πιο συγκεκριμένα, προκειμένου να κατασκευαστούν τα βασικά μοντέλα, εφαρμόζεται η μέθοδος bootstrap στα δείγματα που προέρχονται από το training sample, και το τελικό αποτέλεσμα, προκύπτει από το μέσο όρο των εκτιμήσεων των παραπάνω μοντέλων. Στη μελέτη των Doumpos, et al. (2020), όπου η συγκεκριμένη τεχνική χρησιμοποιείται συγκριτικά με μεθόδους παλινδρόμησης για την εκτίμηση της αξίας των ακινήτων, τα αποτελέσματα δεν είναι ιδιαίτερα ευνοϊκά. Ειδικότερα, ανάγονται στο συμπέρασμα πως απλούστερα γραμμικά μοντέλα, τα οποία έχουν προσεγγιστεί με τοπική βαθμονόμηση, δίνουν γενικότερα καλύτερα αποτελέσματα συγκριτικά με μη γραμμικές προσεγγίσεις της τεχνικής μάθησης όπως το random forests. Μάλιστα, οι τοπικές προσεγγίσεις στην τελευταία μέθοδο δεν επέφεραν βελτίωση, παρά μόνο οι προσθήκη χωρικών όρων στη γενική προσέγγιση (global scheme). Αντιθέτως, σε πιο αισιόδοξα συμπεράσματα αναφορικά με τη μέθοδο random forests, κατέληξαν οι Antipov and Pokryshevskaya (2010). Κάνοντας χρήση ενός σετ δεδομένων αποτελούμενο από 2845 διαμερίσματα στην Αγία Πετρούπολη, γίνεται σύγκριση μεταξύ δέκα αλγορίθμων διαφορετικής φύσεως. Το τελικό αποτέλεσμα της ανάλυσης είναι πως η τεχνική random forests παρέχει τα βέλτιστα αποτελέσματα εξ αυτών, αποδίδοντας σε μεγαλύτερο βαθμό από τις tree-based, neural networks και OLS μεθοδολογίες.

Σε γενικότερο πλαίσιο, προκύπτει ως συμπέρασμα πως η χρήση μη γραμμικών μοντέλων παρουσιάζει δυσκολίες στη σύλληψη και αξιοποίηση της χωρικής ετερογένειας που αναμένουμε να υπάρχει σε δεδομένα ακινήτων. Οι Doumpos, et al. (2020), αναφέρουν ότι οι ισχυρές χωρικές συσχετίσεις της ελληνικής αγοράς δεν αντικατοπτρίζονται πλήρως σε μη γραμμικά μοντέλα, ακόμη και με την προσθήκη χωρικών όρων. Το συμπέρασμα αυτό, συμφωνεί με τους Groben and Thomschke (2018), οι οποίοι υποστήριξαν πως «κατάλληλα σχεδιασμένα τοπικά γραμμικά μοντέλα με γεωγραφική βαθμονόμηση παρέχουν βέλτιστα αποτελέσματα».

3. ΜΕΘΟΔΟΛΟΓΙΑ

3.1. Μοντέλο πολλαπλής γραμμικής παλινδρόμησης

Το πρώτο μοντέλο που θα εφαρμόσουμε προκειμένου να το αξιοποιήσουμε για πρόβλεψη της αξίας των ακινήτων είναι της πολλαπλής γραμμικής παλινδρόμησης. Η συγκεκριμένη μέθοδος έχει ευρεία χρήση σε πλήθος προβλημάτων καθώς αποτελεί ένα προσιτό και εύκολο εργαλείο, τόσο για επεξηγηματικούς, όσο και προβλεπτικούς σκοπούς. Πιο συγκεκριμένα, η απλοϊκή μορφή του το καθιστά ικανό να παραλλαχθεί και να προσαρμοστεί ανάλογα με τη φύση της έρευνας, ενώ τα αποτελέσματα που προκύπτουν είναι εύκολα ερμηνεύσιμα και κατανοητά. Πιο συγκεκριμένα, το μοντέλο πολλαπλής γραμμικής παλινδρόμησης παίρνει την εξής μορφή:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$$

ή

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i$$

όπου Y_i είναι η αξία της παρατήρησης/ακινήτου i (i=1,2,…,n) σε λογαριθμική κλίμακα, β_0 είναι η σταθερά του μοντέλου, X_{ji} η τιμή της j-μεταβλητής για την παρατήρηση/ακίνητο i και β_j οι συντελεστές των αντίστοιχων j-μεταβλητών με j=0,1,…,k. Αντίστοιχα, με τη χρήση πινάκων, η παραπάνω εξίσωση μπορεί να γραφεί ως:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

όπου

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Αναφορικά με την εκτίμηση των παραμέτρων β_j (j = 1,2,…,k), γίνεται χρήση της μεθόδου ελαχίστων τετραγώνων (OLS). Σύμφωνα με αυτή, επιλέγουμε τα β_j τα οποία ελαχιστοποιούν το άθροισμα των τετραγώνων των αποκλίσεων της παρατηρούμενης τιμής Y_i από την αντίστοιχη προβλεπόμενη τιμή της Y, $\widehat{Y}_i$:

$$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

όπου

$$\widehat{Y}_i = b_0 + \sum_{j=1}^{k} b_j X_{ij}$$

Καταλήγουμε ότι ο εκτιμητής ελαχίστων τετραγώνων για το διάνυσμα των παραμέτρων β_j δίνεται από:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Όπως φαίνεται, οι εκτιμήσεις των συντελεστών β_j για κάθε χαρακτηριστικό του μοντέλου, είναι κοινές για όλες τις παρατηρήσεις. Πρακτικά, δηλαδή, θεωρούμε πως οι επεξηγηματικές μεταβλητές επιδρούν με τον ίδιο τρόπο στη διαμόρφωση της τιμής ενός ακινήτου ανεξάρτητα από την τοποθεσία του. Με αυτή την προσέγγιση, αγνοείται η χωρική πληροφορία η οποία στο συγκεκριμένο πρόβλημα ενδέχεται να διαφοροποιήσει τα τελικά αποτελέσματα. Είναι πιθανό να είναι πιο ρεαλιστικό το σενάριο σύμφωνα με το οποίο η αξία των ακινήτων που βρίσκονται σε κοντινές αποστάσεις θα έχουν συγκρίσιμες τιμές εφόσον τα χαρακτηριστικά τους επιδρούν με παρόμοιο τρόπο.

Τα επεξηγηματικά αποτελέσματα που θα προκύψουν από το μοντέλο πολλαπλής γραμμικής παλινδρόμησης, θεωρούμε πως είναι αξιόπιστα εφόσον ικανοποιούνται ορισμένες προϋποθέσεις. Ειδικότερα, είναι αναγκαίο να ελεγχθεί η γραμμικότητα, η κανονικότητα και η ομοσκεδαστικότητα των καταλοίπων προκειμένου να

εμπιστευθούμε το μοντέλο και να εξάγουμε συμπεράσματα βάσει αυτού. Ωστόσο, στη συγκεκριμένη εργασία, κύριος στόχος αποτελεί η εύρεση ενός αξιόλογου εργαλείου προβλεπτικής φύσης, και όχι επεξηγηματικής, για την αξία των ακινήτων. Συνεπώς, εφόσον ενδιαφερόμαστε αποκλειστικά για την προβλεπτική ικανότητα της συγκεκριμένης μεθόδου, δεν θα προχωρήσουμε σε περαιτέρω διερεύνηση και εξέταση των αντίστοιχων υποθέσεων.

Σχετικά με την επιλογή των επεξηγηματικών μεταβλητών που θα πλαισιώσουν το μοντέλο, υπάρχουν διάφορα κριτήρια που μπορούμε να συμβουλευτούμε όπως το R^2, το προσαρμοσμένο R^2, το κριτήριο πληροφορίας Akaike (AIC) και Bayes (BIC) κ.ά. Προκειμένου να βρεθεί το βέλτιστο μοντέλο, μπορεί να γίνει πλήρης απαρίθμηση των υποψήφιων μοντέλων ή να ακολουθήσουμε διαδικασίες πρόσθεσης (forward), αφαίρεσης (backward) και προσθαφαίρεσης (stepwise) μεταβλητών. Η ιδέα πίσω από αυτές τις ενέργειες επιλογής κατάλληλων μεταβλητών είναι, σε κάθε βήμα, να βρίσκονται στο μοντέλο εκείνες οι οποίες βελτιώνουν την τιμή του κριτηρίου που έχουμε επιλέξει.

Αποτελεί ενδιαφέρον ζήτημα να ανακαλύψει κανείς τα χαρακτηριστικά που επηρεάζουν τις τιμές των ακινήτων αλλά και να εμβαθύνει περισσότερο στις μεταξύ τους σχέσεις και ερμηνείες. Όπως έχει γίνει σαφές, όμως, στόχος της παρούσας εργασίας, είναι να επιλέξουμε ως ανεξάρτητες μεταβλητές εκείνες οι οποίες βελτιώνουν την προβλεπτική ικανότητα του μοντέλου για την αξία των ακινήτων. Για το λόγο αυτό, δεν θα χρησιμοποιήσουμε κάποιο από τα παραπάνω κριτήρια καθώς, με βάση εκείνα, θα καταλήξουμε σε μεταβλητές οι οποίες θα έχουν τη βέλτιστη επεξηγηματική σχέση με την αξία των ακινήτων. Στη δική μας περίπτωση, λοιπόν, παρουσιάζουμε ένα μέτρο σύμφωνα με το οποίο, κάθε φορά που εισάγουμε μία επιπλέον επεξηγηματική μεταβλητή στο μοντέλο, η προβλεπτική του ισχύς θα μεγιστοποιείται. Πρόκειται για μία διαφορετική προσέγγιση της διαδικασίας προσθήκης μεταβλητών προσαρμοσμένης στο παρόν πρόβλημα.

Χωρίζουμε, αρχικά, τα δεδομένα που είναι διαθέσιμα σε δύο επιμέρους σετ. Το πρώτο, αποτελείται από το 80% των συνολικών παρατηρήσεων οι οποίες επιλέχθηκαν με τυχαίο τρόπο (train data), ενώ το δεύτερο, από τις υπόλοιπες 20% (test data). Ο διαχωρισμός αυτός έγινε προκειμένου να εξετάζουμε, αν το μοντέλο που θα εφαρμόζουμε στο πρώτο σετ δεδομένων, θα δίνει κάθε φορά προβλέψεις για τις

υπόλοιπες παρατηρήσεις που δεν θα απέχουν πολύ από τις αντίστοιχες παρατηρηθείσες.

Πιο αναλυτικά, η διαδικασία που ακολουθούμε για την επιλογή των κατάλληλων μεταβλητών στο μοντέλο πολλαπλής παλινδρόμησης είναι:

1) Ξεκινάμε από το μοντέλο πολλαπλής γραμμικής παλινδρόμησης, εφαρμοσμένο στο πρώτο σετ δεδομένων, που περιέχει μόνο τη σταθερά.

2) Επιλέγουμε να προσθέσουμε στο παραπάνω, τη μεταβλητή εκείνη, η οποία ελαχιστοποιεί το μέσο της απόλυτης τιμής του σχετικού σφάλματος πρόβλεψης που ορίζεται ως:

$$MAPE_j = \frac{1}{n_2}\sum_{i=1}^{n_2}\left|\frac{y_i - \hat{y_i}}{y_i}\right|$$

όπου, y_i είναι η παρατηρηθείσα αξία του ακινήτου i στο δεύτερο σετ δεδομένων (test), $\hat{y_i}$ η πρόβλεψη για την παρατήρηση i που προέκυψε από το μοντέλο j ενώ το n_2 αντιπροσωπεύει το πλήθος των δεδομένων του δεύτερου σετ.

3) Με τη μεταβλητή που επιλέξαμε στο μοντέλο, επιστρέφουμε στο βήμα 2 προκειμένου να εντοπίσουμε το επόμενο χαρακτηριστικό που η προσθήκη του ελαχιστοποιεί το μέτρο που έχουμε εισάγει.

4) Επαναλαμβάνουμε την παραπάνω διαδικασία προσθήκης μεταβλητών στο μοντέλο μέχρις ότου δεν υπάρχει σημαντική βελτίωση (ελαχιστοποίηση) του MAPE.

Ακολουθώντας τα παραπάνω βήματα, καταλήγουμε σε ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης με το οποίο έχουμε τη δυνατότητα να προβλέψουμε την αξία ακινήτων που δεν συμπεριλαμβάνονται στα διαθέσιμα δεδομένα. Στη δική μας περίπτωση, το τελικό μοντέλο απαρτίζεται από μία ποσοτική μεταβλητή και πέντε ποιοτικές, τέσσερις εκ των οποίων έχουν διατάξιμο χαρακτήρα:

- τύπος περιοχής
- μέγεθος του ακινήτου σε τετραγωνικά μέτρα
- τύπος του ακινήτου

- κατάσταση συντήρησης του ακινήτου
- αριθμός υπνοδωματίων (σε κατηγορίες)
- τύπος θέρμανσης

Είναι σημαντικό να σημειωθεί πως, από ένα σημείο κι έπειτα, η προσθήκη οποιασδήποτε μεταβλητής στο ήδη υπάρχον μοντέλο βελτιώνει σε πολύ μικρό βαθμό την τιμή του MAPE. Επιθυμούμε να έχουμε το βέλτιστο προβλεπτικό μοντέλο, ωστόσο, έχει αξία να το διατηρήσουμε ταυτόχρονα φειδωλό, δηλαδή χωρίς να προσθέσουμε πολλές μεταβλητές που δεν προσφέρουν ουσιαστική βελτίωση στο αποτέλεσμα. Για παράδειγμα, στην παρούσα περίπτωση, η αμέσως επόμενη μεταβλητή που μπορεί να προστεθεί στο παραπάνω μοντέλο, αφορά τον κωδικό της περιοχής στην οποία βρίσκεται το ακίνητο, σε κατηγορίες. Η συγκεκριμένη πληροφορία, ενδέχεται να είναι αρκετά χρήσιμη καθώς συνδέει, κατά κάποιον τρόπο, την τιμή των ακινήτων με την τοποθεσία στην οποία βρίσκονται. Ωστόσο, η προσθήκη του συγκεκριμένου χαρακτηριστικού στην παλινδρόμηση που εφαρμόζουμε, βελτιώνει την προβλεπτική του ικανότητα σε βαθμό 0.6%. Για το λόγο αυτό, κρίνουμε πως θα είναι καλύτερο να παραλειφθεί από το τελικό μοντέλο.

3.2. Μοντέλο γεωγραφικής παλινδρόμησης με βαθμονόμηση

Η επόμενη μεθοδολογία που προτείνεται στην παρούσα εργασία για το πρόβλημα της πρόβλεψης της αξίας των ακινήτων είναι η γεωγραφική παλινδρόμηση με βαθμονόμηση (Geographically Weighted Regression). Το συγκεκριμένο μοντέλο, εφαρμόζεται με στόχο να επιλυθεί η δυσκολία που αντιμετωπίστηκε προηγουμένως με την αξιοποίηση της χωρικής πληροφορίας. Φαίνεται αρκετά ρεαλιστικό, τα ακίνητα τα οποία ανήκουν σε κοντινά σημεία, να εμφανίζουν παρόμοιες τιμές. Προκειμένου, λοιπόν, να ληφθεί υπόψη η διάταξή τους στον χώρο, το εργαλείο που εξετάζουμε βασίζεται στη λογική του να εφαρμοστούν ξεχωριστές πολλαπλές παλινδρομήσεις σε κάθε τοποθεσία. Πιο συγκεκριμένα, το μοντέλο γεωγραφικής παλινδρόμησης θα είναι της μορφής:

$$Y_i = \beta_{i0} + \beta_{i1}X_{i1} + \beta_{i2}X_{i2} + \dots + \beta_{ik}X_{ik} + \varepsilon_i$$

ή

$$Y_i = \beta_{i0} + \sum_{j=1}^{k} \beta_{ij}X_{ij} + \varepsilon_i$$

όπου Y_i η αξία του ακινήτου στην τοποθεσία i, β_{i0} η σταθερά του μοντέλου στην τοποθεσία i, X_{ij} οι τιμές για το j-χαρακτηριστικό στην τοποθεσία i, β_{ij} οι συντελεστές των αντίστοιχων j-μεταβλητών στην τοποθεσία i και ε_i τα κατάλοιπα στην τοποθεσία i. Η τοποθεσία i, ορίζεται με τη βοήθεια των καρτεσιανών συντεταγμένων.

Οι εκτιμητές των συντελεστών β του μοντέλου της κάθε τοποθεσίας, υπολογίζονται με παρόμοιο τρόπο όπως στην πολλαπλή γραμμική παλινδρόμηση, όπου χρησιμοποιήσαμε τη μέθοδο ελαχίστων τετραγώνων. Ωστόσο, η κύρια αλλαγή έγκειται στο γεγονός ότι εισάγεται η έννοια της βαθμονόμησης των παρατηρήσεων. Πιο συγκεκριμένα, κάθε ακίνητο που έχουμε στα δεδομένα, λαμβάνει κάποιο «βάρος» ανάλογα με την θέση του σε σχέση με την τοποθεσία i που αναφέρεται η παλινδρόμηση. Με τον τρόπο αυτό, επιτρέπουμε να έχουμε εκτιμήσεις των συντελεστών β για κάθε χαρακτηριστικό X, οι οποίες θα μεταβάλλονται στο χώρο. Θεωρούμε ότι, παρατηρήσεις που βρίσκονται πιο κοντά σε μία τοποθεσία i θα έχουν μεγαλύτερη επιρροή συγκριτικά με εκείνες που βρίσκονται σε πιο μακρινή απόσταση. Τελικά, οι εκτιμήσεις των συντελεστών β δίνονται από τον τύπο:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} = (\boldsymbol{X}'\boldsymbol{W}_{\boldsymbol{i}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_{\boldsymbol{i}}\boldsymbol{y}_{\boldsymbol{i}}$$

όπου $\boldsymbol{X}$ είναι ο πίνακας που αποτελείται από μία στήλη με μονάδες και τις τιμές των χαρακτηριστικών, $\boldsymbol{y}$ διάνυσμα με την αξία του κάθε ακινήτου, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$, διάνυσμα των k+1 συντελεστών κάθε ανεξάρτητης μεταβλητής και $\boldsymbol{W}_{\boldsymbol{i}}$ τετραγωνικός πίνακας που περιέχει τα βάρη των παρατηρήσεων αναφορικά με την τοποθεσία i που ορίζεται από τις συντεταγμένες (x_i, y_i).

Ο πίνακας $\boldsymbol{W}_{\boldsymbol{i}}$, περιέχει στη διαγώνιο το βάρος που προσδίδεται σε κάθε ακίνητο ανάλογα με τη θέση του σε σχέση με την τοποθεσία i, ενώ τα υπόλοιπα στοιχεία του είναι μηδέν:

$$\boldsymbol{W_i} = \begin{bmatrix} w_{i1} & 0 & 0 & 0 \\ 0 & w_{i2} & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & w_{in} \end{bmatrix}.$$

Εφόσον παρουσιάσαμε την κεντρική ιδέα της γεωγραφικής παλινδρόμησης με βαθμονόμηση, στη συνέχεια, είναι αναγκαίο να οριστούν οι τρεις βασικές παράμετροι για την εφαρμογή της. Αυτές είναι: (α) ο τύπος της απόστασης μεταξύ των παρατηρήσεων (distance), (β) η συνάρτηση βαθμονόμησης (kernel function) και (γ) το εύρος ζώνης κάθε περιοχής/ γειτονιάς (bandwidth).

Αναφορικά με τον τρόπο βάσει του οποίου θα οριστεί η απόσταση μίας παρατήρησης/ ακινήτου από την τοποθεσία i, το μέτρο που χρησιμοποιείται συχνότερα με καρτεσιανές συντεταγμένες είναι η Ευκλείδεια απόσταση, η οποία ορίζεται ως:

$$d_2 = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2}$$

όπου εκφράζει την απόσταση μεταξύ ενός σημείου $(x_1, x_2, \dots, x_n)$ και ενός σημείου $(y_1, y_2, \dots, y_n)$.

Το συγκεκριμένο μέτρο θα χρησιμοποιηθεί και στην παρούσα περίπτωση για τον υπολογισμό της απόστασης, ωστόσο έχει προταθεί και πλήθος άλλων. Για παράδειγμα, αντί για την Ευκλείδεια απόσταση, δύναται να χρησιμοποιηθεί η απόσταση Μανχάταν, η οποία επίσης ανήκει στην οικογένεια των αποστάσεων Minkowski, ενώ σε περίπτωση των σφαιρικών συντεταγμένων προτείνονται οι αποστάσεις μεγάλου κύκλου (Great Circle). Πιο περίπλοκα μέτρα είναι τα μη Ευκλείδεια, τα οποία ορίζουν την απόσταση των σημείων λαμβάνοντας υπόψη πιο ρεαλιστικές πληροφορίες όπως τουριστικά αξιοθέατα, δίκτυα μετακινήσεων, ποταμών ή πιο περίπλοκων γεωγραφικών συνθηκών.

Στη συνέχεια, η επόμενη παράμετρος για την εφαρμογή της γεωγραφικής παλινδρόμησης με βαθμονόμηση, είναι η συνάρτηση με βάση την οποία ορίζονται τα βάρη των παρατηρήσεων. Είναι λογικό, πως αν κάθε ακίνητο λάβει βάρος ίσο με τη μονάδα στο μοντέλο γεωγραφικής παλινδρόμησης, θα εφαρμόζουμε την κλασική πολλαπλή γραμμική παλινδρόμηση, που αποτελεί υποπερίπτωση της πρώτης και οι

συντελεστές των χαρακτηριστικών είναι ίδιοι για κάθε ακίνητο. Τα πρώτα σχέδια βαθμονόμησης αφορούν συνεχείς συναρτήσεις όπως η Γκαουσιανή και η εκθετική. Με τη χρήση αυτών, το σημείο βαθμονόμησης i κάθε φορά (calibration point) θα παίρνει βάρος ίσο με τη μονάδα, ενώ η επιρροή των υπόλοιπων σημείων θα μειώνεται σταδιακά και ομαλά όσο αυξάνει η απόσταση. Πιο συγκεκριμένα, η Γκαουσιανή συνάρτηση βαθμονόμησης, σύμφωνα με την οποία θα ορίσουμε τα βάρη στην παρούσα εργασία, είναι της μορφής:

$$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right),$$

ενώ αντίστοιχα η εκθετική:

$$w_{ij} = \exp\left(-\frac{|d_{ij}|}{b}\right)$$

όπου και στις δύο περιπτώσεις, w_{ij} τα στοιχεία του πίνακα W_i της τοποθεσίας i, d_{ij} η απόσταση μεταξύ της παρατήρησης j και της τοποθεσίας i και b το εύρος ζώνης της περιοχής.

Από την άλλη, υπάρχει η επιλογή το σχέδιο βαθμονόμησης των παρατηρήσεων να βασίζεται σε μη συνεχείς συναρτήσεις όπως η Box-car και η Bi-square. Σύμφωνα με την πρώτη, οι παρατηρήσεις που εμπεριέχονται στο εύρος ζώνης που έχουμε ορίσει, και θα παρουσιαστεί αναλυτικά στη συνέχεια, θα λαμβάνουν μοναδιαίο βάρος, ενώ όσες βρίσκονται εκτός αυτού θα έχουν μηδενικό βάρος:

$$w_{ij} = \begin{cases} 1, & |d_{ij}| < b \\ 0, & διαφορετικά \end{cases}$$

Η Bi-square συνάρτηση βασίζεται στην ίδια ιδέα, με τη διαφορά πως η βαθμονόμηση των σημείων μέσα στο εύρος ζώνης μειώνεται σταδιακά όσο αυξάνει η απόσταση από το σημείο βαθμονόμησης:

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{b}\right)^2\right)^2, & |d_{ij}| < b \\ 0, & διαφορετικά \end{cases}$$

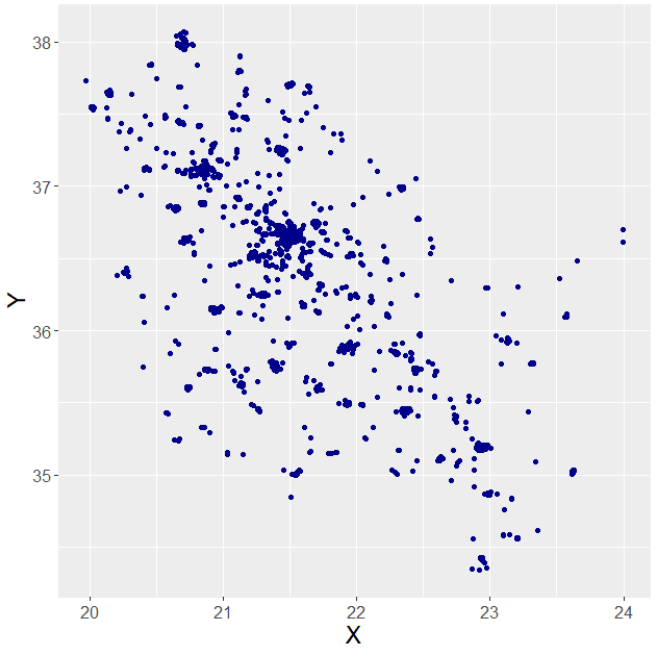
Η τρίτη και πιο σημαντική παράμετρος που είναι αναγκαίο να οριστεί στη γεωγραφική παλινδρόμηση με βαθμονόμηση είναι το εύρος ζώνης (bandwidth), το οποίο αποτελεί και καθοριστικό στοιχείο της συνάρτησης στάθμισης. Όπως στις μη συνεχείς, έτσι και στις συνεχείς συναρτήσεις βαθμονόμησης, το εύρος ζώνης μπορεί να προσεγγιστεί με δύο τρόπους. Ειδικότερα, ο πρώτος τρόπος υποδεικνύει το εύρος ζώνης να οριστεί ως μία συγκεκριμένη απόσταση (fixed bandwidth), ανεξάρτητα από το πόσες παρατηρήσεις θα περιέχει κάθε φορά. Με αυτόν τον τρόπο, είναι λογικό, περιοχές στις οποίες παρατηρούμε περισσότερα ακίνητα, να κάνουν χρήση περισσότερων σημείων στην αντίστοιχη παλινδρόμηση, ενώ πιο αραιά κατοικημένες περιοχές, λιγότερα. Στη δεύτερη περίπτωση, ελλοχεύει ο κίνδυνος να έχουμε μη επαρκή αριθμό παρατηρήσεων σε κάποια τοπική γεωγραφική παλινδρόμηση με αποτέλεσμα να προκύψουν μεγάλα σφάλματα στους αντίστοιχους εκτιμητές.

Ο δεύτερος τρόπος αναφέρεται σε ένα προσαρμοστικό εύρος ζώνης (adaptive bandwidth), το οποίο θα περιλαμβάνει συγκεκριμένο αριθμό παρατηρήσεων και άρα τα όρια του θα μεταβάλλονται ανάλογα. Αυτό σημαίνει πως, το μέγεθος της ζώνης θα αυξάνεται όταν τα παρατηρούμενα ακίνητα είναι πιο αραιά κατανεμημένα, ενώ θα μειώνεται όταν είναι πιο πυκνά. Η συγκεκριμένη επιλογή, μας εξασφαλίζει επαρκή αριθμό παρατηρήσεων για να εφαρμοστεί η τοπική παλινδρόμηση και να προκύψουν αξιόπιστα αποτελέσματα.

Το εύρος ζώνης μιας περιοχής/ γειτονιάς, είναι δυνατό να οριστεί τόσο από τον ερευνητή όσο και μέσω μιας αυτοματοποιημένης διαδικασίας. Πιο συγκεκριμένα, μέσω συναρτήσεων και κριτηρίων όπως είναι το κριτήριο πληροφορίας Akaike ή της διαδικασίας cross validation, προκύπτει το βέλτιστο εύρος ζώνης για το αντίστοιχο μοντέλο γεωγραφικής παλινδρόμησης που εφαρμόζουμε.

4. ΔΕΔΟΜΕΝΑ

Τα δεδομένα που έχουμε διαθέσιμα αφορούν σπίτια στην Ελλάδα αλλά για λόγους εμπιστευτικότητας και για να μην είναι εφικτή η ταυτοποίηση τους, τα γεωγραφικά τους χαρακτηριστικά έχουν αλλοιωθεί με τέτοιον τρόπο ώστε να είναι αδύνατη η εύρεση των γεωγραφικών συντεταγμένων. Αυτός ήταν και ο βασικός όρος για να αποκτήσουμε πρόσβαση στα δεδομένα της εταιρείας. Παρόμοια αλλοίωση έχει συμβεί και σε κάποιες κωδικοποιήσεις των μεταβλητών.



Διάγραμμα 1. Τα ακίνητα στο χώρο

Το δείγμα που χρησιμοποιείται για την εφαρμογή των μοντέλων απαρτίζεται από 3978 ακίνητα. Τα δεδομένα αυτά έχουν συλλεχθεί μετά την εμφάνιση της οικονομικής κρίσης σε παγκόσμιο αλλά και πανελλήνιο επίπεδο και πιο συγκεκριμένα, κατά την περίοδο 2010-2019. Συνεπώς, δεν θα προχωρήσουμε σε σύγκριση ή συμπερασματολογία για τις τιμές των ακινήτων αναφορικά με το συγκεκριμένο γεγονός. Το μέγεθος του δείγματος είναι αξιόλογο συγκριτικά με την πλειοψηφία των

ερευνών στο συγκεκριμένο τομέα και είναι αρκετά ενδιαφέρον να εξετάσουμε τα συμπεράσματα που θα προκύψουν βάσει αυτών.

Όνομα μεταβλητής	Περιγραφή
IID	Κωδικός ακινήτου
year	Έτος εκτίμησης
month	Μήνας εκτίμησης
areacode	Κωδικός περιοχής
urban	Τύπος πόλης (Διατάξιμη)
price	Τιμή του ακινήτου (σε €)
type	Τύπος ακινήτου (Κατηγορική)
sqm	Έκταση ακινήτου (σε m^2)
landarea	Έκταση γης/κήπου (σε m^2)
yearofconstr	Έτος κατασκευής
floor	Όροφος
totalfloors	Σύνολο ορόφων του κτιρίου
parking	Θέσεις στάθμευσης
heating	Τύπος θέρμανσης
bedrooms2	Δωμάτια (Κατηγορική)
storage	Αποθήκη (Ναι/Όχι)
maintenance	Συντήρηση (Διατάξιμη)
view	Θέα (Ναι/Όχι)

Πίνακας 1. Αρχικές μεταβλητές

Αναφορικά με τα γνωρίσματα των ακινήτων, έχουμε διαθέσιμα τόσο ποσοτικά όσο και ποιοτικά χαρακτηριστικά, τα οποία περιγράφονται στον Πίνακα 1. Προκειμένου να εντοπιστούν οι καταλληλότερες μεταβλητές που αφορούν τη βελτίωση της προβλεπτικής ικανότητας του εκάστοτε μοντέλου, έγιναν επιπλέον τροποποιήσεις των αρχικών μεταβλητών. Ωστόσο, δεν παρουσιάζονται στον παραπάνω πίνακα, παρά μόνο η αλλαγή που αφορά των αριθμό των δωματίων του ακινήτου (bedrooms). Στα αρχικά χαρακτηριστικά, αποτελεί μία ποσοτική μεταβλητή, όμως στα μοντέλα παλινδρόμησης προστίθεται ως μεταβλητή διατάξιμης φύσεως (bedrooms2). Ακόμη, σημειώνεται πως οι τιμές των ακινήτων, τις οποίες ενδιαφερόμαστε να προβλέψουμε, είναι συγκρίσιμες μεταξύ τους ανεξάρτητα του έτους εκτίμησής τους, καθώς είναι όλες αντιστοιχισμένες με περίοδο βάσης το Δεκέμβριο του 2019. Τέλος, για κάθε ακίνητο δίνονται οι καρτεσιανές συντεταγμένες του έτσι ώστε να εντοπιστεί και να αξιοποιηθεί η πληροφορία της θέσης του στο χώρο.

Αριθμητική	Ελάχιστη τιμή	Μέση τιμή	Μέγιστη τιμή	Τυπική απόκλιση
price	5000	55869.72	490000	779.27
sqm	12.49	89.41	345.6	0.94
landarea	0	213.98	17563	10.79
yearofconstr	1867	1984	2019	0.36
floor	-1	1.99	19	0.04
totalfloors	0	3.64	23	0.06
parking	0	0.34	2	0.01
Κατηγορική	**Επικρατούσα τιμή**			
urban	Πρωτεύουσα			
type	Διαμέρισμα			
bedrooms2	2			
maintenance	Μεσαία			
heating	Κεντρική θέρμανση			
Δίτιμη	**Ποσοστό 1 (Ναι)**	**Ποσοστό 2 (Όχι)**		
storage	99.6%	0.4%		
view	82%	18%		

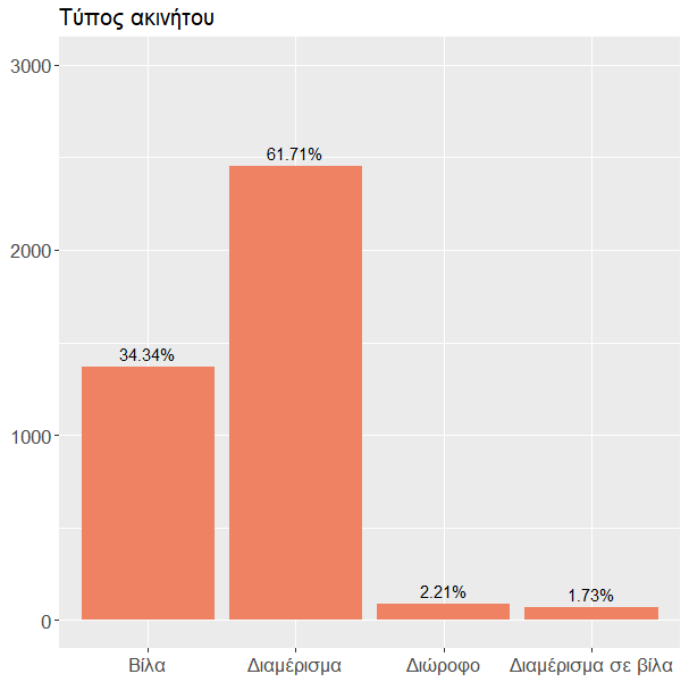
Πίνακας 2. Περιγραφικά μέτρα των μεταβλητών

Σύμφωνα με τα περιγραφικά μέτρα, τα οποία παρουσιάζονται στον Πίνακα 2, το δείγμα μας αποτελείται από ακίνητα που, κατά μέσο όρο, είναι κατασκευασμένα το 1984 και κοστίζουν 56000€. Η αξία τους, ωστόσο, την οποία ενδιαφερόμαστε να προβλέψουμε, φαίνεται πως έχει μεγάλη διακύμανση, καθώς κυμαίνεται από 5000€ έως και 490000€. Ακόμη, συμπεριλαμβάνονται παρατηρήσεις οι οποίες αφορούν τόσο μεγάλα, σε ηλικία, σπίτια όσο και πιο πρόσφατα κατασκευασμένα, με το τελευταίο έτος κατασκευής να είναι το 2019. Αναφορικά με την έκταση των ακινήτων σε τετραγωνικά μέτρα, χαρακτηριστικό που προκύπτει ότι έχει βασικό ρόλο στη διαμόρφωση της τιμής τους, παρατηρούμε ορισμένα με έκταση μόλις 12.5m^2, αλλά και άλλα που ξεπερνούν τα 300m^2. Ακόμη, σύμφωνα με τον πίνακα, οι ιδιοκτησίες που μελετάμε στην παρούσα εργασία, εντοπίζονται, κατά μέσο όρο, στον δεύτερο όροφο και διαθέτουν από μηδέν έως δύο θέσεις παρκαρίσματος.

Εστιάζουμε, στη συνέχεια, την προσοχή μας στις κατηγορικές μεταβλητές, πολλές από τις οποίες έχουμε συμπεριλάβει στα μοντέλα παλινδρόμησης προκειμένου να καταλήξουμε σε βελτιωμένες προβλέψεις της αξίας. Με τη βοήθεια των περιγραφικών μέτρων, παρατηρούμε πως ένα πολύ υψηλό ποσοστό των ιδιοκτησιών παρέχει επιπλέον αποθηκευτικό χώρο, ενώ μόνο το 18% του συνόλου δεν βλέπει σε θέα. Η

πλειοψηφία των παρατηρήσεων, ακόμη, διαθέτει δύο υπνοδωμάτια, με τον ανώτερο αριθμό αυτών να αγγίζει τα έξι με επτά.

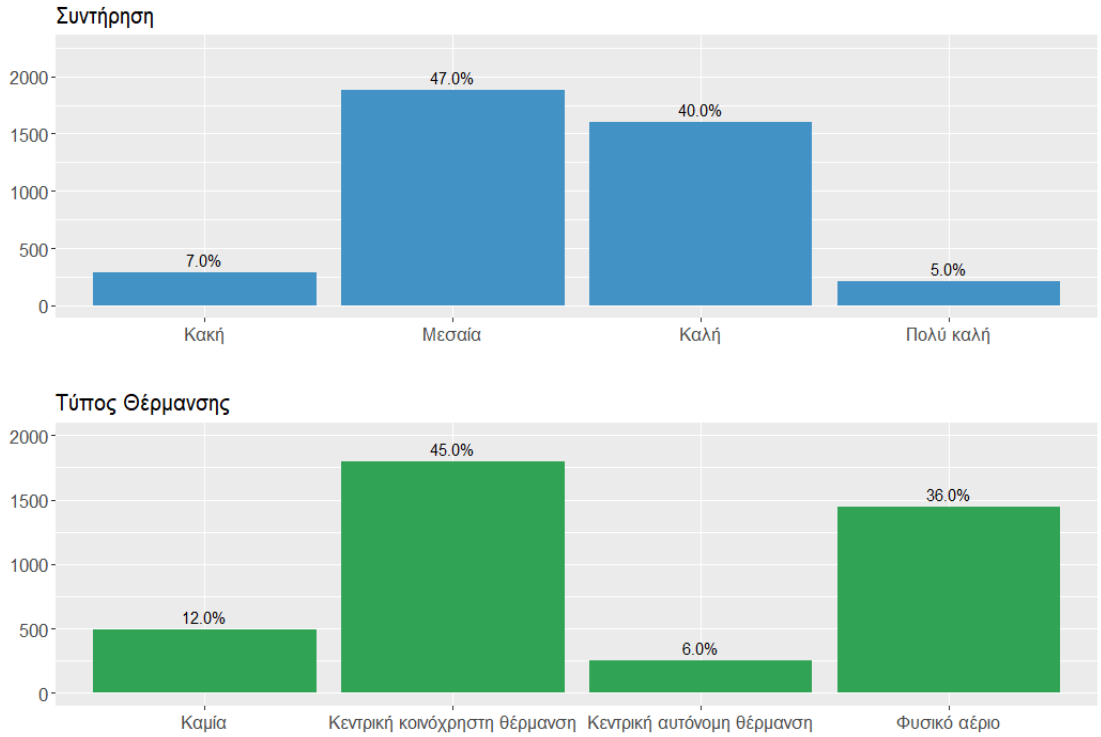
Μία ακόμη σημαντική πληροφορία σχετικά με τα ακίνητα τα οποία διαμορφώνουν το δείγμα μας, είναι ο τύπος τους. Όπως φαίνεται στο Διάγραμμα 2, οι περισσότερες παρατηρήσεις αφορούν διαμερίσματα, ενώ ακολουθούν οι βίλες με ποσοστό 34%. Αντιθέτως, τα διώροφα ακίνητα, καθώς και τα διαμερίσματα που συναντώνται σε βίλες, αποτελούν ένα πολύ μικρό μέρος του συνολικού δείγματος, με ποσοστά που δεν ξεπερνούν το 3%.



Διάγραμμα 2. Ραβδόγραμμα για τον τύπο ακινήτου

Από το Διάγραμμα 3, αντλούμε πληροφορίες σχετικά με την κατάσταση στην οποία διατηρείται το ακίνητο αλλά και τον τύπο θέρμανσης. Πιο συγκεκριμένα, το 88% των ιδιοκτησιών που έχουμε διαθέσιμες βρίσκονται σε μεσαία ή καλή κατάσταση, ενώ 209 από το συνολικό δείγμα, διατηρούνται σε πολύ καλή κατάσταση. Αναμένουμε, πως οι συνθήκες στις οποίες συναντώνται τα ακίνητα, επηρεάζουν ανάλογα και την αξία τους στον κτηματομεσιτικό τομέα. Αναφορικά με τον τύπο θέρμανσης, το 45% των ακινήτων που έχουμε στο δείγμα μας διαθέτουν κεντρική κοινόχρηστη θέρμανση, ενώ περίπου

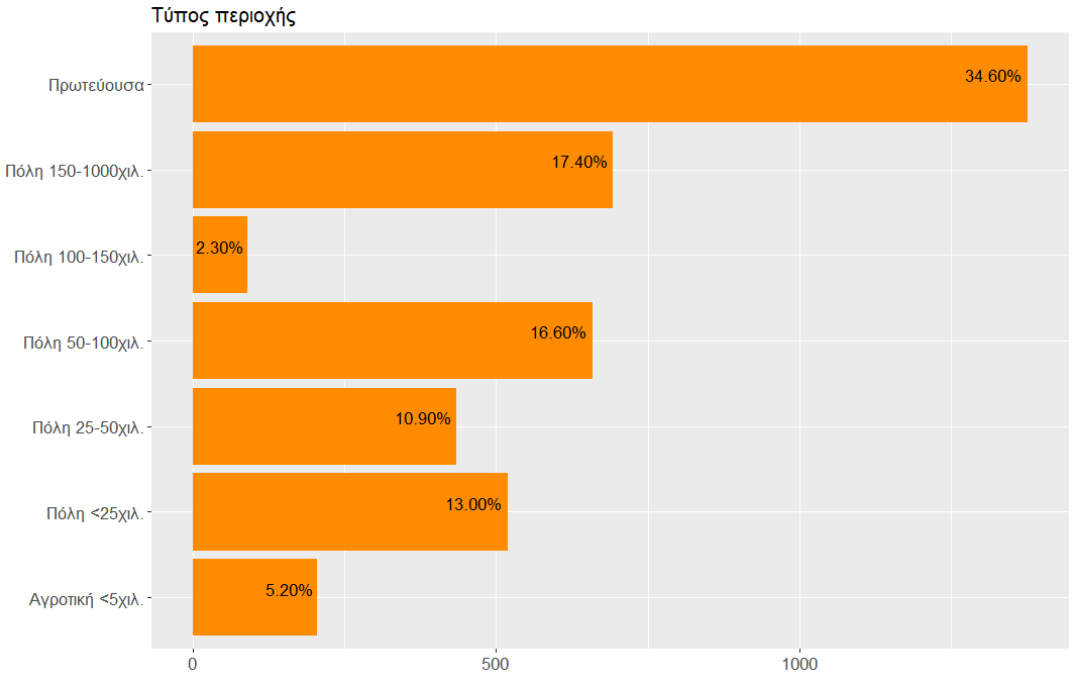
1500 από αυτά έχουν εγκατεστημένο φυσικό αέριο. Τέλος, σημειώνεται πως στο 12% των ιδιοκτησιών δεν παρέχεται κεντρική θέρμανση.



Διάγραμμα 3. Ραβδογράμματα για την κατάσταση συντήρησης και τον τύπο θέρμανσης των ακινήτων

Τελευταίο αλλά εξίσου σημαντικό στοιχείο, είναι ο τύπος της περιοχής στην οποία βρίσκεται κάθε ακίνητο. Όπως είναι λογικό, περίπου το 1/3 των παρατηρήσεων που διαθέτουμε, συναντάται σε πρωτεύουσες και ένας μικρός αριθμός αυτών βρίσκεται σε αγροτική περιοχή με λιγότερους από πέντε χιλιάδες κατοίκους (Διάγραμμα 4). Ικανοποιητικό ποσοστό των ακινήτων εντοπίζεται, στη συνέχεια, σε πόλεις με 150-1000 και 50-100 χιλιάδες κατοίκους, ενώ οι λιγότερες ιδιοκτησίες είναι σε αστική περιοχή με πληθυσμό 100-150 χιλιάδων. Το συγκεκριμένο χαρακτηριστικό, κατέχει πολύ σημαντικό ρόλο στο πρόβλημα που εξετάζουμε για την πρόβλεψη της αξίας των ακινήτων. Όπως και ο κωδικός της περιοχής μίας ιδιοκτησίας, έτσι και ο τύπος της περιοχής που εντοπίζονται, συλλαμβάνουν ένα μέρος της χωρικής πληροφορίας, την οποία θέλουμε να αξιοποιήσουμε κατάλληλα για βέλτιστες προβλέψεις. Στο δείγμα μας, περιλαμβάνονται παρατηρήσεις από επτά διαφορετικούς τύπους περιοχών και

40 κωδικούς τοποθεσίας. Αξίζει να σημειωθεί, μάλιστα, πως η μεταβλητή που αφορά τον τύπο της περιοχής, προστέθηκε αρχικά στο πρώτο μοντέλο παλινδρόμησης που εφαρμόσαμε, κάνοντας αμέσως αισθητή την χρησιμότητά της.



Διάγραμμα 4. Ραβδόγραμμα για τον τύπο περιοχής των ακινήτων

Εκτός από τη διερεύνηση και κατανόηση των δεδομένων που έχουμε διαθέσιμα, είναι χρήσιμο να εξεταστεί εάν υφίσταται κάποιο χωρικό μοτίβο στη μεταβλητή που ενδιαφερόμαστε να προβλέψουμε. Πιο συγκεκριμένα, μας ενδιαφέρει να εντοπίσουμε αν η τιμή των ακινήτων, διαφοροποιείται στον χώρο έχοντας συγκεκριμένη μορφή και όχι με τυχαιοποιημένο τρόπο. Για παράδειγμα, οι τιμές των ακινήτων στις παρατηρήσεις που βρίσκονται βόρεια των δεδομένων ενδέχεται να έχουν αυξημένες τιμές συγκριτικά με εκείνες που παρατηρούνται σε νοτιότερα σημεία. Με αυτόν τον τρόπο, παίρνουμε μία πρώτη ιδέα για το αν υπάρχει χωρική συσχέτιση, έτσι ώστε να τη λάβουμε υπόψη στη μοντελοποίηση της αξίας των ακινήτων, διορθώνοντας ως προς τα χαρακτηριστικά που την επηρεάζουν. Από το Διάγραμμα 5, παρατηρούμε πως ακίνητα με αξία μεγαλύτερη των 67000€, εντοπίζονται, κατά κύριο λόγο στα κεντρικά σημεία του συνόλου. Το ίδιο μπορεί να ισχυριστεί κανείς και για ιδιοκτησίες που κοστίζουν από 42000 έως 67000€, σε λιγότερο, ωστόσο βαθμό. Αντιθέτως, φαίνεται

πως παρατηρήσεις με μικρή αξία κατανέμονται τυχαία στον χώρο, χωρίς να ακολουθούν κάποιο συγκεκριμένο μοτίβο.

Προκειμένου να ποσοτικοποιήσουμε τα παραπάνω συμπεράσματα που προκύπτουν από το διάγραμμα και να έχουμε μία πιο καθαρή εικόνα, υπολογίζουμε τον δείκτη Moran's ως ένα μέτρο χωρικής αυτοσυσχέτισης σε όλο το πλήθος των δεδομένων. Με βάσει αυτό, έχουμε τη δυνατότητα να ελέγξουμε πόσο συσχετίζονται οι τιμές των ακινήτων ανάλογα με την τοποθεσία στην οποία βρίσκονται. Το εύρος των πιθανών τιμών του παραπάνω δείκτη είναι από -1 έως 1. Η τιμή 0 υποδηλώνει πως δεν υπάρχει χωρική συσχέτιση και άρα οι τιμές της μεταβλητής που εξετάζουμε εντοπίζονται με τυχαίο τρόπο. Από την άλλη, τιμές του δείκτη κοντά στο 1, δίνουν ένδειξη θετικής χωρικής αυτοσυσχέτισης, ενώ κοντά στο -1, αρνητικής χωρικής αυτοσυσχέτισης. Ο δείκτης Moran's I ορίζεται ως:

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

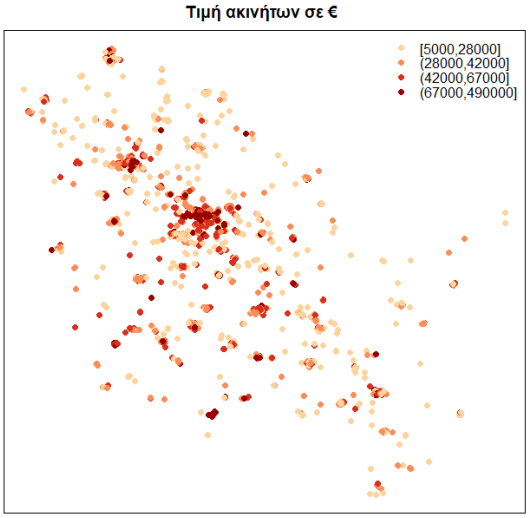
όπου, N ο αριθμός των παρατηρήσεων, x_i και x_j οι παρατηρούμενες τιμές της μεταβλητής που εξετάζουμε στις τοποθεσίες i και j αντίστοιχα και w_{ij} είναι τα στοιχεία του πίνακα W που περιέχει τα βάρη που αντιστοιχούν σε κάθε παρατήρηση.

Στην παρούσα εφαρμογή του μέτρου, ο πίνακας W περιέχει μηδενικά στη διαγώνιο $(w_{ii} = 0)$ εξ'ορισμού και τα υπόλοιπα στοιχεία υπολογίζονται με τη βοήθεια της γκαουσιανής συνάρτησης που παρουσιάστηκε προηγουμένως, έχοντας ορίσει ως εύρος ζώνης μίας περιοχής την απόσταση 0.23. Συνεπώς, σημεία τα οποία βρίσκονται σε κοντινότερες αποστάσεις, είναι λογικό να έχουν αυξημένα βάρη σε σχέση με τα πιο απομακρυσμένα. Αξίζει να σημειωθεί πως η επιλογή του πίνακα των βαρών για τις παρατηρήσεις επηρεάζει σημαντικά το αποτέλεσμα του δείκτη. Άρα, αν επιλέξουμε διαφορετική συνάρτηση βαθμονόμησης των ακινήτων, αναμένουμε διαφορετικές τιμές. Τα αποτελέσματα του συγκεκριμένου δείκτη παρουσιάζονται στον Πίνακα 3. Όπως φαίνεται, λαμβάνει την τιμή 0.1884172, όπου αποτελεί ένδειξη πως, έστω και σε μικρό βαθμό, η αξία των ακινήτων παρουσιάζει θετική αυτοσυσχέτιση στον χώρο. Αυτό σημαίνει, ουσιαστικά, πως ιδιοκτησίες που βρίσκονται σε κοντινά σημεία τείνουν να έχουν παρόμοιες τιμές. Άρα, είναι ενδιαφέρον να εξετάσουμε αν και σε τι

βαθμό, η ένταξη της χωρικής πληροφορίας στο μοντέλο θα βελτιώσει τις εκτιμήσεις της αξίας των ακινήτων. Υπολογίζεται ακόμη η αναμενόμενη τιμή του ως:

$$E(I) = \frac{-1}{N-1}$$

καθώς η τυπική απόκλιση και το p-value, που αναφέρεται στον έλεγχο της υπόθεσης πως δεν υπάρχει χωρική αυτοσυσχέτιση, έναντι της εναλλακτικής ότι υπάρχει θετική χωρική αυτοσυσχέτιση. Ωστόσο, ο παραπάνω έλεγχος βασίζεται στην υπόθεση κανονικότητας της μεταβλητής που εξετάζουμε, η οποία δεν ικανοποιείται στη δική μας περίπτωση. Συνεπώς, δεν είμαστε σε θέση να εμπιστευτούμε και να ερμηνεύσουμε το συγκεκριμένο αποτέλεσμα.



Διάγραμμα 5. Τιμή των ακινήτων σε €

Moran's I	
Παρατηρούμενη τιμή	0.1884172
Αναμενόμενη τιμή	-0.0002514458
Τυπική απόκλιση	0.001361685
p-value	<0.001

Πίνακας 3. Αποτελέσματα του δείκτη Moran's

Το παραπάνω σετ παρατηρήσεων που μελετήσαμε εκτενώς, χωρίζεται με τυχαιοποιημένη διαδικασία σε δύο μέρη. Το πρώτο (train set), χρησιμοποιείται για την εφαρμογή και εκτίμηση των παραμέτρων β των δύο προβλεπτικών μοντέλων που εξετάζουμε και αποτελείται από 3182 ακίνητα. Στη συνέχεια, με τη χρήση των παλινδρομήσεων, προβλέπουμε τις τιμές των ακινήτων που αφορούν το δεύτερο μέρος (test set) των δεδομένων και αποτελείται από 796 παρατηρήσεις. Σημειώνεται, λοιπόν, πως η παρακάτω ανάλυση και τα συμπεράσματα βάσει αυτής, προκύπτουν από ένα συγκεκριμένο training και test δείγμα.

5. ΑΠΟΤΕΛΕΣΜΑΤΑ

Έχοντας παρουσιάσει αναλυτικά τη μεθοδολογία καθώς και τα δεδομένα που είναι διαθέσιμα για την ανάλυσή μας στα προηγούμενα κεφάλαια, θα προχωρήσουμε στη μελέτη των αποτελεσμάτων στα οποία καταλήγουμε. Η πρώτη πρόταση που γίνεται στην παρούσα εργασία σχετικά με την πρόβλεψη των τιμών των ακινήτων, είναι η χρήση πολλαπλής γραμμικής παλινδρόμησης, η εφαρμογή της οποίας παρουσιάζεται στον Πίνακα 4. Υπενθυμίζεται πως, μέσω αυτής της προσέγγισης, οι συντελεστές β του κάθε χαρακτηριστικού που συμμετέχει στο μοντέλο, και επηρεάζει τον λογάριθμο της τιμής, είναι ίδιοι για όλα τα ακίνητα του δείγματος, ανεξαρτήτως της τοποθεσίας τους (global regression). Για παράδειγμα, το κόστος του τετραγωνικού για κάθε ιδιοκτησία είναι 1.006606 (exp(0.0065846)) ανεξαρτήτων των άλλων γνωρισμάτων ή της θέσης του.

Για την ένταξη των κατηγορικών μεταβλητών στο μοντέλο γραμμικής παλινδρόμησης, έχουν δημιουργηθεί $k-1$ ψευδομεταβλητές για κάθε χαρακτηριστικό, όπου k αναφέρεται στο πλήθος των επιπέδων/κατηγοριών της κάθε μεταβλητής. Η εκτίμηση της σταθεράς του μοντέλου (β_0=8.47) αφορά την αναμενόμενη τιμή, σε λογαριθμική κλίμακα, ιδιοκτησιών με μηδενική έκταση, που βρίσκονται σε αγροτική περιοχή, είναι βίλες σε κακή κατάσταση, χωρίς δωμάτια και θέρμανση (επίπεδα αναφοράς των ποιοτικών μεταβλητών).

Σύμφωνα με τις τιμές των p-value για τις παραμέτρους β του μοντέλου, σε επίπεδο σημαντικότητας $\alpha = 0.05$, οι ψευδομεταβλητές που αφορούν ιδιοκτησίες με φυσικό αέριο για θέρμανση και 7 υπνοδωμάτια αντίστοιχα, είναι στατιστικά μη σημαντικές. Η εκτίμηση του συντελεστή της δεύτερης, ωστόσο, προέκυψε από μία μόνο παρατήρηση που ανήκει σε αυτή την κατηγορία. Ακόμη, φαίνεται πως το μοντέλο προσαρμόζει καλύτερα τα δεδομένα συγκριτικά με το σταθερό (p-value<0.001), ενώ το 76% της μεταβλητότητας της αξίας των ακινήτων εξηγείται από τη συγκεκριμένη παλινδρόμηση. Παρ ’όλα αυτά, ενδιαφερόμαστε για την προβλεπτική και όχι ερμηνευτική ικανότητα του μοντέλου, και άρα θα το αξιολογήσουμε με τη βοήθεια άλλων μέτρων.

Μοντέλο: log(price) ~ urban+sqm+type+maintenance+bedrooms2+heating				
Κατάλοιπα				
Ελάχιστη τιμή -1.36066	**1ο τεταρτημόριο** -0.20501	**Διάμεσος** -0.00184	**3ο τεταρτημόριο** 0.20287	**Μέγιστη τιμή** 1.45487
Συντελεστές				
	Εκτίμηση	**Τυπικό σφάλμα**	**t value**	**p-value**
Σταθερά	8.4736596	0.0476429	177.858	<2e-16
urban<25k	0.2363725	0.0316201	7.475	9.91e-14
urban25-50	0.1786991	0.0322557	5.540	3.27e-08
urban50-100	0.3209931	0.0308212	10.415	<2e-16
urban100-150	0.2648468	0.0472653	5.603	2.28e-08
urban150-1000	0.4973451	0.0316998	15.689	<2e-16
urbancapital	1.0291128	0.0309255	33.277	<2e-16
sqm	0.0065846	0.0001743	37.771	<2e-16
typeapartment	0.4180796	0.0210690	19.843	<2e-16
typeduplex	0.4905609	0.0448214	10.945	<2e-16
typeapartment in villa	0.3261878	0.0465530	7.007	2.97e-12
maintenancemedium	0.2307582	0.0243771	9.466	<2e-16
maintenancegood	0.4145198	0.0253418	16.357	<2e-16
maintenanceverygood	0.7179428	0.0353196	20.327	<2e-16
bedrooms21	0.2639173	0.0317380	8.316	<2e-16
bedrooms22	0.4314485	0.0323274	13.346	<2e-16
bedrooms23	0.5326175	0.0362131	14.708	<2e-16
bedrooms24	0.4950793	0.0428548	11.552	<2e-16
bedrooms25	0.4299056	0.0507005	8.479	<2e-16
bedrooms26	0.3194603	0.0614014	5.203	2.09e-07
bedrooms27	0.2522322	0.3410299	0.740	0.459586
heatinglowshared	0.1923058	0.0211916	9.075	<2e-16
heatinglowauto	0.1146407	0.0311519	3.680	0.000237
heatinghigh	0.0211611	0.0203971	1.037	0.299603
Τυπικό σφάλμα καταλοίπων: 0.3371 σε 3158 βαθμούς ελευθερίας				
R^2= 0.7615		Προσαρμοσμένο R^2= 0.7598		
F-statistic: 438.4 σε 23 και 3158 βαθμούς ελευθερίας, p-value: < 2.2e-16				

Πίνακας 4. Αποτελέσματα του μοντέλου πολλαπλής γραμμικής παλινδρόμησης

Στη συνέχεια, προκειμένου να αξιοποιήσουμε την χωρική πληροφορία, που αγνοείται στην παραπάνω πολλαπλή γραμμική παλινδρόμηση, και να προβλέψουμε την αξία των ιδιοκτησιών, εφαρμόζουμε το μοντέλο γεωγραφικής παλινδρόμησης με βαθμονόμηση. Οι μεταβλητές που συμπεριλαμβάνονται στο μοντέλο, τόσο οι ποσοτικές όσο και οι ψευδομεταβλητές για τα ποιοτικά γνωρίσματα, είναι οι ίδιες με εκείνο της πολλαπλής γραμμικής παλινδρόμησης. Αναφορικά με τις αποστάσεις μεταξύ των σημείων, επιλέγουμε να υπολογιστούν μέσω της ευκλείδειας απόστασης

με τη βοήθεια των καρτεσιανών συντεταγμένων που προσδιορίζουν την τοποθεσία της κάθε παρατήρησης. Η συνάρτηση βαθμονόμησης, σύμφωνα με την οποία ορίζονται τα βάρη της κάθε παρατήρησης και διαμορφώνουν τον πίνακα W, είναι η γκαουσιανή. Η τελευταία, και ίσως σημαντικότερη, παράμετρος που είναι αναγκαίο να καθοριστεί για την εφαρμογή της βαθμονομημένης παλινδρόμησης, είναι το εύρος ζώνης που θέτει τα όρια μίας περιοχής/γειτονιάς. Στην παρούσα εργασία, επιλέγουμε να είναι καθορισμένο από μία απόσταση, η οποία θα παραμένει σταθερή μεταξύ των τοπικών παλινδρομήσεων. Προκειμένου να επιλέξουμε το κατάλληλο εύρος ζώνης, χρησιμοποιούμε μια αυτοματοποιημένη διαδικασία η οποία προτείνει εκείνο το οποίο ελαχιστοποιεί την τιμή του κριτηρίου πληροφορίας AIC στο συγκεκριμένο μοντέλο. Σύμφωνα με αυτή, το βέλτιστο καθορισμένο εύρος ζώνης έχει την τιμή 2.711. Σημειώνεται, ωστόσο, ότι η τιμή αυτή προκύπτει με βάση το μοντέλο που δεν περιέχει την κατηγορική μεταβλητή για τον αριθμό των υπνοδωματίων. Για το λόγο αυτό, και προκειμένου να έχουμε μία πιο πλήρη εικόνα σχετικά με το πόσο μπορεί η επιλογή του εύρους ζώνης να επηρεάσει την προβλεπτική ικανότητα του μοντέλου, εξετάζουμε εναλλακτικές τιμές.

Πριν προχωρήσουμε στην παρουσίαση των αποτελεσμάτων που προκύπτουν από το βαθμονομημένο γεωγραφικό μοντέλο, εισάγουμε τα κριτήρια με τα οποία θα αξιολογηθούν οι δύο προσεγγίσεις. Προκειμένου να εντοπίσουμε ποιο από τα δύο έχει καλύτερη προβλεπτική ικανότητα, και άρα θα δώσει πιο ακριβείς προβλέψεις για την αξία των ακινήτων, χρησιμοποιείται η μέση τιμή του απολύτου σχετικού σφάλματος πρόβλεψης:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y_i}}{y_i}\right|$$

Με βάση αυτό, συμπεραίνουμε σε τι ποσοστό οι προβλέψεις που προέκυψαν από το συγκεκριμένο μοντέλο απέχουν από τις τιμές τις οποίες παρατηρήσαμε. Συνεπώς, μικρότερες τιμές του μέτρου υποδεικνύουν και πιο βελτιωμένη απόδοση του αντίστοιχου μοντέλου. Επιλέγουμε να χρησιμοποιήσουμε το παραπάνω εργαλείο καθώς τα μέτρα βάσει αναλογίας είναι πιο σχετικά και δίνουν καλύτερη πληροφορία για το σφάλμα πρόβλεψης, ενώ είναι και λιγότερο ευαίσθητα σε ακραίες τιμές.

Ένα άλλο μέτρο αξιολόγησης, αποτελεί το P20, το οποίο μετράει το ποσοστό των περιπτώσεων όπου η τιμή του MAPE είναι μικρότερη από 20%. Ορίζεται ως:

$$P20 = \frac{100}{N}\sum_{i=1}^{N} 1_{|PE_i|\leq 0.2}$$

με $PE = \frac{y_i - \hat{y}_i}{y_i}$, το ποσοστό σφάλματος, και $1_{|PE_i|\leq 0.2}$ δείκτρια συνάρτηση όπου:

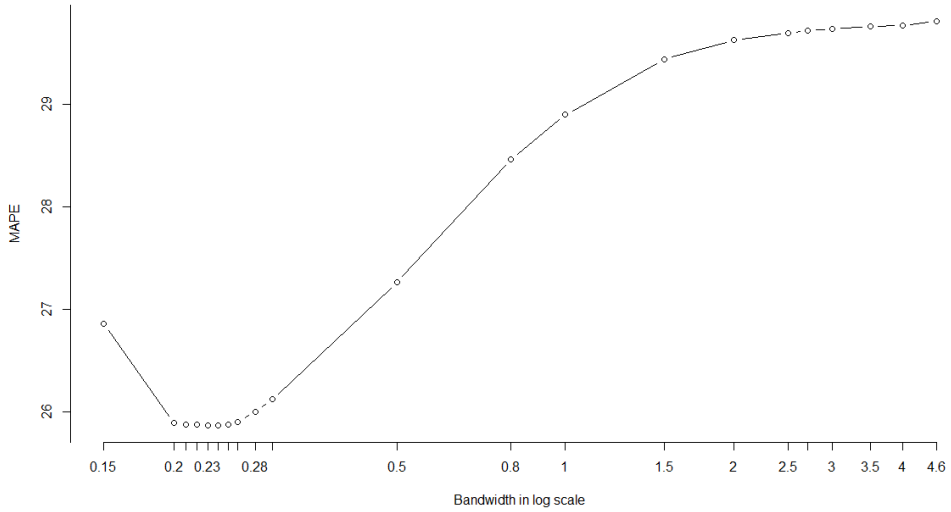
$$1_{|PE_i|\leq 0.2} = \begin{cases} 1, & |PE_i| \leq 0.2 \\ 0, & |PE_i| > 0.2 \end{cases}$$

Όπως είναι λογικό, μεγάλες τιμές του P20 είναι προτιμότερες.

Στο Διάγραμμα 6, παρουσιάζονται οι τιμές του παραπάνω μέτρου για το μοντέλο πολλαπλής παλινδρόμησης (bandwidth=4.6) καθώς και για το γεωγραφικό με βάρη, για διάφορες τιμές του εύρους ζώνης. Στο πρώτο μοντέλο, η τιμή του MAPE είναι 29.81%, ενώ ιδιαίτερο ενδιαφέρον φαίνεται να υπάρχει στις αντίστοιχες τιμές του δεύτερου για διαφορετικές επιλογές της περιοχής/ γειτονιάς. Ξεκινώντας ορίζοντας την απόσταση ως 0.15, έχουμε σημαντική βελτίωση συγκριτικά με το global μοντέλο, με το MAPE να λαμβάνει την τιμή 26.86%. Η τιμή αυτή μειώνεται καθώς αυξάνουμε σταδιακά το μέγεθος της ζώνης μέχρι το δεύτερο να πάρει τις τιμές 0.23 και 0.24, όπου και επιτυγχάνεται η ελάχιστη τιμή του μέτρου αξιολόγησης (25.87%). Στη συνέχεια, ενώ συνεχίζει να αυξάνεται η απόσταση, παρατηρούμε πως η τιμή του MAPE αυξάνεται, επίσης, και πλησιάζει βαθμιαία την αντίστοιχη του global μοντέλου. Σύμφωνα με τα παραπάνω, αντιλαμβανόμαστε πως το βέλτιστο, ως προς την προβλεπτική ικανότητα του μοντέλου, εύρος ζώνης είναι εκείνο που ορίζεται από την απόσταση 0.23, και όχι 2.71 όπως πρότεινε αρχικά η αυτοματοποιημένη διαδικασία. Όπως γίνεται εύκολα αντιληπτό, η επιλογή του κατάλληλου εύρους ζώνης παίζει καθοριστικό ρόλο στο βαθμό βελτίωσης του αντίστοιχου μοντέλου όταν το αξιολογούμε με βάση την προβλεπτική του ικανότητα. Ωστόσο, καταλήγουμε στο συμπέρασμα πως ανεξάρτητα από τα όρια της περιοχής που θα ορίσουμε, τα

γεωγραφικά μοντέλα με βαθμονόμηση συμπεριφέρονται, έστω και σε μικρό βαθμό, καλύτερα σε σύγκριση με το μοντέλο πολλαπλής γραμμικής παλινδρόμησης.

Στη συνέχεια, εστιάζουμε στις τιμές του μέτρου P20 για το μοντέλο πολλαπλής παλινδρόμησης και το αντίστοιχο γεωγραφικό με βαθμονόμηση και εύρος ζώνης 0.23. Τα αποτελέσματα έρχονται να συμφωνήσουν με τα συμπεράσματα που εξήγαμε παραπάνω, βάσει του MAPE, σχετικά με τη βελτιωμένη προβλεπτική ικανότητα που παρουσιάζει το γεωγραφικό μοντέλο. Η πρώτη παλινδρόμηση έχει P20 ίσο με 48.11%, ενώ η αντίστοιχη με σχέδιο βαθμονόμησης 53.39%. Αυτό σημαίνει, ουσιαστικά, πως οι περιπτώσεις στις οποίες το MAPE είναι μικρότερο του 20% στη δεύτερη προσέγγιση είναι αυξημένες κατά 5%.



Διάγραμμα 6. Τιμές του MAPE του global και του γεωγραφικού μοντέλου για διαφορετικά bandwidths

Εφαρμόζοντας το γεωγραφικό μοντέλο βαθμονόμησης, ορίζοντας την παράμετρο που αναφέρεται στο εύρος ζώνης ίση με 0.23, παίρνουμε τα αποτελέσματα που παρουσιάζονται στον Πίνακα 5 και στο Διάγραμμα 7.

Μοντέλο: log(price) ~ urban+sqm+type+maintenance+bedrooms2+heating					
Kernel function: gaussian					
Fixed bandwidth: 0.23					
Regression points: the same locations as observations are used.					
Distance metric: A distance matrix is specified for this model calibration.					
Συντελεστές					
	Ελάχιστη τιμή	**1ο τεταρτημόριο**	**Διάμεσος**	**3ο τεταρτημόριο**	**Μέγιστη τιμή**
Σταθερά	7.36034631	8.30187545	8.60457804	8.69448014	10.4181
urban<25k	-0.97214582	0.19433583	0.25123106	0.42349334	1.3100
urban25-50	-1.14849318	0.26647081	0.36272968	0.38919817	1.0323
urban50-100	-0.64845019	0.12113719	0.16722784	0.38252210	1.3912
urban100-150	-0.73815926	-0.02653786	0.18226093	0.41694892	1.4840
urban150-1000	-0.80271477	0.15865363	0.25975084	0.65298781	1.5382
urbancapital	-0.20626712	0.86930417	0.87941580	1.02589242	1.8937
sqm	0.00037752	0.00545704	0.00704518	0.00954118	0.0099
typeapartment	-0.02602026	0.34285009	0.42310562	0.43346776	0.6924
typeduplex	-0.06441811	0.44136074	0.45227882	0.48588202	0.9748
typeapartmentinvilla	-0.52752890	0.06260958	0.36595480	0.37782137	0.7947
maintenancemedium	-0.14320317	0.08282592	0.17452699	0.30305978	0.8083
maintenancegood	0.11286981	0.27930581	0.36502498	0.48848919	1.1717
maintenanceverygood	-0.40646276	0.58280073	0.60688124	0.85434600	1.7434
bedrooms21	-0.12870886	0.18187579	0.22246027	0.36265120	0.8040
bedrooms22	-0.00561831	0.29589221	0.36974414	0.58057067	0.8624
bedrooms23	0.03602609	0.37682447	0.47609596	0.68136312	1.1353
bedrooms24	0.02434300	0.24385479	0.40521848	0.67417126	1.2556
bedrooms25	-0.54941683	0.18997634	0.32763760	0.64299605	1.2430
bedrooms26	-0.20739579	0.00959487	0.20013469	0.58670662	1.3392
bedrooms27	-0.84659286	-0.75672895	0.08908433	0.51468295	1.1880
heatinglowshared	-0.09060138	0.15605771	0.18747457	0.20325976	0.3142
heatinglowauto	-0.27126025	0.01349015	0.03334978	0.12122777	1.1574
heatinghigh	-0.16106847	-0.02810128	-0.00506852	0.06981702	0.2195
AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 1510.707					
AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 1059.25					
R^2= 0.8456088		Προσαρμοσμένο R^2= 0.8202008			

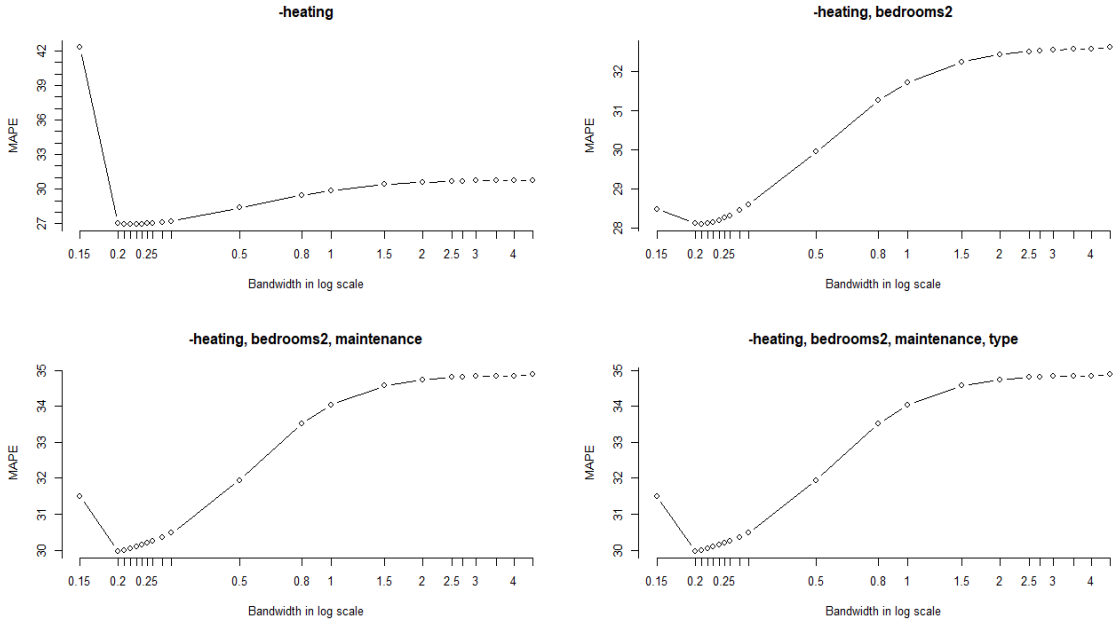
Πίνακας 5. Αποτελέσματα του γεωγραφικού μοντέλου με βαθμονόμηση

Από τον Πίνακα 5, εξάγουμε πληροφορίες τόσο για τις εκτιμήσεις που προκύπτουν για τις παραμέτρους του μοντέλου, καθώς και μέτρα σχετικά με την προσαρμοστικότητά του στα δεδομένα. Πιο συγκεκριμένα, έχουμε περιγραφικά μέτρα για τις παραμέτρους β του μοντέλου για κάθε χαρακτηριστικό, οι οποίες μεταβάλλονται σε κάθε παλινδρόμηση που εφαρμόζεται. Φαίνεται πως μεγαλύτερη διακύμανση παρουσιάζει η σταθερά, όπου σχετίζεται με τα επίπεδα αναφοράς των κατηγορικών μεταβλητών και ακίνητα με μηδενική έκταση (ελάχιστη τιμή=7.36034631, μέγιστη τιμή=10.4181).

Ωστόσο, είναι πολύ πιο εύκολο να αντιληφθεί κανείς και να εξετάσει πιο αναλυτικά τα παραπάνω αποτελέσματα μέσω γραφημάτων. Στο Διάγραμμα 7 παρουσιάζονται οι εκτιμήσεις των παραμέτρων β κάθε ανεξάρτητης (ψευδό)μεταβλητής που προκύπτουν από όλες τις γεωγραφικές παλινδρομήσεις. Κάθε χρώμα αντιπροσωπεύει τις τιμές του αντίστοιχου συντελεστή που εντάσσονται σε ένα τεταρτημόριο, με το ανοιχτόχρωμο να αναφέρεται στο πρώτο, το αμέσως πιο σκούρο στο δεύτερο κ.ο.κ. Για παράδειγμα, από το γράφημα της έκτασης του ακινήτου, παρατηρούμε πως στις ιδιοκτησίες που βρίσκονται στα κεντρικά σημεία, η αύξηση ενός τετραγωνικού μέτρου, με τα υπόλοιπα χαρακτηριστικά σταθερά, φαίνεται να επηρεάζει σε μεγαλύτερο βαθμό την λογαριθμική αξία του ακινήτου συγκριτικά με τις παρατηρήσεις που βρίσκονται στον περίγυρο. Αντίθετα, ο λογάριθμος της τιμής των κεντρικών ιδιοκτησιών που βρίσκονται σε αστική περιοχή 100-150 χιλιάδων κατοίκων, μεταβάλλεται λιγότερο από τις αντίστοιχες σε άλλες τοποθεσίες, σε σχέση με ακίνητα που βρίσκονται σε αγροτικές περιοχές και με τα υπόλοιπα γνωρίσματα σταθερά. Σε άλλες ψευδομεταβλητές, ωστόσο, όπως, για παράδειγμα, εκείνη που αναφέρεται σε τύπο θέρμανσης την αυτόνομη κεντρική θέρμανση, παρατηρούμε πως η μεταβολή στην τιμή είναι μικρή σε σχέση με τα επίπεδα αναφοράς, ανεξάρτητα από την τοποθεσία των ακινήτων. Σε γενικές γραμμές, φαίνεται πως το χωρικό μοτίβο που παρατηρείται κυρίως στα ακίνητα που βρίσκονται στα κεντρικά σημεία του δείγματος, αντικατοπτρίζεται στα διαφορετικά β που προκύπτουν με τη βοήθεια του πλάνου βαθμονόμησης. Ακόμη, επαληθεύεται η υπόθεση πως ιδιοκτησίες που βρίσκονται σε κοντινή απόσταση επηρεάζονται με τον ίδιο τρόπο από τα περισσότερα γνωρίσματα συγκριτικά με εκείνες που είναι απομακρυσμένες.

Διάγραμμα 7. Εκτιμήσεις των παραμέτρων β του γεωγραφικού μοντέλου παλινδρόμησης με βαθμονόμηση

Παρουσιάσαμε τα αποτελέσματα του μοντέλου γεωγραφικής παλινδρόμησης, το οποίο με την κατάλληλη επιλογή του εύρους ζώνης, βελτιώνει ικανοποιητικά τις προβλέψεις για την αξία των ιδιοκτησιών. Στη συνέχεια, έχει ενδιαφέρον να ελέγξουμε αν η συγκεκριμένη μέθοδος δίνει ακόμη καλύτερα αποτελέσματα όταν αφαιρούμε κάποια από τα χαρακτηριστικά που έχουν ρόλο ανεξάρτητης μεταβλητής, αλλά και αν υπάρχει γενικότερα βελτίωση συγκριτικά με το γενικό μοντέλο σε κάθε περίπτωση. Η διερεύνηση αυτή, γίνεται και πάλι για διαφορετικά όρια περιοχών, καθώς, όπως καταλήξαμε προηγουμένως, αποτελεί μία παράμετρο που επηρεάζει σε μεγάλο βαθμό την προβλεπτική ικανότητα. Κάθε φορά που ερευνούμε την παραπάνω υπόθεση, αφαιρούμε σταδιακά τη μεταβλητή η οποία προστέθηκε τελευταία στο μοντέλο. Τα αποτελέσματα παρουσιάζονται στο Διάγραμμα 8, όπου τα βαθμονομημένα μοντέλα αξιολογούνται με βάση το MAPE.



Διάγραμμα 8. Τιμές του MAPE του αντίστοιχου global μοντέλου και του γεωγραφικού για διαφορετικά bandwidths

Όπως παρατηρείται, τα μοντέλα όπου στο πρώτο έχουμε αφαιρέσει τη μεταβλητή για τον τύπο θέρμανσης και στο δεύτερο και τον αριθμό των υπνοδωματίων, έχουν παρόμοια συμπεριφορά. Πιο συγκεκριμένα, παρουσιάζουν καλύτερη προβλεπτική

ικανότητα από τη global πολλαπλή παλινδρόμηση που περιέχει όλες τις μεταβλητές, όταν επιλέξουμε να ορίσουμε την παράμετρο του εύρους ζώνης να έχει τιμές μικρότερες του 0.5. Σε περίπτωση που λάβει μεγαλύτερες τιμές, φαίνεται πως συμπεριφέρονται χειρότερα από το global, καθώς το MAPE αυξάνεται και αγγίζει ποσοστά των 30, 31 και 32%. Σημειώνεται, ακόμη, πως η ελάχιστη τιμή για το πρώτο μοντέλο επιτυγχάνεται όταν η απόσταση των ορίων κάθε περιοχής είναι 0.22, ενώ για το δεύτερο 0.21. Στη συνέχεια, εστιάζοντας στις δύο τελευταίες γεωγραφικές παλινδρομήσεις όπου έχουν αφαιρεθεί επιπλέον η κατάσταση συντήρησης και ο τύπος της περιοχής αντίστοιχα κάθε φορά, γίνεται εύκολα αντιληπτό πως δεν έχουν καλή απόδοση. Ειδικότερα, φαίνεται πως και τα δύο, ανεξάρτητα από το μέγεθος της γειτονιάς, εμφανίζουν τιμές για το MAPE που ξεπερνούν την αντίστοιχη του global με όλα τα γνωρίσματα. Μάλιστα, το μοντέλο που περιέχει, μόνο τις μεταβλητές για τον τύπο της περιοχής και την έκταση του ακινήτου, έχει τιμές από 35-41%. Αυτό σημαίνει, ουσιαστικά πως, κατά μέσο όρο, το 41% των προβλέψεων για την αξία των ιδιοκτησιών με βάση αυτό το μοντέλο είναι εκτός. Το τελευταίο αποτέλεσμα φαίνεται λογικό καθώς με την αφαίρεση αρκετών ανεξάρτητων χαρακτηριστικών από τα μοντέλα, παραλείπεται μεγάλο μέρος σημαντικής πληροφορίας που διαμορφώνει τις τιμές των ακινήτων. Συνεπώς, σε αυτή την περίπτωση, η αξιοποίηση της χωρικής πληροφορίας ενδέχεται να παραλειφθεί καθώς θα μπορούσε κανείς να αρκεστεί στην εφαρμογή μιας global γραμμικής παλινδρόμησης με όλες, ωστόσο, τις μεταβλητές. Τέλος, παρατηρώντας τις τιμές που παίρνει το MAPE στην αντίστοιχη global παλινδρόμηση των παραπάνω περιπτώσεων που εξετάζουμε, προκύπτει ότι η προσθήκη της τοποθεσίας σε συνδυασμό με ένα σχέδιο βαθμονόμησης, δίνει κάθε φορά καλύτερα αποτελέσματα. Αυτό συμβαίνει ανεξάρτητα από το εύρος ζώνης, το οποίο, ωστόσο, όταν οριστεί κατάλληλα θα βελτιώσει σε μεγαλύτερο βαθμό την ακρίβεια των προβλέψεων.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία παρουσιάσαμε δύο διαφορετικές προσεγγίσεις παλινδρόμησης για την πρόβλεψη της αξίας των ακινήτων. Στόχος ήταν, με τη σύγκριση αυτών βάσει κατάλληλων μέτρων, να τα αξιολογήσουμε ως προς την προβλεπτική τους ικανότητα. Πιο συγκεκριμένα, επικεντρωθήκαμε και εξετάσαμε αν με τη χρήση γεωγραφικής παλινδρόμησης με βαθμονόμηση, έχουμε καλύτερες προβλέψεις των τιμών, συγκριτικά με ένα μοντέλο που αγνοεί την τοποθεσία των ακινήτων.

Στην πρώτη μεθοδολογία της πολλαπλής γραμμικής παλινδρόμησης, όπου οι ανεξάρτητες μεταβλητές ήταν έξι, το μοντέλο προσάρμοζε σε ικανοποιητικό βαθμό τα δεδομένα. Αναφορικά με την προβλεπτική του ικανότητα, η μέση τιμή του απολύτου σχετικού σφάλματος πρόβλεψης υπολογίστηκε 29.81%, ενώ το μέτρο P20 στο 48.11%. Στη συνέχεια, με την ένταξη της χωρικής πληροφορίας μέσω ενός σχεδίου βαθμονόμησης των σημείων ανάλογα με την τοποθεσία τους και την εφαρμογή γεωγραφικών παλινδρομήσεων καταλήξαμε σε βελτιωμένα αποτελέσματα. Πιο συγκεκριμένα, είχαμε πιο ακριβείς προβλέψεις κατά 4% και τιμή του P20 αυξημένη κατά 5%. Ωστόσο, το αποτέλεσμα αυτό προέκυψε ύστερα από την επιλογή του κατάλληλου εύρους ζώνης, που όπως καταλήξαμε, αποτελεί παράμετρο η οποία επηρεάζει το βαθμό ακρίβειας στις προβλέψεις του γεωγραφικού μοντέλου. Παρ ’όλα αυτά, ανεξάρτητα από την απόσταση που θα οριοθετήσει μία περιοχή, το γεωγραφικό μοντέλο παρουσιάζει κάθε φορά βελτιωμένη απόδοση συγκριτικά με την πρώτη προσέγγιση. Το συμπέρασμα αυτό, έρχεται να συμφωνήσει με προηγούμενες έρευνες, καθώς η πολλαπλή γραμμική παλινδρόμηση αγνοεί πλήρως την τοποθεσία των ακινήτων, γεγονός που είναι μη ρεαλιστικό, ειδικά στον κτηματομεσιτικό κλάδο. Από την άλλη, μέσω της δεύτερης προσέγγισης, καταφέρνουμε να αξιοποιήσουμε τη χωρική συσχέτιση με τη σκέψη πως ακίνητα που είναι σε κοντινά σημεία τείνουν λογικά να έχουν σχετικές τιμές. Με αυτόν τον τρόπο, διορθώνοντας τα χαρακτηριστικά και επιτρέποντας να αλλάζει η επιρροή τους κάθε φορά ανάλογα με την τοποθεσία του ακινήτου, έχουμε πιο ακριβείς προβλέψεις για την τιμή τους.

Ακόμη, εξετάστηκε πώς επηρεάζεται η εκτίμηση της αξίας των ακινήτων με τη διαδοχική αφαίρεση ανεξάρτητων μεταβλητών. Ειδικότερα, καταλήξαμε πως με την παράλειψη μίας και δύο μεταβλητών αντίστοιχα και τον ορισμό του κατάλληλου εύρους ζώνης, έχουμε καλύτερα αποτελέσματα από την πολλαπλή γραμμική

παλινδρόμηση με όλες τις μεταβλητές που είχαμε εξαρχής. Στη δεύτερη περίπτωση, μάλιστα, όπου έλειπαν δύο από τα επεξηγηματικά χαρακτηριστικά, παρατηρούμε πως ανεξάρτητα από τα όρια της γειτονιάς, η τιμή του μέτρου MAPE είναι βελτιωμένη συγκριτικά με το αντίστοιχο global μοντέλο, έστω και σε μικρό βαθμό. Το ίδιο πόρισμα προκύπτει και από τις δύο τελευταίες περιπτώσεις όπου έχουμε αφαιρέσει τρεις και τέσσερις από τις αρχικές μεταβλητές αντίστοιχα. Ωστόσο, στα συγκεκριμένα μοντέλα, με τη χρήση χωρικής παλινδρόμησης, δεν έχουμε βελτίωση στις προβλέψεις συγκριτικά με το global που περιέχει όλες τις μεταβλητές. Φαντάζει λογικό, ωστόσο, καθώς παραλείπεται αρκετή πληροφορία που είναι χρήσιμη για την εκτίμηση της αξίας των ακινήτων. Υπενθυμίζεται, ωστόσο, πως όλα τα παραπάνω συμπεράσματα βασίζονται σε ένα μόνο training και testing δείγμα δεδομένων αντίστοιχα. Για το λόγο αυτό, παραμένουμε επιφυλακτικοί ότι η ανάλυση με διαφορετικά δείγματα θα έχει ακριβώς την ίδια κατάληξη.

Με βάση τα ευρήματα της παρούσας εργασίας, προκύπτουν ορισμένα θέματα για περαιτέρω διερεύνηση. Πιο συγκεκριμένα, όπως αποδείχθηκε, η επιλογή του εύρους ζώνης αποτελεί μία σημαντική παράμετρο, η οποία επηρεάζει την προβλεπτική ικανότητα του μοντέλου. Παρουσιάζει ενδιαφέρον, να εξετάσει κανείς τι συμβαίνει όταν γίνει χρήση ενός προσαρμοσμένου ορίου περιοχής (adaptive bandwidth) που θα μεταβάλλεται ανάλογα με τον αριθμό των ακινήτων που σχηματίζουν μία γειτονιά. Αναφορικά με τις αποστάσεις μεταξύ των σημείων, εκτός από την Ευκλείδεια, μπορεί να ερευνήσει κανείς το ίδιο πρόβλημα με τη βοήθεια πιο ρεαλιστικών προσεγγίσεων, εφόσον είναι διαθέσιμα, όπως δίκτυα μετακινήσεων ή περίπλοκων γεωγραφικών συνθηκών, που έχουν προταθεί στη βιβλιογραφία. Προτείνεται, επιπλέον, η σύγκριση των δύο μοντέλων ως προς την προβλεπτική τους ικανότητα βάσει κι άλλων μέτρων, όπως, για παράδειγμα το μέσο τετραγωνικό σφάλμα πρόβλεψης (MSPE), προκειμένου να υπάρχει μια πιο πλήρης και σαφής εικόνα του βέλτιστου. Ακόμη, ισχύει πως το βαθμονομημένο γεωγραφικό μοντέλο είναι πολύ ειδικευμένο για να επιτρέψει την χωρική επίδραση. Για το λόγο αυτό, παρουσιάζει ενδιαφέρον να εξεταστεί, σε μελλοντικές εργασίες, εάν ο συνδυασμός των δύο παραπάνω ή άλλων προσεγγίσεων επιφέρει ακριβέστερα αποτελέσματα. Τέλος, όπως χρησιμοποιείται στη βιβλιογραφία, μπορεί κανείς να ασχοληθεί με τη μοντελοποίηση της τιμής των ακινήτων ανά τετραγωνικό μέτρο έναντι της τιμής του σε κανονική ή λογαριθμική κλίμακα.

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

Ελληνική Βιβλιογραφία

Enikonomia (2020). Ακίνητα: Ελληνική πρωτιά στους φόρους - Η σύγκριση με χώρες της Ε.Ε. [online] Διαθέσιμο από: <http://www.enikonomia.gr/economy/236670,akinita-elliniki-protia-stous-forous-i-sygkrisi-me-chores-tis-ee.html> [10 Ιουλίου 2020]

Ξενόγλωσση Βιβλιογραφία

Alexandridis, A. K., Karlis, D., Papastamos, D. and Andritsos, D. (2019). Real Estate valuation and forecasting in non-homogeneous markets: A case study in Greece during the financial crisis, *Journal of the Operational Research Society*, 70(10), 1769-1783. https://doi.org/10.1080/01605682.2018.1468864

Anselin, L. (1995). Local indicators of spatial association—LISA, *Geographical analysis*, *27*(2), 93-115.

Antipov, E. A., and Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics, *Expert Systems with Applications*, *39*(2), 1772-1778.

Bitter, C., Mulligan, G. F. and Dall'erba, S. (2006). Incorporating spatial variation in housing attribute prices: A comparison of geographically weighted regression and the spatial expansion method, *Journal of Geographical Systems*, *9*(1), 7-27.

Brunsdon, C., Fotheringham, S. and Charlton, M. (1998). Geographically Weighted Regression-Modelling Spatial Non-Stationarity, *Journal of the Royal Statistical Society. Series D (The Statistician),* 47(3), 431-443.

Cassetti, E. (1972). Generating models by the expansion method: Applications to geographical research, *Geographical Analysis*, 4, 81-92.

Charlton, M., Fotheringham, S., and Brunsdon, C. (2009). Geographically weighted regression. *White paper. National Centre for Geocomputation. National University of Ireland Maynooth*.

Curry, B., Morgan, P. and Silver, M. (2002). Neural networks and non-linear statistical methods: an application to the modelling of price–quality relationships, *Computers & Operations Research*, 29(8), 951-969. https://doi.org/10.1016/S0305-0548(00)00096-4

Doumpos, M., Papastamos, D., Andritsos, D. and Zopounidis, C. (2020). Developing automated valuation models for estimating property values: a comparison of global and locally weighted approaches, *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03556-1

Dubin, R. A. (1992). Spatial autocorrelation and neighborhood quality, *Regional Science and Urban Economics*, 22(3), 433-452. https://doi.org/10.1016/0166-0462(92)90038-3

Gröbel, S., & Thomschke, L. (2018). Hedonic pricing and the spatial structure of housing data—an application to berlin, *Journal of Property Research*, 35(3), 185–208.

Helbich, M. and Griffith, D.A. (2016). Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches, Computers, Environment and Urban Systems, 57, 1-11.

Lu, B., Charlton, M., Harris, P. and Fotheringham, A. S. (2014). Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data, *International Journal of Geographical Information Science*. http://dx.doi.org/10.1080/13658816.2013.865739

McCluskey, W., McCord, M., Davis, P., Haran, M., and McIlhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches, *Journal of Property Research*, 30(4), 239–265. https://doi.org/10.1080/09599916.2013.781204

Mimis, A., Rovolis, A. and Stamou, M. (2013). Property valuation with artificial neural network: the case of Athens, *Journal of Property Research*, 30(2), 128-143. http://dx.doi.org/10.1080/09599916.2012.755558

Nguyen, N. and Cripps, Al. (2001). Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks, *Journal of Real Estate Research*, 22(3), 313–336.

Odland, J. (1988). Spatial Autocorrelation. Reprint. Edited by Grant Ian Thrall. WVU Research Repository, 2020.

Pace, R.K., Barry, R. and Sirmans, C. F. (1998). Spatial Statistics and Real Estate, *Journal of Real Estate Finance and Economics*, (17)1, 5-13.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T. and French, N. (2003). Real estate appraisal: a review of valuation methods, *Journal of Property Investment & Finance*, 21(4), 383-401.

Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, *Expert Systems with Applications*, 42(6), 2928-2934. https://doi.org/10.1016/j.eswa.2014.11.040

Pavlov, A. D. (2000). Space-varying regression coefficients: A semi-parametric approach applied to real estate markets. *Real Estate Economics*, *28*(2), 249-283. https://doi.org/10.1111/1540-6229.00801

Sulekan, A. and Jamaludin, S.S.S. (2020). Review on Geographically Weighted Regression (GWR) approach in spatial analysis, *Malaysian Journal of Fundamental and Applied Sciences*, 16(2), 173-177.

Szymanowski, M. and Kryza, M. (2012). Local regression models for spatial interpolation of urban heat island—an example from Wrocław, SW Poland, *Theoretical and Applied Climatology,* 108, 53–71. https://doi.org/10.1007/s00704-011-0517-6

Tiefelsdorf, M. and Boots, B. (1997). A Note on the Extremities of Local Moran's Iis and Their Impact on Global Moran's I, *Geographical Analysis*, 29(3), 249-257.