Universidade de Aveiro
2014

Departamento de Eletrónica, Telecomunicações e Informática

Luís Miguel Casal Carta

Sistema de Apoio à Decisão: Análise Espacial Aplicada ao Mercado da Habitação

Decision Support System: Spatial Analysis applied to Housing Market

Universidade de Aveiro
2014
Departamento de Eletrónica, Telecomunicações e Informática

Luís Miguel Casal Carta

Sistema de Apoio à Decisão: Análise Espacial Aplicada ao Mercado da Habitação

Decision Support System: Spatial Analysis applied to Housing Market

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Sistemas de Informação, realizada sob a orientação científica do Doutor José Manuel Matos Moreira, Professor Auxiliar do Departamento de Eletrónica, Telecomunicações e Informática e do Doutor João José Lourenço Marques, Professor Auxiliar do Departamento de Ciências Sociais, Políticas e Território.

Dedico este trabalho à minha família que sempre me apoiaram em tudo incondicionalmente, e em especial aos meus avós falecidos dos quais sinto muita falta.

o júri / the jury

presidente / presidente

Professora Doutora Ana Maria Perfeito Tomé

Professora Associada do Instituto de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro

vogais / examiners committee

Professor Doutor João Pedro Carvalho Leal Mendes Moreira

Professor Auxiliar da Faculdade de Engenharia da Universidade do Porto

Professor Doutor João José Lourenço Marques

Professor Auxiliar do Departamento de Ciências Sociais, Políticas e Território da Universidade de Aveiro

Agradecimentos / Acknowledgements

Gostaria desde já agradecer ao Professor Doutor José Moreira, ao Professor Doutor João Marques e ao Mestre Paulo Batista pela possibilidade que me foi concebida de realizar este trabalho, pela disponibilidade, motivação e apoio.

Gostaria de agradecer à minha família e à Margarida por todo o apoio, paciência e sorrisos que deram ao longo do meu percurso académico.

Agradeço a todos os meus amigos pelos incentivos, pelos conselhos e pela ajuda. Nunca deixaria de vos mencionar porque vocês todos são importantes para mim.

Um obrigado a todos!

palavras-chave *sistemas de apoio à decisão, análise de dados espaciais, modelos hedónicos, mercado habitação, preferências habitacionais,*

resumo

A presente dissertação apresenta uma nova abordagem aos modelos hedónicos e à questão da heterogeneidade espacial, no contexto do mercado habitacional. Os modelos de avaliação do preço da habitação recorrentemente utilizados, os modelos hedónicos, são considerados limitados; entre outras razões que são apontadas o fato de não considerarem convenientemente as particularidades do território (heterogeneidades e iteração). Neste contexto, as técnicas habitualmente usadas são regressões múltiplas (OLS). Existe um outro conjunto de métodos, por exemplo o GWR, que já considera a localização das habitações como fator de ponderação dos coeficientes de regressão, mas estes baseiam-se tradicionalmente em noções euclidianas e bidimensionais do espaço. A proposta apresentada neste trabalho passa pela utilização do método GWR, estendendo a noção de espaço a lógicas que ultrapassam a perspetiva física e geográfica do território. Esta nova abordagem aumenta a capacidade explicativa dos modelos hedónicos, permitindo a inclusão de métricas de vizinhança não exclusivamente geográficas. A aplicação desta metodologia e respetiva ferramenta ao mercado habitacional de Aveiro-Ílhavo permitiu obter melhores resultados quando comparado com os métodos tradicionais, como o OLS e o GWR.

keywords *decision support system, spatial data analysis, hedonic models, housing market, housing preferences,*

Abstract

In this work a novel approach to hedonic models and spatial heterogeneity was followed, as concerns the housing market. The models often employed for the evaluation of house prices, hedonic models, still exhibit several limitations. Moreover, their inability to take into account the territory particularities (heterogeneities and interactions) also represents a drawback. In this context, the method frequently employed is the multiple regression method (OLS). Also another approach can be followed, the geographic weighted regression, GWR, which considers the location of houses as a weighting factor on the regression coefficients, although in bi-dimensional Euclidean notions of space. The proposal presented in this work involves the use of GWR, extending the notion of space into a logical that goes beyond physical and geographical perspective of the territory. This new approach increases the explicative abilities of the hedonic models, allowing thus the inclusion of neighborhood metrics not solely geographic. This methodology and respective tool to housing market of Aveiro-Ílhavo yielded better results when comparing to conventional methods, such as OLS and GWR.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Símbolos

AIC	Critério de informação de Akkaike
API	Application Programming Interface
CBD	Distância ao Centro da Cidade
CSS	Cascading Style Sheets
CV	Cross-Validation
EPSG:3763	Código EPSG para o sistema global de Portugal Continental
EPSG:4326	Código EPSG para o sistema geodético mundial utilizado pelos GPS
GeoJSON	Geographic JavaScript Object Notation
GPS	Sistema de Posicionamento Global
GTWR	Regressão Geográfica e Temporal Ponderada
GWPR	Regressão Geográfica Ponderada em Painel
GWR	Regressão Geográfica Ponderada
h	Largura de banda
HTML	HyperText Markup Language
JSON	JavaScript Object Notation
OLS	Método dos Mínimos Quadrados
RSS	Soma dos quadrados dos resíduos
TOM	Tempo no Mercado
TSS	Soma dos quadrados totais
T-stat	Teste T de *student*
WLS	Mínimos Quadrados Ponderados
β	Coeficiente de Regressão
ε	Erro
R^2	Coeficiente de determinação

1. Introdução

Esta dissertação apresenta uma nova abordagem de análise do mercado habitacional construindo uma ferramenta de avaliação do valor patrimonial de uma habitação. Tendo como ponto de partida os modelos hedónicos, esta investigação, estende a aplicação de modelos GWR (Regressão Geográfica Ponderada) considerando o espaço territorial numa lógica não-euclidiana e multidimensional.

Ao longo dos anos, o interesse pela análise do mercado da habitação tem sido crescente em resultado da importância que a habitação assume nos dias de hoje, acentuado pelo contexto atual de recessão económica. Cresce a necessidade de saber quanto vale uma habitação e a necessidade de perceber a sua evolução temporal e variabilidade territorial. Uma habitação representa o mais importante valor para uma família, sendo ao mesmo tempo um fator de elevados encargos. O custo e a qualidade das habitações influência substancialmente a qualidade de vida das famílias. Por estas razões torna-se importante compreender as racionalidades que estão na base da formação dos preços de uma habitação, que por si só é um grande desafio mas que não se esgota no âmbito desta dissertação. O mercado habitacional é composto por habitações que possuem características únicas, ou seja, não existem duas habitações idênticas (Batista 2010, Marques 2012)

A abordagem mais comum para a avaliar e compreender o valor patrimonial de uma habitação é através de preferências reveladas (Marques 2012), recorrendo a modelos regressão (método dos mínimos quadrados – OLS). Estas técnicas permitem avaliar a forma como o valor de uma variável dependente (variável resposta - neste caso concreto, o preço de uma habitação) se relaciona com um conjunto de variáveis independentes (variáveis preditivas - neste caso concreto, as caraterísticas habitacionais). Uma análise de regressão é normalmente uma combinação linear de variáveis independentes, onde os coeficientes são considerados estacionários sobre o espaço e produzem um modelo global (Demšar, Fotheringham et al. 2008). Por esta razão, os modelos OLS não têm em consideração os dados espaciais (apenas atributos locacionais que se possam incluir no modelo para caracterizar uma habitação). Assim, outras técnicas, ou mesmo tentativas, foram ao longo do tempo desenvolvidas, considerando a questão de que os dados espaciais não são estacionários:

- Casetti (1972) introduziu a expansão dos parâmetros;
- Foster e Gorr (1986) introduziu a filtragem adaptada espacial;
- Anselin (1988) introduziu o *spatial lag* e *spatial error*, um exemplo é o caso de estudo da criminalidade (Figura 1);
- McMillen (1996) introduziu a regressão localmente ponderada;
- Brunsdon, Fotheringham et al. (1996) introduziram a regressão geográfica ponderada.

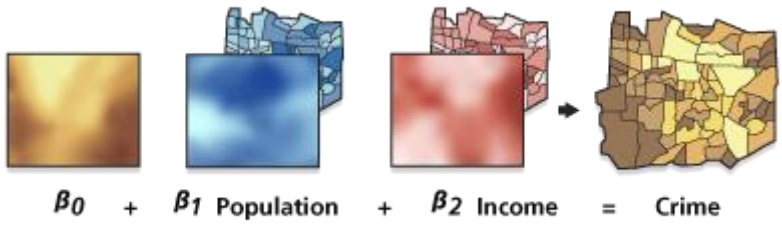


Figura 1 – Modelo de avaliação da criminalidade[1].

A regressão geográfica ponderada, ou GWR, introduzida por Brunsdon, Fotheringham et al. (1996) desenvolve uma nova abordagem. Ao contrário do método OLS, o GWR estima os coeficientes de regressão, calibrando-os pela localização geográfica das observações, podendo esta variar espacialmente (Brunsdon, Fotheringham et al. 1999). Desde então, novas extensões do método GWR foram introduzidas ao longo dos anos: Huang, Wu et al. (2010) introduziu o método GTWR, regressão geográfica e temporal ponderada, que resulta de uma análise espacial e temporal dos dados; e Yu (2010) introduziu o método GWPR, regressão geográfica ponderada em painel. Contudo, um dos grandes desafios da análise espacial é a forma como se medem as distâncias e forma como se definem vizinhanças. O mais comum é utilizarem-se distâncias Euclidianas e noções exclusivamente geográficas para determinar níveis de vizinhança.

[1] Fonte: http://resources.arcgis.com/en/help/main/10.1/index.html#//005p00000021000000

A ambição deste estudo é assim dar mais um pequeno no que toca à expansão do método tradicional GWR, sendo desta forma possível obterem-se melhores resultados e consequentemente avaliações de preços habitacionais mais assertivas. Acresce ainda o facto de ser possível aferir quais os submercados habitacionais (oferta) que vão de encontro às expectativas e preferências dos potenciais compradores (procura).

1.1 Motivação

Os modelos hedónicos foram introduzidos por Rosen (1974) em que os preços das habitações são influenciados por várias características/ atributos, tais como: i) físicos, ii) localização e iii) temporais (Crespo R. 2007, Huang, Wu et al. 2010). Pelo fato da localização ser um elemento fundamental da caracterização do mercado habitacional, confere a este objeto de estudo (a habitação) uma enorme complexidade, não só na determinação de quais os atributos relevantes a incluir nos modelos, como também, a forma como são modelados (por exemplo, considerando a heterogeneidade e iteração espacial) (Marques 2012).

Os modelos de regressão podem ser agrupados em duas grandes categorias: i) modelos globais e ii) modelos locais. Os modelos globais são habitualmente uma forma improvisada do modelo hedónico tradicional (Can 1992, Dubin 1992, Anselin 1998). Estes modelos assumem que os parâmetros estimados são constantes ao longo do espaço (Brunsdon, Fotheringham et al. 1996, Huang, Wu et al. 2010). Para um modelo específico, como o preço de uma habitação, a suposição da estabilidade estacionária ou estrutural ao longo do tempo e do espaço é geralmente irrealista, já que os parâmetros tendem a ser diferentes na área de estudo (Huang, Wu et al. 2010). Ao longo dos anos, várias metodologias foram desenvolvidas de modo a considerar esta variação dos parâmetros no espaço. Assim, surgiu a metodologia da variação da regressão de modelos locais (*local regression model*) a regressão geográfica ponderada (*geographically weighted regression*) (Brunsdon, Fotheringham et al. 1996, Fotheringham, Charlton et al. 1996, Fotheringham, Brunsdon et al. 2003).

A metodologia de regressão geográfica ponderada alarga o quadro de referência teórico da regressão tradicional excluindo a premissa da estacionariedade e considera os parâmetros dos modelos funções da localização (Demšar, Fotheringham et al. 2008).

Sumariamente, a regressão geográfica ponderada tem tido bastante atenção nos últimos anos (Huang, Wu et al. 2010) tendo sido intensamente aplicada à avaliação do mercado imobiliário (Pavlov 2000, Fotheringham, Brunsdon et al. 2003, Yu 2006).

As metodologias para estudar o mercado imobiliário têm tido grandes desenvolvimentos no sentido de determinar o ajuste das preferências de quem procura casa, ou a nível profissional, tal como empresas, por exemplo bancos, no qual pretendem saber qual o real valor de uma habitação, no sentido de reavaliar hipotecas ou mesmo para a concessão de novos créditos. Acredita-se que esta é a melhor altura para o desenvolvimento de uma ferramenta que ofereça tais possibilidades, isto é, o de tornar mais compreensível o modo como se formam os preços de uma habitação.

Existem algumas ferramentas que aplicam os métodos de regressão até agora referidos. O ArcGIS[2] é uma ferramenta de mapeamento, análise e gestão de dados geográficos. Desenvolvida pela ESRI, está disponível em várias plataformas – desktop, web, *smartphone* – e possui funções para a análise estatística espacial, onde estão presentes os métodos OLS e GWR. R[3] é uma linguagem e um sistema de computação estatística multiplataforma. É livre e *open-source*, no qual é possível utilizar e modificar o código fonte. Esta ferramenta tem disponível um conjunto de pacotes (funções) *online*, onde é possível aceder ao pacote de funções para o método OLS e GWR, e permite o desenvolvimento de funções que, podem ser disponibilizadas a outros utilizadores. GWR4[4] é uma ferramenta desenvolvida com o propósito de calibrar modelos GWR, sejam gaussianos, de *Poisson* ou logísticos, e apenas pode ser executado em sistemas Windows. A ferramenta Matlab[5] disponibiliza uma função que aplica o método OLS e, adicionalmente, encontra-se disponível um pacote de funções econométricas que aplicam o método GWR.

As ferramentas indicadas cobrem grande parte das funcionalidades dos métodos OLS e GWR, porém estas apresentam lacunas quanto à flexibilização e custo. As ferramentas ArcGIS e GWR4 são proprietárias, não sendo possível aceder e modificar o

[2] https://www.arcgis.com/features/

[3] http://www.r-project.org/

[4] https://geodacenter.asu.edu/gwr_software

[5] http://www.mathworks.com/products/matlab/

código fonte. Relativamente aos custos, as ferramentas ArcGIS e Matlab tornam-se dispendiosas. No Matlab é possível alterar o código relativo ao método GWR, no entanto, a aquisição deste revela custos elevados. R é uma ferramenta que possui a sua própria linguagem e ambiente de desenvolvimento. Este fator leva a que a curva de aprendizagem seja mais acentuada.

Assim, por um lado, a relevância do território para a compreensão dos mercados habitacionais e a complexidade que daí resulta; e por outro a pouca flexibilidade dos instrumentos existentes para inclusão desta dimensão espacial, conferem a esta investigação um papel relevante, no avanço do conhecimento aplicado a uma realidade concreta – o mercado da habitação.

1.2 Objetivos

O objetivo principal desta dissertação é desenvolver uma ferramenta que determine o valor de uma habitação, diferenciado por vários submercados, utilizando uma versão modificada da metodologia de regressão geográfica ponderada. Os pontos focados são os apresentados a seguir:

1. Analisar o método GWR e verificar como pode ser estendido.
2. Desenvolver uma ferramenta que aplique o método OLS e a versão estendida do GWR e a estudos na área do mercado da habitação.
3. Desenvolver uma aplicação web que apresente num mapa os resultados obtidos a partir das metodologias apresentadas no ponto 2.

Esta dissertação compreende duas componentes, uma analítica e outra visual. Na primeira, pretende-se desenvolver uma biblioteca que ofereça uma forma de determinar um modelo hedónico utilizando uma versão estendida da metodologia GWR. Além de ser aplicado o método GWR tradicional com a distância Euclidiana foram aplicadas funções de distância não Euclidianas e uma função de similaridade (que traduz noções multidimensionais de espaço), aplicando ponderações aos vizinhos. Na componente visual, procura-se desenvolver uma aplicação web que demonstre a utilização da biblioteca referida, utilizando os coeficientes de regressão do modelo GWR, e sejam apresentados, num mapa, as zonas com as habitações de potencial compra e/ou as mais rentáveis. As

pesquisas permitem selecionar um conjunto de atributos habitacionais, estabelecer um preço limite e determinar habitações no mercado com determinados atributos.

1.3 Estrutura do Documento

Este documento, para além desta introdução, encontra-se dividido em outras seis secções.

Na secção 2 é descrito o estado de arte das metodologias para o cálculo dos modelos de avaliação do preço de uma habitação, incluindo uma descrição dos modelos hedónicos e da importância da heterogeneidade espacial. Na secção 3 são apresentadas ferramentas de avaliação do preço de uma habitação, onde é descrito um caso de estudo e é realizada uma síntese das metodologias analisadas, expondo as alternativas e o foco desta dissertação. Na secção 4 são apresentados os requisitos e a arquitetura do sistema global, acompanhados por diagramas que esclarecem as funcionalidades e o funcionamento da solução final. Na secção 5 é explicado o processo de implementação das ferramentas a desenvolver, indicando as linguagens e APIs utilizadas. Por fim, nas secções 6 e 7 é dado foco aos resultados obtidos e conclusões.

2. Modelos de Avaliação do Preço da Habitação

Nesta secção são apresentadas algumas definições no contexto dos mercados habitacionais, nomeadamente os modelos hedónicos e a heterogeneidade espacial. São apresentadas as metodologias estudadas, tanto a nível de investigação como na área de sistemas de informação.

2.1 Modelos Hedónicos e a Heterogeneidade Espacial

O termo hedónico, ou mesmo modelo de preços hedónicos, é muito utilizado e conhecido no contexto do mercado imobiliário (Goodman 1978, Cheshire e Sheppard 1995, Meen, Andrew et al. 1998). Este tipo de modelo é utilizado para determinar ou explicar os preços das habitações a partir das suas caraterísticas (Rosen 1974, Marques, Castro et al. 2012). Um modelo de preços hedónicos é tipicamente representado da seguinte na forma:

$$\boldsymbol{p = Hv + \varepsilon} \qquad \text{Equação 1}$$

Nesta equação, p é um vetor de m preços de habitações (particularmente logaritmizado); H é uma matriz que contém n atributos para m habitações, relacionado com características intrínsecas e à localização; v é o vetor de preços hedónicos, refletindo a avaliação de m habitações; e ε representa o erro.

A heterogeneidade espacial possui elevada importância no mercado da habitação, que é caracterizado por ser segmentado e estruturado por um padrão complexo e inter-relacionado de elementos, em vez de ser determinado por um único e homogéneo processo de organização espacial. No contexto da habitação, a heterogeneidade espacial ocorre quando existe uma segmentação territorial no mercado da habitação e, ou os preços hedónicos associados aos diferentes atributos, ou as características das habitações, não são constantes ao longo do território (Marques, Castro et al. 2012). E desta forma, se quisermos incluir esta perspetiva nos típicos modelos hedónicos a equação fica:

$$\boldsymbol{p = d + Hv + \varepsilon}$$ Equação 2

Onde d corresponde à *dummy* de cada submercado. Independentemente de considerar ou não a heterogeneidade espacial nos modelos hedónicos, estes podem ser determinados de várias formas. Nas secções seguintes apresenta-se essa diversidade metodológica que pode ser na sua forma mais simples, recorrendo a modelos OLS, ou evoluírem para modelos econométricos mais complexos.

2.2 Método dos Mínimos Quadrados (OLS)

O método dos mínimos quadrados (ou, *Ordinary Least Squares*) é uma técnica de regressão em que se pretende estimar valores desconhecidos de uma ou mais variáveis independentes e minimizar a soma dos quadrados de n observações, dando origem a uma equação matemática. O método OLS pode também ser linear ou não linear. Enquanto a regressão linear pretende estimar valores de uma variável dependente a partir de uma ou várias variáveis independentes, a regressão não-linear estima esses valores através de um processo iterativo, onde sucessivas aproximações são efetuadas até o erro ser minimizado.

O modelo da regressão linear pode tomar duas formas, simples ou múltipla. A regressão linear simples é uma equação em que se pretende estimar dois parâmetros β, uma constante e uma variável independente. A forma deste tipo de regressão linear é dada através da seguinte expressão:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \; para \; i = 1, \dots, n}$$ Equação 3

Por sua vez, é possível aplicar a mesma expressão da regressão linear simples para obter valores estimados para várias variáveis independentes. Assim, de um modo geral, a regressão linear múltipla tem a seguinte expressão:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} + \varepsilon_i, para \; i = 1, \dots, n}$$ Equação 4

Nas duas equações verifica-se que y_i é a variável dependente para uma determinada observação i, x_i é a variável independente, ε_i é o erro e os β são parâmetros a ser estimados que minimizam a à soma dos mínimos quadrados do conjunto de n observações. No contexto do mercado habitacional, os valores estimados (β) representam as ponderações ou pesos que cada característica da habitação tem no preço final da casa. No caso particular da habitação, o valor do coeficiente de regressão, por exemplo da área da habitação, mede o impacto no preço quando aumenta uma unidade (m^2) de área; ou no caso do número de quartos, mede o impacto no preço devido ao facto da habitação ter mais um quarto. A soma dos quadrados dos resíduos (ou RSS, *Residual Sum of Squares*) é dada pela seguinte expressão:

$$\sum_{i=1}^{n}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\boldsymbol{2}} \qquad \text{Equação 5}$$

Verifica-se que na equação 5 que $\hat{y}_i$ é o valor previsto para a i-ésima observação, dado o valor i-ésimo de x. O termo $(y_i - \hat{y}_i)$ é o resíduo para a i-ésima observação. Os resíduos devem ser independentes e seguem a Distribuição Normal Gaussiana com média de zero e variância constante, ou seja, $error \sim N(0, \sigma^2)$ (Mathworks 2013). A expressão da soma dos quadrados dos resíduos é o objetivo deste método, já que se pretende que o valor deste somatório seja mínimo para que haja um melhor ajuste da linha de regressão.

O estimador dos coeficientes para o Método dos Mínimos Quadrados é apresentado na forma matricial, tendo a expressão seguinte:

$$\boldsymbol{\hat{\beta}} = (\boldsymbol{X^T X})^{\boldsymbol{-1}}\boldsymbol{X^T y} \qquad \text{Equação 6}$$

Na equação 6, $\hat{\beta}$ é o vetor dos parâmetros estimados, X é a matriz dos valores das variáveis independentes com uma coluna adicional de uns que representa a constante, y é o vetor dos valores observados e a expressão $(X^T X)^{-1}$ é a inversa da matriz variância-covariância.

Os coeficientes determinados na equação 5 necessitam de ser avaliados de modo a ser verificado se o seu ajuste é ótimo. O coeficiente de determinação, R^2, explica quanto da

proporção de variação de Y é determinado pelo modelo X (Bennoit 2010). Este coeficiente é determinado pelo rácio entre a soma dos quadrados dos resíduos (RSS) e a soma total dos quadrados (TSS), no qual resulta a expressão seguinte:

$$\boldsymbol{R^2 = 1 - \frac{RSS}{TSS} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y_i})^2}{\sum_{i=1}^{N}(y_i - \overline{y_i})^2}} \qquad \text{Equação 7}$$

O valor do R^2 é medido entre os valores 0 e 1. Quando mais perto este valor estiver de 1, melhor o Y é explicado pelo modelo. Para uma habitação, o valor do coeficiente de determinação é importante, porque permite verificar se o conjunto de características habitacionais é explicativo do preço final da habitação.

2.3 Mínimos Quadrados Ponderados (WLS)

O método OLS, apresentado na secção 2.2, estima valores para as variáveis independentes através de uma regressão linear ou não linear, tanto simples como múltipla, dando origem a um vetor dos parâmetros estimados $\hat{\beta}$. Caso seja necessário, é possível atribuir pesos às observações de uma regressão para melhorar o ajuste de modo a determinar a forma como cada valor de resposta influencia a estimativa dos parâmetros finais (Mathworks 2013). Tendo em conta a equação 4 da soma dos mínimos quadrados, o método WLS minimiza os erros através da equação seguinte:

$$\boldsymbol{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2} \qquad \text{Equação 8}$$

Na equação 8, w_i representa os pesos. São aplicados pesos a cada diferença de quadrados antes de minimizar para que a imprecisão de algumas estimativas sofram maior correção que outras. Adicionando uma matriz W, uma matriz quadrada onde os pesos se encontram na diagonal da mesma, resulta na seguinte expressão:

$$\boldsymbol{\hat{\beta} = (X^T W X)^{-1} X^T W y} \qquad \text{Equação 9}$$

Verifica-se na equação 9 que os valores estimados sofrem uma modificação. Tal como para o método OLS, o método WLS, aplicado ao mercado habitacional, pode influenciar o preço final da habitação atribuindo pesos superiores a características que se considerem pertinentes.

2.4 Análise Geoestatística

2.4.1 Regressão Geográfica Ponderada (GWR)

Regressão geográfica ponderada (ou *Geographically Weighted Regression*) é uma técnica simples que estende o método tradicional da regressão, permitindo que os parâmetros sejam estimados localmente, para que os coeficientes do modelo, em vez de serem estimativas globais, sejam específicos para um ponto i (localização) (Brunsdon, Fotheringham et al. 1996). A principal ideia desta metodologia passa por estimar estes parâmetros dado uma variável dependente y e um conjunto de uma ou mais variáveis independentes x, medidas em locais específicos (Charlton, Fotheringham et al. 2009). Nesta fase, verifica-se que uma dada localização u influencia no cálculo da regressão pelo método dos mínimos quadrados. Assim, na equação do OLS com regressão múltipla são atribuídas localizações u a cada uma das variáveis dando origem à seguinte equação:

$$\boldsymbol{y_i(u) = \beta_{0i}(u) + \sum_m \beta_{mi}(u) x_{mi}} \qquad \text{Equação 10}$$

Na equação 10, $\beta_{0i}(u)$ representa o valor de interceção e $\beta_{mi}(u)$ representa o conjunto de parâmetros para a localização i. A variável m varia conforme o número de variáveis independentes que se pretende passar para a regressão do GWR.

As localizações de cada um dos pontos são medidos conforme a sua proximidade e similaridade, isto é, observações que se encontrem mais perto de uma determinada localização tenham maior influência que pontos mais afastados (Charlton, Fotheringham et al. 2009), e o modelo é calibrado utilizando o método dos mínimos quadrados ponderados (WLS). Além disso, os pesos variam conforme a localização das observações, existindo uma diferente calibração para cada observação. O estimador do GWR aplica, então, o método WLS com pesos nas localizações e resulta na seguinte equação:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{u}) = (\boldsymbol{X^T W(u) X})^{-1} \boldsymbol{X^T W(u) y} \quad \text{Equação 11}$$

Na equação 11, $W(u)$ é a matriz de pesos relativa à localização u, a expressão $X^T W(u) X$ é a matriz de variância-covariância geográfica com pesos e y é o vetor dos valores da variável dependente.

Os pesos relativos à matriz $W(u)$ são calculados de maneira diferente em relação ao método OLS, conhecidos também por *kernel*[6]. De acordo com Charlton, Fotheringham et al. (2009) existem dois tipos de esquemas de ponderação: *kernel* fixo e *kernel* adaptado. Para o *kernel* fixo, assume-se que a distância é constante ao longo do espaço enquanto o número de vizinhos próximos varia. Relativamente ao *kernel* adaptado, a situação é contrária ao *kernel* fixo. Tipicamente, é utilizado um *kernel* na forma Gaussiana e este é calculado pela seguinte expressão:

$$\boldsymbol{w_{ij}(u)} = \boldsymbol{e}^{-0.5\left(\frac{d_{ij}(u)}{h}\right)^2} \quad \text{Equação 12}$$

Na equação 12, $w_{ij}(u)$ é o peso geográfico da i-ésima observação relativamente à localização u, $d_{ij}(u)$ é a distância medida entre a observação i e a observação j com a localização u e h é o parâmetro não negativo que representa a largura de banda (Huang, Wu et al. 2010). As distâncias podem ser calculadas com qualquer métrica, no entanto, é geralmente utilizada a distância Euclidiana com coordenadas cartesianas, sendo que, as coordenadas do ponto i são (x_i, y_i) e do ponto j são (x_j, y_j).

$$\boldsymbol{d_{ij}} = \sqrt{\left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)^2 + \left(\boldsymbol{y_i} - \boldsymbol{y_j}\right)^2} \quad \text{Equação 13}$$

[6] No contexto do GWR, *kernel* é a forma de determinar os pesos na matriz W a partir de um esquema de ponderação, por exemplo, a expressão apresentada na equação 12.

Tal como referido anteriormente podem ser adaptadas outras métricas no cálculo da distância entre dois pontos, contudo esta pode depender no tipo de coordenadas que são passadas, por exemplo, se as coordenadas forem esféricas (tridimensionais) é utilizada a distância ortodrómica (Charlton, Fotheringham et al. 2009), ou até não Euclidianas.

Voltando à equação 11, a largura de banda h é expressa no mesmo sistema de coordenadas utilizado para o cálculo da distância (Charlton, Fotheringham et al. 2009). A largura de banda é um parâmetro importante no cálculo dos pesos na matriz de ponderação W, devido ao espaçamento dos dados. A regressão geográfica ponderada pode utilizar uma largura de banda fixa ou adaptada conforme esse espaçamento dos dados. Assim, para dados mais dispersos, é apropriado utilizar uma largura de banda fixa, enquanto, para dados mais densos, é conveniente que o *kernel* escolha uma largura de banda adaptado. A cada iteração das n observações, para uma observação específica da localização u, são ordenadas as distâncias em relação a essa observação e a largura de banda é atribuída. Às observações cuja distância calculada seja superior à largura de banda é atribuído peso zero. Uma adaptação *kernel* que pode ser utilizada é a ponderação *bisquare*:

$$\boldsymbol{w_{ij}} = \begin{cases} \left(\mathbf{1} - \left(\frac{\boldsymbol{d_{ij}(u)}}{\boldsymbol{h_i}}\right)^{\mathbf{2}}\right)^{\mathbf{2}}, \boldsymbol{para\ d_{ij}(u)} < h \\ \mathbf{0}\ , \boldsymbol{para\ d_{ij}(u)} > h \end{cases} \qquad \text{Equação 14}$$

O processo de calibração para a escolha da largura de banda adequada é realizado através de *Cross-Validation*, isto, para que o modelo seja calibrado com as observações perto da localização i e não dessa mesma localização i (Brunsdon, Fotheringham et al. 1996, Brunsdon, Fotheringham et al. 1998, Huang, Wu et al. 2010). Se o valor previsto y_i em função de h é $\hat{y}_i(h)$, então a soma dos quadrados dos erros é dada pela seguinte expressão:

$$\sum_{\boldsymbol{i}}\left(\boldsymbol{y_i} - \boldsymbol{\hat{y}_{\neq i}(h)}\right)^{\mathbf{2}} \qquad \text{Equação 15}$$

Na equação 15, $\hat{y}_{\neq i}(h)$ é o valor ajustado de y_i com as observações da localização i omitidas do processo de calibração (Brunsdon, Fotheringham et al. 1996). Considerar a

expressão da equação 13 ao parâmetro h pode proporcionar uma orientação para a escolha de um valor apropriado do parâmetro (Brunsdon, Fotheringham et al. 1996, Huang, Wu et al. 2010) e, para testar a sua avaliação de ajuste (*goodness of fit*), utiliza-se o critério de informação de *Akaike* (AIC) (Charlton, Fotheringham et al. 2009, Huang, Wu et al. 2010).

2.4.2 Regressão Geográfica e Temporal Ponderada (GTWR)

B. Huang et al. (2010) introduziu uma nova metodologia que incorpora a informação temporal das habitações, capturando a heterogeneidade espacial e temporal (Huang, Wu et al. 2010). Convencionalmente, a variável tempo é acomodada separadamente ajustando o preço de vendas das observações para uma data em comum (Wang 2006). No entanto, para incorporar o fator tempo das habitações, tem-se em conta a matriz ponderada com base nas distâncias das coordenadas x, y e t entre a observação i com as restantes observações. A equação do modelo GTWR é expressa:

$$\boldsymbol{Y_i = \beta_0(x_i, y_i, t_i) + \sum_k \beta_k(x_i, y_i, t_i) X_{ik} + \varepsilon_i} \qquad \text{Equação 16}$$

Por sua vez, a estimação dos valores para $\beta_k(x_i, y_i, t_i)$ é dada pela expressão:

$$\boldsymbol{\widehat{\beta}(x_i, y_i, t_i) = [X^T W(x_i, y_i, t_i)]^{-1} X^T W(x_i, y_i, t_i) Y} \qquad \text{Equação 17}$$

Na equação 17, $W(x_i, y_i, t_i) = diag(\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{in})$ e n é o número de observações. Os elementos da diagonal $\alpha_{ij} (1 \leq j \leq n)$ são as distâncias espaciais e temporais da função (x, y, t), correspondendo aos pesos quando se calibra uma regressão ponderada adjacente à observação i. A função para calcular as distâncias, incorporando o fator tempo, é dada pela seguinte equação:

$$\boldsymbol{d_{ij} = \sqrt{\lambda \left[(x_i - x_j)^2 + (y_i - y_j)^2 \right] + \mu (t_i - t_j)^2}} \qquad \text{Equação 18}$$

Na equação 18, λ e μ são fatores de escala para balancear os diferentes efeitos utilizados para medir a distância espacial e temporal nos seus respetivos sistemas métricos. Por sua vez, a equação anterior pode ser melhorada de forma a contornar o problema da degradação do cálculo das distâncias quando λ ou μ são zero. Assim, introduzindo τ para determinar o rácio μ/λ, com $\lambda \neq 0$, obtém-se a equação:

$$\frac{(d_{ij})^2}{\lambda} = \left[(x_i - x_j)^2 + (y_i - y_j)^2\right] + \tau(t_i - t_j)^2 \qquad \text{Equação 19}$$

$W(x_i, y_i, t_i)$ multiplicado pela constante da equação 17 não altera a estimação $\beta_k(x_i, y_i, t_i)$, pode-se observar que o parâmetro $\tau = \mu/\lambda$ executa um papel importante no cálculo dos pesos (Huang, Wu et al. 2010).

2.4.3 Regressão Geográfica Ponderada em Painel (GWPR)

A análise de dados em painel tem tido elevada importância na área da análise econométrica devido às suas vantagens sobre as técnicas de análise de dados de cortes transversais ou de séries temporais convencionais (Hsiao 2003, Baltagi 2008, Yu 2010).

A ideia central da análise da regressão geográfica ponderada em painel é relativamente similar ao corte transversal da análise da regressão geográfica ponderada (Yu 2010). Na metodologia da regressão geográfica ponderada em painel, é assumido que a série temporal das observações numa dada localização é uma realização de um processo espaciotemporal suave (Yu 2010), isto é, considera os efeitos de vizinhança da mesma forma que em dados *cross-section*. Neste caso os dados são organizados em painel, isto é ano a ano. Este processo espaciotemporal segue uma distribuição em que as observações vizinhas (tanto geograficamente ou temporalmente) estão mais relacionadas do que as observações mais distantes (Yu 2010).

A regressão geográfica ponderada em painel pode ser vista como uma versão expandida da análise do corte transversal da regressão geográfica ponderada. Assim, a função *kernel* e a largura de banda são utilizados da mesma forma que na regressão geográfica ponderada e determinam o tamanho do subconjunto em torno de uma localização geográfica particular e atribuem-se pesos aos dados existentes (Yu 2010).

Um dos pontos-chave em aplicar uma regressão localmente ponderada em painel é o tamanho do conjunto local, por terminologia da regressão geográfica ponderada, a largura de banda da função do *kernel* fixo ou o mais próximo dos vizinhos da função do *kernel* adaptado. Assim, são aplicados dois critérios (Yu 2010):

1. Classificação por *Cross-Validation* (CV);
2. Critério de Informação de Akaike (AIC).

Sumariamente, a metodologia da regressão geográfica ponderada em painel aplica os seguintes passos para estimar as variáveis (Bruna e Yu 2013):

1. Selecionar h_i, sendo h_i a largura de banda para a observação i;
2. Subdividir os dados em níveis para estimação dos locais i;
3. Ponderar as observações temporais das variáveis j em níveis para estimação dos locais i;
4. Aplicar a estimação dos dados em painel para os dados ponderados e subdivididos (Croissant e Millo 2008);
5. Mapear coeficientes significantes (Mennis 2006).

2.5 Síntese

A componente analítica, isto é, a determinação de um modelo hedónico para caraterizar o mercado habitacional, passa por aplicar os métodos OLS ou GWR, com a aplicação de diferentes funções de distância. No método OLS é utilizada uma regressão múltipla sem que sejam consideradas, ou ponderadas, as distâncias das habitações, enquanto, no método GWR, estas localizações são levadas em linha de conta na determinação do preço de uma habitação. Contudo, pretende-se que com a regressão geográfica estendida seja fornecida uma matriz com as diferentes ponderações de pesos espaciais, W, a uma noção não euclidiana e multidimensional. Acredita-se que, com estas abordagens se obtenham resultados com uma correlação entre as variáveis mais elevado em relação aos modelos com as metodologias anteriormente referidas.

Relativamente ao GWR, a distância habitualmente utilizada é a Euclidiana e os coeficientes de regressão são ponderados por noções exclusivamente geográficas. Contudo, o método de cálculo da distância pode ser diferente conforme o tipo de coordenadas que são passadas. Uma mais-valia para a regressão geográfica ponderada é implementar

métodos diferentes no cálculo das distâncias, dando possibilidade ao utilizador em utilizar outros tipos de distâncias conhecidos, por exemplo, a distância de Manhattan ou até introduzir noções de vizinhança não exclusivamente geográficas.

A matriz ponderada de similaridade (vizinhança) é a base da metodologia da regressão geográfica ponderada. O resultado dá origem às ponderações relativas às distâncias utilizando o cálculo do *kernel* na forma gaussiana. Nesta dissertação procura-se estender as ponderações da matriz, não só relativamente às distâncias com coordenadas geográficas mas também a outros níveis, tais como, contextos territoriais, por exemplo, características socioeconómicas, entre outros. Acredita-se que estas abordagens permitem tratar a heterogeneidade espacial de forma mais efetiva e consequentemente aferir mais rigorosamente o valor patrimonial de uma habitação. As métricas estipuladas passíveis de serem utilizadas para a matriz são: nominais (se é vizinho ou não); ordinal (1 para muito perto, 2 para perto, 3 para longe e 4 para muito longe); e escalar (com o valor real da distância em metros ou quilómetros).

3. Ferramentas e Caso de Estudo de Avaliação do Preço da Habitação

No capítulo anterior apresentaram-se várias técnicas de construção dos modelos hedónicos. Nesta secção apresentam-se ferramentas existentes para o cálculo do OLS e GWR. Além disso, é apresentado a base de dados do mercado da habitação de Aveiro-Ílhavo que serve de caso de estudo da presente dissertação.

3.1 Ferramentas Econométricas

A ferramenta Matlab detém a funcionalidade de cálculo para o método OLS, através do comando "*regress*"[7].

$$[\boldsymbol{b}, \boldsymbol{bint}, \boldsymbol{r}, \boldsymbol{rint}, \boldsymbol{stats}] = \boldsymbol{regress}(\boldsymbol{y}, \boldsymbol{X}) \quad \text{Equação 20}$$

Na equação 20, o comando recebe dois argumentos: y representa o vetor dos valores observados e X representa a matriz de n observações das variáveis independentes. A sintaxe utilizada permite devolver cinco parâmetros: b representa o vetor dos coeficientes estimados, $bint$ representa uma matriz n por 2 dos coeficientes estimados com intervalo de confiança de 95%, r representa o vetor dos resíduos, $rint$ representa uma matriz n por 2 para diagnosticar *outliers* e $stats$, onde inclui o R^2.

LeSage (1998) introduziu uma nova ferramenta econométrica para calcular o método GWR tendo por base Brunsdon, Fotheringham et al. (1996). A ferramenta foi desenvolvida com recurso ao Matlab através de scripts m[8]. Na secção 2.4.1 foi referida a forma de que a matriz dos pesos W é determinada, no qual se baseia na distância entre a observação i e as restantes observações dos pontos da amostra. A matriz de pesos W é obtida a partir de uma função de pesos. A equação 10 apresentada na secção 2.4.1 foi apresentada por Brunsdon, Fotheringham et al. (1996) no entanto, podem ser utilizadas

[7] Mais informações sobre o comando "*regress*": http://www.mathworks.com/help/stats/regress.html

[8] Scripts m são ficheiros onde constam comandos Matlab que servem para criar rotinas.

outras funções de peso na regressão geográfica ponderada. Na ferramenta apresentada nesta secção, a função de pesos utilizada é a seguir apresentada:

$$\boldsymbol{W_i^2 = \phi(d_i/\sigma\theta)} \qquad \text{Equação 21}$$

A equação 21 depende de uma função gaussiana ϕ que representa a função densidade de probabilidade normal *standard*, onde σ representa o desvio padrão, θ a largura de banda e d_i o vetor das distâncias.

A ferramenta pretende resolver o problema de otimização relativamente à minimização da função *score* para encontrar a largura de banda ótima por *cross-validation*. A função GWR em *Matlab* determina as pontuações associadas a diferentes larguras de banda. Esta função uni-variada do parâmetro escalar da largura de banda é então minimizada utilizando o algoritmo *simplex* do Matlab "fmin" (LeSage 1998).

3.2 Caso de Estudo

Os dados utilizados para caso de estudo provêm da base de dados da CasaSapo – Janela Digital com informações de habitações dos concelhos de Aveiro e Ílhavo entre 2005 e 2010. O modelo hedónico é composto por um conjunto de atributos das quais foram utilizadas as variáveis a seguir descritas:

1. Preço (Preço da habitação, logaritmizado);
2. Área (Área total da habitação, logaritmizado);
3. Tipo de habitação (variável *dummy*[9]: moradia = 1, apartamento = 0);
4. Estado de conservação (variável *dummy*: novo = 1, usado = 0);
5. Número de quartos (Número de quartos, logaritmizado);
6. Distância ao centro da cidade de Aveiro (logaritmizado);
7. Tempo da habitação no mercado em dias;
8. Variáveis *dummy* relativas aos anos entre 2005 e 2010;

[9] Uma variável *dummy* é uma variável que só pode ter dois estados.

9. Praias (variável *dummy*: perto = 1, longe = 0).

Com base nesta informação, o modelo hedónico geral é representado através da seguinte fórmula:

$$\ln Preço_{€} = f(\ln Area, Tipo, Estado, \ln NQuartos, \ln CBD, \ln TOM, Anos_{2005-2010}, Praias)$$

Partindo do modelo acima, executou-se o método dos mínimos quadrados através do comando "*regress*" do Matlab e, com recurso à ferramenta econométrica do Matlab desenvolvido por James P. LeSage (1998), executaram-se duas instâncias a regressão geográfica ponderada: uma com *kernel* fixo e outra com *kernel* adaptado. Nesta fase, apenas foram analisados os coeficientes de determinação, de modo a verificar se o modelo hedónico utilizado é capaz de explicar convenientemente o preço de uma habitação. Para o modelo foi excluída a variável relativa ao ano de 2005 (por ser uma variável *dummy* complementar às restantes *dummy* do tempo, ou seja, quando as restantes variáveis relativas aos anos entre 2006 e 2010 forem 0, garante-se que a habitação pertence ao ano de 2005. Relativamente à da regressão geográfica ponderada ao modelo hedónico, esta foi aplicada com o *kernel* fixo e o *kernel* adaptado, para posterior análise e comparação dos valores estimados e o R^2. No primeiro caso foi utilizado o método OLS e obteve-se os resultados apresentados na Tabela 1.

Tabela 1 – Valores estimados com o método OLS.

	Coeficientes
Preço	11.2844
Área	0.0023
Tipo	0.1190
Estado	0.1789
Nº Quartos	0.3224
CBD	−0.0458
TOM	0.0097
2005	–
2006	0.0327
2007	0.0494
2008	0.0086
2009	0.0030
2010	0.0211
Praias	0.4410
Nº Observações	**7288**
Erro Variância	**1810.6**
$\boldsymbol{R^2}$	**0.7492**

Na Tabela 1 verifica-se que as variáveis independentes (características habitacionais) do modelo hedónico explicam 74.92% (R^2) do preço da uma habitação.

Para a aplicação da regressão geográfica ponderada, foi utilizado o mesmo modelo hedónico aplicado com o método dos mínimos quadrados. Adicionalmente, existe o parâmetro relativo à largura de banda que implica ponderar diferenciadamente um conjunto de dados. Foi decidido executar a regressão geográfica ponderada com largura de banda igual a 1, já que a largura de banda é expressa nas mesmas unidades métricas das coordenadas utilizadas no conjunto de dados (Charlton, Fotheringham et al. 2009). Na Tabela 2 são apresentados os coeficientes resultantes da aplicação da regressão geográfica ponderada para largura de banda igual a 1.

Tabela 2 – Valores estimados com o método GWR com largura de banda 1.

	Coeficientes
Preço	11.2617
Área	0.0026
Tipo	0.1022
Estado	0.1831
Nº Quartos	0.3113
CBD	−0.0466
TOM	0.0114
2005	–
2006	0.0301
2007	0.0437
2008	0.0054
2009	−0.0046
2010	0.0121
Praias	0.3681
Nº Observações	7288
$\boldsymbol{R^2}$	**0.7639**

Por sua vez, a regressão geográfica ponderada permite determinar de forma otimizada a largura de banda. Assim, foi executado o método utilizando o *kernel* adaptado para obter a largura de banda otimizada e verificar o coeficiente de determinação R^2. A Tabela 3 apresenta os resultados com os coeficientes de regressão e o respetivo R^2.

Tabela 3 – Método GWR com *kernel* adaptado.

	Coeficientes
Preço	10.5897
Área	0.0030
Tipo	0.0837
Estado	0.1684
Nº Quartos	0.2693
CBD	0.0616
TOM	0.0081
2005	–
2006	0.0132
2007	0.0633
2008	0.0318
2009	0.0159
2010	0.0359
Praias	0.0745
Nº Observações	7288
Largura de Banda	$\boldsymbol{h = 0.4303}$
$\boldsymbol{R^2}$	$\mathbf{0.8283}$

Apesar de não serem ponderados os dados espaciais, o método dos mínimos quadrados obteve uma explicação 74.92% para o modelo hedónico analisado, o que indica que as variáveis independentes explicam razoavelmente o preço da uma habitação nos concelhos de Aveiro e Ílhavo. Na regressão geográfica ponderada com largura de banda igual a 1 verificou-se um aumento de quase 1.5% na razão do R^2. Este facto mostra que os dados espaciais são importantes. Por fim, a utilização de um *kernel* adaptado para obter uma largura de banda otimizada revelou resultados tanto positivos como negativos. Por um lado, em relação a uma largura de banda 1, o R^2 obteve um aumento acentuado para os 82.83%. Por outro lado, os coeficientes resultantes do modelo hedónico foram ambíguos. No modelo verifica-se que a distância ao centro da cidade de Aveiro possui uma ponderação com valor positivo, o que sugere que uma habitação mais afastada do centro da cidade de Aveiro tem um preço mais elevado do que uma habitação perto do centro da cidade, o que não corresponde à realidade. O facto de ter sido utilizada uma estrutura de submercados no modelo explica o valor do coeficiente.

3.3 Síntese

Dos modelos hedónicos obtidos a partir de cada metodologia, a regressão geográfica ponderada com largura de banda 1 foi a que obteve melhores resultados e será este o modelo base de estudo para o desenvolvimento de uma biblioteca.

O cálculo da matriz ponderada de similaridade é concebida com base nas variáveis independentes. Para o método em questão, estas representam os atributos das casas e distâncias a determinados pontos, como é o caso da distância ao centro da cidade. Com isto, pretende-se avaliar o quanto estas explicam a variável dependente, o preço de uma habitação. As variáveis independentes utilizadas foram a área, o tipo de habitação, o estado de conservação, a tipologia, a distância ao centro da cidade (CBD), número de quartos, o tempo no mercado da habitação (TOM), as variáveis *dummy* dos anos desde 2005 até 2010 e as distâncias às zonas das praias (Barra e Costa Nova). Para as coordenadas geográficas foram utilizadas as variáveis x e y. Em suma, a expressão a determinar é:

$$\ln Preço_{€} = f(\ln Area, Tipo, Estado, \ln NQuartos, \ln CBD, \ln TOM, Anos_{2005-2010}, Praias)$$

Por norma, os critérios que são passados para a regressão geográfica ponderada são as coordenadas cartesianas. São medidas as distâncias em termos de vizinhança de forma escalar. Porém, estes critérios podem ser medidos utilizando, em vez de coordenadas, características socioeconómicas, rendimentos, entre outros. A incorporação de várias variáveis nos critérios seria vantajoso para o caso da utilização de dados não geográficos.

A matriz de ponderação gerada é no final avaliada para mostrar se as variáveis independentes explicam a variável dependente, neste caso, o preço de uma habitação. Contudo, deve ser sempre verificado o ajuste ótimo ("*goodness of fit*") do modelo. Este fator é explicado pelo coeficiente R^2. A aplicabilidade desta função traz benefícios em termos de análise das larguras de bandas utilizadas, ou seja, quanto maior R^2, menor a largura de banda.

A biblioteca tem o propósito de determinar os modelos hedónicos a partir dos métodos OLS e GWR. Como referido anteriormente, pretende-se que sejam tomados em conta os critérios passados para o método GWR. Quando são fornecidos critérios com dados das distâncias ou não espaciais, então a biblioteca aplica uma métrica de ponderação

para determinar a similaridade (vizinhança). Caso os critérios contenham dados das distâncias Euclidianas ou não Euclidianas, a biblioteca permite a utilização de funções de distância além da Euclidiana.

A aplicação web serve de prova da metodologia a utilizar que obtenha um R^2 superior das restantes metodologias aplicadas. Pretende-se que sejam utilizados na interface web os coeficientes de regressão obtido a partir do modelo hedónico e obter uma visualização das habitações/zonas de acordo com as preferências impostas pelo utilizador.

4. Modelação do Sistema

Nesta secção são apresentados diagramas que compõem as funcionalidades da biblioteca e da aplicação web. Foi seguida a metodologia de desenvolvimento de *software* ICONIX, que consiste na produção de um conjunto de artefactos que retratam a visão dinâmica e estática de um sistema e que vão sendo desenvolvidos incrementalmente e em paralelo (Silva e Videira 2001). Esta metodologia passa essencialmente por quatro fases: levantamento e análise de requisitos, análise e desenho preliminar, desenho detalhado e a fase de implementação. No decorrer do processo do desenvolvimento, os diagramas elaborados descrevem apenas funcionalidades simples, tanto da biblioteca como a aplicação web, e, por isso, não foram elaborados diagramas de robustez, já que estes aprofundam as funcionalidades de um sistema de informação.

4.1 Requisitos do Sistema

Na secção 3.3 foi realizada uma síntese dos métodos e identificaram-se alguns pontos onde se pode acrescentar valor ao método GWR. As ferramentas existentes fornecem pouco ou nenhum controlo sobre o método em questão, o que torna relevante o desenvolvimento de uma ferramenta que forneça essa capacidade. Nesta fase decidiu-se pelo desenvolvimento de uma biblioteca que implemente as metodologias referidas. Assim, esta dissertação foca em duas ferramentas desenvolvidas: uma biblioteca e uma aplicação web.

Biblioteca

- Disponibilizar uma ferramenta que preencha a necessidade de utilizar os métodos OLS e GWR tradicional de forma a otimizar a estimação dos preços de uma habitação.
- Quando o analista utilizar a biblioteca, esta deve fornecer uma forma de determinar diferentes ponderações de pesos espaciais com o método GWR, isto é, utilizar outros tipos de distâncias, como tipicamente é utilizado no método GWR tradicional.

- O método GWR deve possuir a capacidade de estimar valores a partir de dados não espaciais, aplicando ponderações de similaridade nominais. Por exemplo, duas habitações que estejam num determinado submercado são ponderadas, caso contrário tomam valor zero.
- Os resultados obtidos a partir da biblioteca devem ser disponibilizados através de um ficheiro JSON para posteriormente ser consumido pela aplicação web.

Aplicação Web

- Pretende-se que seja possível avaliar habitações que se encontram a um preço justo, isto é, abaixo do valor do mercado do mesmo, ajustar valores da procura com valores da oferta e localizar zonas de potencial satisfação da procura por parte do cliente.
- Procurar habitações que se encontram abaixo do preço limite estabelecido pelo cliente. É utilizado o preço estimado da habitação, ou seja do modelo hedónico, e não o preço de mercado praticado.
- Definir os submercados de habitações de acordo com a pesquisa realizada pelo cliente, ou seja, pretende-se sombrear num mapa os submercados de potencial compra por parte do cliente e visualizar as habitações encontradas.

4.1.1 Atores

Para a biblioteca, o utilizador final é um analista, ao qual espera-se que possua um elevado conhecimento acerca do negócio e sobre os modelos hedónicos de preços. Em relação à aplicação web, além do analista, esta destina-se também ao cliente, onde é alimentada a filosofia de uma interface minimalista e robusta, atendendo aos diferentes níveis de experiência destes atores. Devido à experiência do analista, espera-se que este não tenha quaisquer problemas na interação com a aplicação web. A aplicação web foi desenvolvida como prova de conceito, aplicando um cenário de estudo.

Os atores no sistema são apresentados a seguir:

- **Analista**. Representa o ator que possui conhecimentos acerca dos métodos OLS e GWR e tem permissões para controlar e modificar as funções de distância. Este ator possui também permissões para modificar o modo de funcionamento da aplicação

web, acrescentando funcionalidades, modificar o *layout*, entre outros. O analista pode exportar o resultado do modelo hedónico, para posteriormente ser consumido pela aplicação web.

- **Cliente**. Representa o ator que realiza apenas pesquisas na aplicação web. O cliente não possui qualquer conhecimento acerca das tecnologias e métodos utilizados nas duas ferramentas. Este ator apenas pretende que lhe sejam apresentadas as zonas com habitações de potencial compra por parte deste.

4.1.2 Diagrama de Casos de Uso

Foi considerado o desenvolvimento de um serviço API para receber e disponibilizar os resultados obtidos da biblioteca, no entanto considerou-se um passo não prioritário.

A Figura 2 apresenta o diagrama de casos de uso do sistema global, expondo os dois pacotes: a biblioteca e a aplicação web.

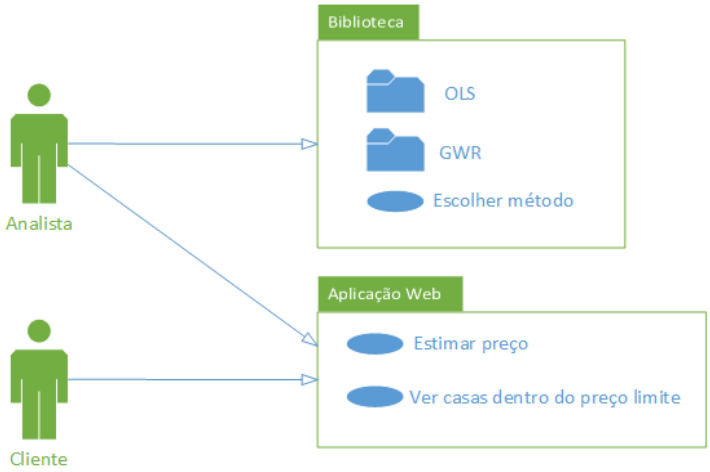


Figura 2 – Diagrama de casos de uso do sistema.

Observa-se na Figura 2 que o analista interage com as duas componentes e o cliente com a aplicação web. De seguida são apresentados os diagramas de casos de uso para cada uma das componentes com uma descrição dos casos de uso.

Biblioteca

Relativamente à biblioteca, a Figura 3 apresenta o diagrama de casos de uso com as operações que o analista pode realizar.

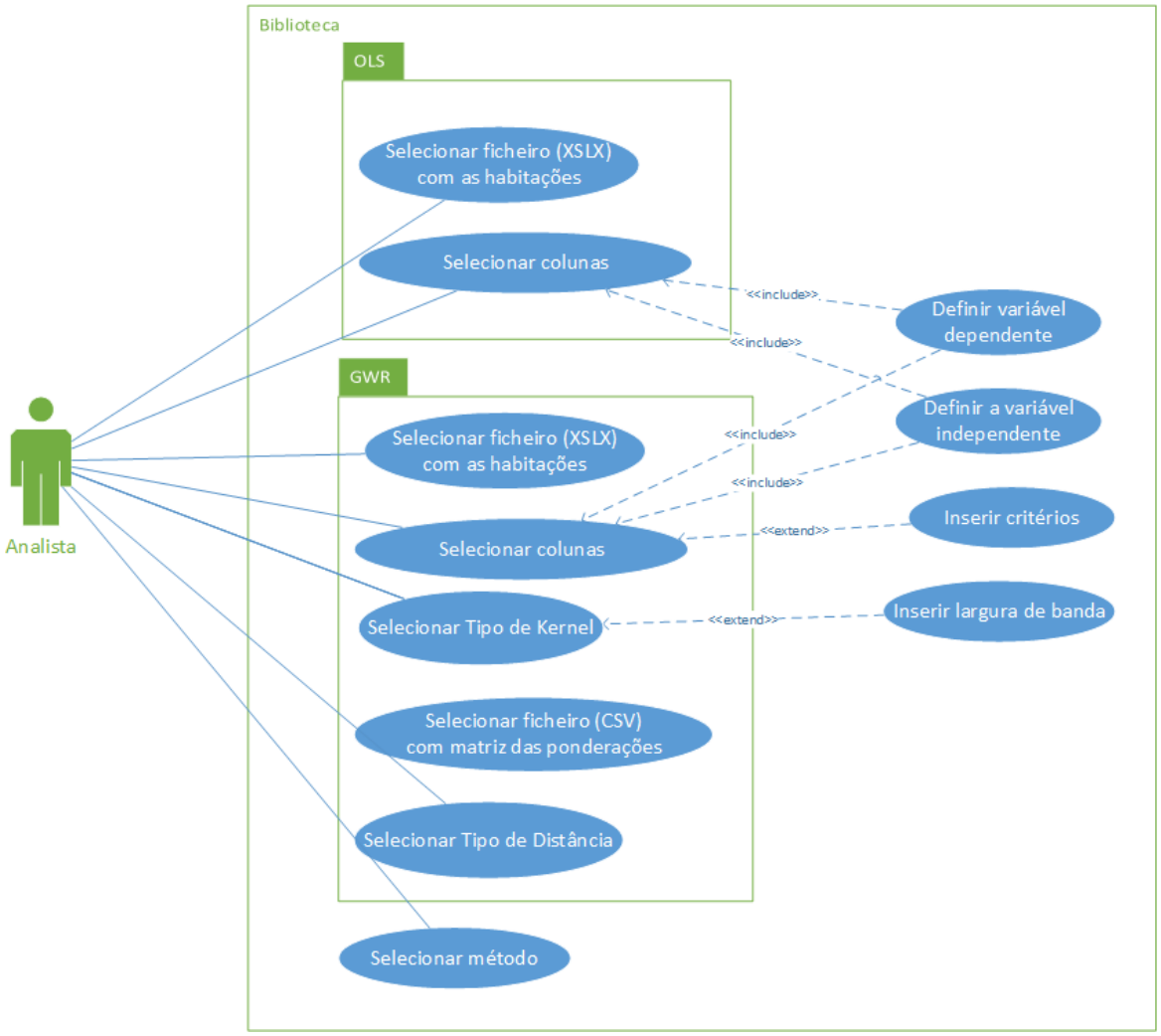


Figura 3 – Diagrama de casos de uso da biblioteca.

Como pode ser observado na Figura 3, o utilizador pode escolher uma das metodologias que pretende executar com a biblioteca, OLS e GWR. De seguida são descritas as funcionalidades apresentadas no diagrama da Figura 3 definindo a prioridade (alta, média ou baixa) de cada um. Relativamente ao método OLS:

- Selecionar um ficheiro Excel (em formato XLSX) – O analista deve fornecer ao programa um ficheiro Excel para posterior seleção das colunas. Prioridade: Alta.

- Selecionar as colunas – O analista deve especificar as colunas que pretende analisar, onde cada coluna representa uma variável. Nesta fase, é apenas pedido ao utilizador o índice da posição da coluna. Prioridade: Alta.

Para o método GWR, o utilizador pode executar as mesmas operações aplicadas ao OLS mas com funcionalidades adicionais da própria metodologia. Assim, as operações possíveis são:

- Selecionar um ficheiro Excel (em formato XLSX) – Funcionalidade idêntica ao método OLS. Prioridade: Alta.
- Selecionar as colunas – Para além de selecionar a variável dependente e a(s) variável(eis) independente(s), é possível inserir os critérios. Nesta fase, é pedido ao utilizador a posição da coluna no ficheiro Excel. Prioridade: Alta;
- Selecionar o tipo de *kernel* a utilizar – Caso seja especificada uma largura de banda é utilizado o *kernel* fixo, caso contrário, é utilizado o *kernel* adaptado. Prioridade: Baixa;
- Selecionar o tipo de distância – Possibilidade de serem especificadas distâncias não Euclidianas, por exemplo, a distância de Manhattan. Prioridade: Alta;
- Selecionar um ficheiro (CSV) com a matriz das ponderações – Opcional. É pedido ao utilizador que especifique um ficheiro CSV que contenha informação acerca das distâncias (pode ser por similaridade ou geográfica). A matriz deve ser quadrada com tamanho idêntico ao número de dados a analisar no ficheiro Excel. Prioridade: Baixa.

Aplicação Web

A aplicação web serve como meio para aplicar os resultados obtidos da biblioteca a um caso de estudo. A Figura 4 apresenta o diagrama de casos de uso com as operações que o utilizador pode realizar na aplicação web.

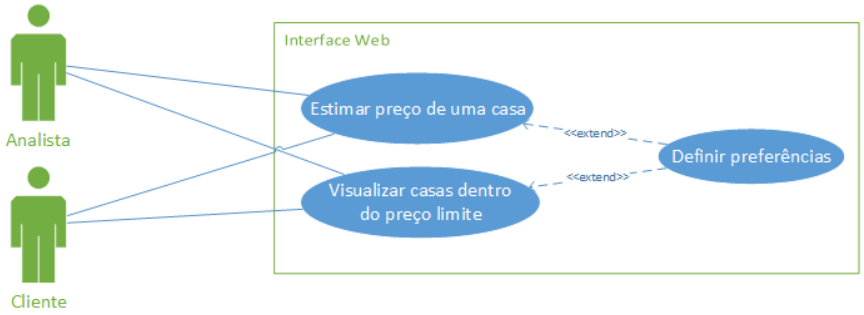


Figura 4 – Diagrama de casos de uso para a aplicação web.

Observa-se na Figura 4 que o utilizador pode realizar duas operações na aplicação web:

- Identificar habitações abaixo do preço limite estabelecido – O analista e o cliente podem introduzir preferências habitacionais, especificando um valor para o preço limite, a área, o tipo de habitação, o estado de conservação, o número de quartos e as distâncias ao centro da cidade e praias. Prioridade: Alta.
- Estimar o preço de uma habitação – Esta funcionalidade permite avaliar se as habitações que se encontram a um preço justo em relação ao preço da mesma no mercado. São apresentadas ao utilizador, as zonas e as habitações cujo preço está abaixo do preço de mercado. Tal como no caso de uso anterior, o analista e o utilizador podem especificar as mesmas preferências sem o preço limite. Prioridade: Alta.

4.1.3 Requisitos não-funcionais

Para as duas ferramentas desenvolvidas, são apresentados os requisitos não-funcionais, descrevendo as propriedades que as ferramentas devem apresentar e os constrangimentos que devem respeitar (Wiegers 2003).

- Confiabilidade. A ferramenta deve estar também preparada para lidar com erros e criar um sistema que dê feedback ao utilizador, evitando que este se desinteresse pelo sistema. A biblioteca recebe um ficheiro XLSX com dados das características das habitações fornecido pelos utilizadores. Prioridade: Alta.

- *Open-Source*. Todas as ferramentas a utilizadas devem respeitar este requisito, bem como as ferramentas desenvolvidas. A ideia passa por fornecer acesso universal através de uma licença livre ao *design* da ferramenta e distribuição do mesmo, dando a possibilidade de ser alterada por terceiros (Lakhani e von Hippel 2003, Gerber, Molefe et al. 2010). Prioridade: Alta.
- Otimização e Performance. A biblioteca executa tarefas computacionalmente intensivas sendo desejável otimizar o tempo de execução.
- Prioridade: Baixa.
- Portabilidade. Tanto o analista como o cliente acedem à aplicação web através de uma ligação à internet. Esta deve ser compatível com qualquer *browser*. A biblioteca deve ser multiplataforma, dando a possibilidade de ser executada em qualquer sistema operativo.
- Prioridade: Média.
- Segurança. A segurança dos dados na aplicação web não foi tomada em conta, já que esta pretende ser utilizada como prova de conceito.
- Prioridade: Baixa.
- Usabilidade. Dado que se considera o cliente como um utilizador pouco experiente, a aplicação web deve possuir uma interface de utilização simples. A biblioteca aplica a mesma filosofia apesar do analista ser considerado como utilizador com experiência entre média e alta.
- Prioridade: Alta.

4.2 Arquitetura do Sistema

Nesta secção é apresentada a arquitetura do sistema, exibindo o modelo de domínio com as classes e relações entre elas, o modelo tecnológico e o diagrama de componentes.

4.2.1 Modelo de Domínio

Os diagramas gerados provêm do *software* Intellij IDEA da Jetbrains[10]. Partindo dos modelos de domínio apresentados é possível descrever o papel que cada classe desempenha. Relativamente ao método OLS é utilizada apenas uma classe, a classe *main*, visto que se pretende utilizar uma biblioteca externa – Apache Commons Math – para o cálculo do método dos mínimos quadrados. A Figura 5 apresenta o modelo de domínio de alto nível com a classe "Main" e os três pacotes "gwr", "ols" e "user_dist_functions".

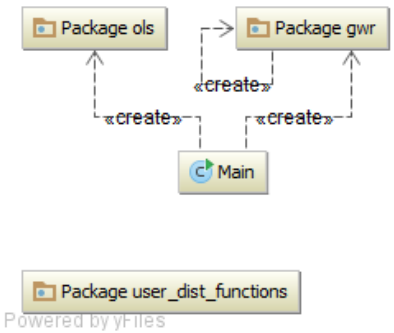


Figura 5 – Modelo de domínio de alto nível.

- "Main" – Classe *main* da biblioteca. Aplica os menus apresentados na linha de comandos e executa as classes e funções dos pacotes "ols" e "gwr". Também executa o método OLS a partir da biblioteca externa referida anteriormente.
- Pacote "ols" – Pacote onde se encontram as funcionalidades analisar e carregar em memória os dados do ficheiro Excel.
- Pacote "gwr" – Pacote de classes onde se encontram as funcionalidades para calcular o método GWR. Tal como no pacote "ols", possui a funcionalidade de analisar e carregar em memória os dados do ficheiro Excel.
- Pacote "user_dist_functions" – Pacote com as classes e ficheiros JAR das distâncias Euclidiana e de Manhattan que servem de exemplo para futuras classes desenvolvidas.

[10] http://www.jetbrains.com/idea/

- “Manhattan” – Classe que aplica a distância de Manhattan.
- “Euclidean” – Classe que aplica a distância Euclidiana.

Apesar de estarem no pacote “user_dist_functions” apenas duas classes para as funções de distância, é possível adicionar outras funções definidas pelo utilizador.

No pacote “ols” encontra-se somente uma classe “AnalyseExcelFileOLS”.

- “AnalyseExcelFileOLS” – Classe que realiza um *parsing*[11] ao ficheiro Excel que extrai os valores das colunas que representam a variável dependente e as variáveis independentes.

A Figura 6 apresenta o modelo de domínio do pacote “gwr”.

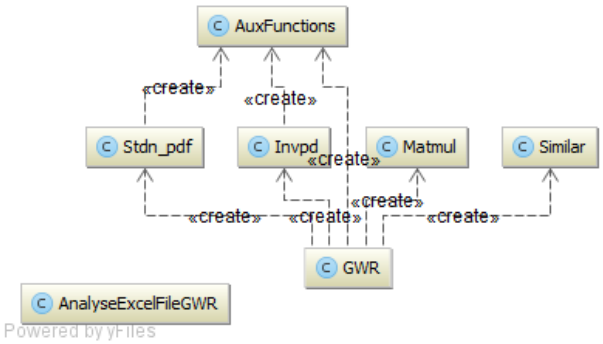


Figura 6 – Modelo de domínio do método GWR.

- “FuncoesAux” – Classe com funções auxiliares para realizar operações com matrizes. Comandos utilizados no Matlab, como o *reshape*[12].
- “Stdn_PDF” – Classe que aplica a função densidade de probabilidade normal.
- “Invpd” – Classe que aplica a pseudo inversão matricial de Moore-Penrose.

[11] *Parsing* é a forma de ler uma fonte (página web, documentos, ficheiros) e extrair informação.

[12] Comando *reshape* do Matlab resulta em obter uma matriz m por n onde os elementos são atribuídos coluna por coluna.

- "Matmul" – Classe que realiza a multiplicação de matrizes mesmo que não tenham a mesma dimensão. Caso as matrizes não sejam compatíveis em colunas ou linhas, é aplicada a função para replicar a matriz descompensada.
- "Similar" – Classe que aplica ponderações conforme a vizinhança de uma habitação.
- "GWR" – Classe que aplica a metodologia da regressão geográfica ponderada com auxílio das classes mencionadas anteriormente.
- "AnalyseExcelFileGWR" – Tal como para o método OLS, a classe realiza um *parsing* no ficheiro Excel fornecido e extrai os valores das colunas especificadas.

Tendo os modelos de domínio definidos, estes evoluem para diagramas de classe que representam a estrutura interna de cada classe, isto é, as propriedades e os construtores e as funções de cada classe. No Anexo A encontra-se uma visualização do diagrama de classes para o método GWR em maior escala.

4.2.2 Diagrama de Componentes

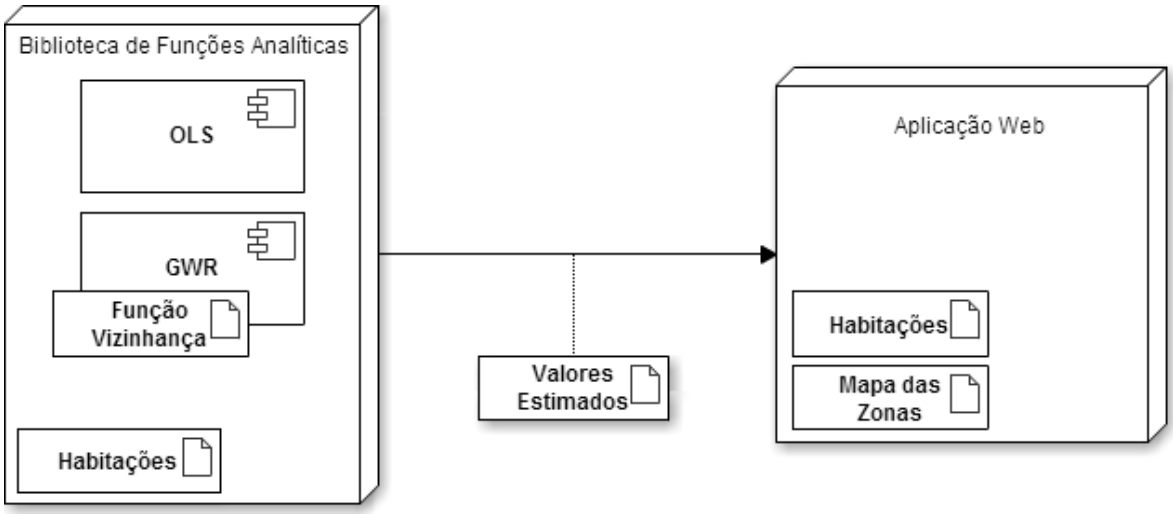


Figura 7 – Diagrama de Componentes.

Na Figura 7 é apresentado um diagrama de componentes que descreve como estes interagem entre si. A biblioteca de funções analíticas utiliza dois módulos implementados em linguagem Java: OLS e GWR. As funções de vizinhança são inseridas no módulo

GWR como ficheiro JAR. A informação das habitações é facultada à biblioteca sob forma de um ficheiro XLSX que, por sua vez, é utilizada para análise dos dois módulos referidos anteriormente. No final são enviados na forma de JSON os valores estimados resultantes dos cálculos.

A aplicação web utiliza, tal como na biblioteca de funções analíticas, informações das habitações sob forma de um ficheiro JSON. O mapa das zonas onde são exibidos os polígonos com os submercados provêm de um ficheiro GeoJSON. A aplicação web, por sua vez, consome o ficheiro JSON com os valores estimados enviada pela biblioteca de funções analíticas.

5. Implementação

Nesta secção são apresentadas as linguagens e tecnologias e utilizadas para o desenvolvimento da componente analítica e componente visual. São descritas as implementações que foram consideradas para cada componente, tomando em consideração possíveis cenários de utilização. No final é apresentado o *workflow* do funcionamento com diagramas de sequência.

5.1 Linguagem e APIs

Tendo por base a estrutura das funções da regressão geográfica ponderada para o Matlab descrita no capítulo 3.1, foi desenvolvida uma biblioteca de funções analíticas e uma aplicação web. A Figura 8 apresenta o modelo tecnológico com a representação das tecnologias utilizadas nos dois componentes.

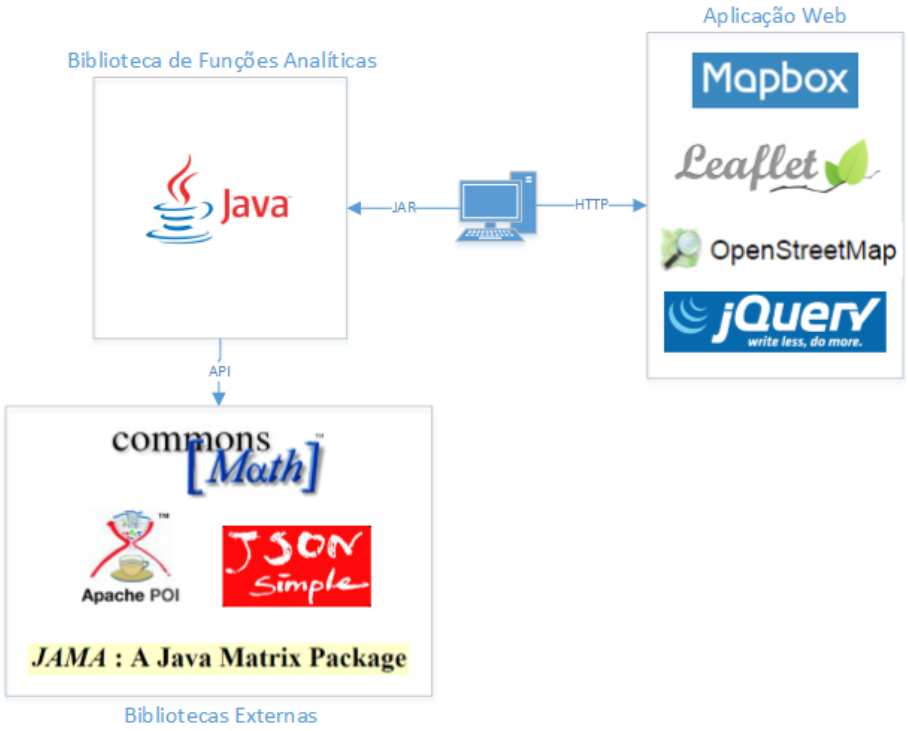


Figura 8 – Modelo Tecnológico.

A componente analítica que aplica as metodologias do método dos mínimos quadrados, da regressão geográfica ponderada e a versão estendida desta, com recurso à

linguagem Java. Além disso, a biblioteca utiliza recursos externos para manipulação de matrizes, a biblioteca JAMA[13], e para a leitura de ficheiros Excel, tanto em formato XLSX como XLS, a biblioteca utiliza a ferramenta Apache POI[14]. Além da metodologia anterior, pretende-se que a biblioteca a desenvolver implemente o método dos mínimos quadrados e, para isso, utilizou-se a biblioteca externa Commons Math[15] da Apache para o cálculo da regressão múltipla. Foi utilizada a biblioteca JSON Simple[16] para a criação de ficheiros JSON.

Relativamente à componente de visualização, a aplicação web foi desenvolvida com recurso às bibliotecas Leaflet[17] e Mapbox[18] utilizando para o efeito as linguagens Javascript, HTML5 e CSS3.

O Java foi escolhido para o desenvolvimento da biblioteca devido a ser uma linguagem utilizada largamente (Deitel e Deitel 2011) e está presente em variadas plataformas, tais como, em *smartphones* (*Android*), em aplicações *standalone* e em aplicações Web. Java é uma linguagem de distribuição livre, multiplataforma e que o autor da dissertação possui conhecimento e não carece de esforço adicional de aprendizagem.

Na componente de visualização foram consideradas algumas ferramentas para apresentar os dados num mapa. Inicialmente, começou-se por recorrer às ferramentas *open-source* OSGeo4W, MapServer, OpenLayers e p.mapper, para trabalhar com os dados das habitações no formato *shapefiles*. Contudo, após a realização de algumas experiências inicias foram escolhidas as ferramentas *open-source* Mapbox.js e Leaflet devido à sua simplicidade, facilidade de utilização e performance. Mapbox.js é um *plugin* baseado no Leaflet que estende e simplifica código proveniente do próprio Leaflet. A opção desde teve por base o formato dos dados a utilizar das habitações, ou seja, o formato GeoJSON.

[13] http://math.nist.gov/javanumerics/jama/

[14] http://poi.apache.org/

[15] http://commons.apache.org/proper/commons-math/

[16] https://code.google.com/p/json-simple/

[17] http://leafletjs.com/

[18] https://www.mapbox.com/

5.2 Componente Analítica - Biblioteca

Na fase inicial foi criada uma biblioteca com o objetivo de aplicar o método dos mínimos quadrados, a regressão geográfica ponderada e a versão estendida desta última, tendo por base a ferramenta já existente em Matlab.

Tendo em conta o caso de estudo referido na secção 3.2, pretende-se estimar valores para as ponderações de cada variável do modelo, sendo depois multiplicadas pelos valores que o utilizador introduzirá na interface web. Assim, procura-se obter a seguinte expressão para o modelo hedónico:

$$Y_p = \alpha + \beta_{area}.X_{area} + \beta_{tipo}.X_{tipo} + \beta_{quartos}.X_{quartos} + \beta_{cbd}.X_{cbd} + \beta_{tom}.X_{tom} + \beta_{2005}.X_{2005} + \beta_{2006}.X_{2006} + \beta_{2007}.X_{2007} + \beta_{2008}.X_{2008} + \beta_{2009}.X_{2009} + \beta_{2010}.X_{2010} + \beta_{praias}.X_{praias}$$

O Y_p é o preço estimado de uma habitação, α é o valor da constante de regressão, os β são os coeficientes de regressão das variáveis correspondentes e os X os valores obtidos a partir da interface web especificados pelo utilizador.

Os dados utilizados para a determinação do modelo hedónico provêm de um ficheiro Excel. A Figura 9 apresenta um excerto do conteúdo do ficheiro Excel.

	A	B	C	D	E	F	G	H	I	J	K
1	X	Y	Preço_ln	Area_ln	Tipo_de_	Estado_d	LnNumQuartos	Dist_CBD_Ave	dCentralityleve2	TOM_ln	ANO_2005
2	-43241,0	108828,0	12,17044547	182,00	0,00	1,00	0,693147181	3,49	7,22	7,561122	0,00
3	-43241,0	108828,0	11,42409425	77,00	0,00	1,00	0	3,49	7,22	7,561122	0,00
4	-43241,0	108828,0	11,45105006	84,00	0,00	1,00	0	3,49	7,22	7,561122	0,00
5	-43241,0	108828,0	12,03231432	160,00	0,00	1,00	1,098612289	3,49	7,22	7,561122	0,00
6	-43241,0	108828,0	11,51292546	94,00	0,00	1,00	0	3,49	7,22	7,561122	0,00
7	-43241,0	108828,0	11,46163217	84,00	0,00	1,00	0	3,49	7,22	7,561122	0,00
8	-43241,0	108828,0	11,44035477	77,00	0,00	1,00	0	3,49	7,22	7,561122	0,00
9	-43241,0	108828,0	12,17044547	187,00	0,00	1,00	0,693147181	3,49	7,22	7,561122	0,00
10	-43241,0	108828,0	12,62237588	186,00	0,00	1,00	1,098612289	3,49	7,22	7,575585	0,00
11	-43241,0	108828,0	11,83982214	130,00	0,00	1,00	0,693147181	3,49	7,22	7,575585	0,00
12	-43241,0	108828,0	11,7651606	110,00	0,00	1,00	0,693147181	3,49	7,22	7,575585	0,00

Figura 9 – Conteúdo do ficheiro Excel.

Observa-se na Figura 9 que cada linha do ficheiro Excel representa uma habitação, excluindo a primeira linha. Cada coluna representa um atributo da habitação. Note-se que nas duas primeiras colunas da Figura 9 encontram-se as coordenadas cartesianas. O

ficheiro Excel a ser utilizado pela biblioteca deve respeitar o formato exemplar apresentado na Figura 9, para ser considerado válido.

A biblioteca deve ser desenvolvida de modo a que seja disponibilizado um menu para que o utilizador possa introduzir os dados na seguinte ordem:

1. Metodologia a utilizar;
2. Ficheiro Excel (.xlsx);
3. Função da distância (.jar);
4. Variável dependente (número da coluna no ficheiro Excel);
5. Variáveis independentes (número das colunas correspondentes no ficheiro Excel);
6. Critérios (onde é aplicada a função de distância ou métrica de vizinhança);
7. Largura de banda a utilizar.

Note-se que, os passos 3, 6 e 7 não se aplicam à metodologia OLS. A função da distância é o foco principal desta dissertação. O propósito é estender o método GWR a outras distâncias que não a Euclidiana aplicada a uma matriz multidimensional. Esta nova aproximação permite a estimação da matriz de pesos W de forma controlada, em vez de uma forma arbitrária, subjetiva e *ad hoc*.

A biblioteca encontra-se preparada para analisar uma pasta onde devem estar colocados os respetivos ficheiros JAR e utilizar a técnica *reflection*[19] conhecida em Java. O analista ao escolher uma função de distância, a biblioteca analisa o conteúdo que se encontra dentro do ficheiro JAR e extrai as classes, os construtores e os métodos. Primeiramente, foram definidas algumas regras para o desenvolvimento de funções de distância:

- A classe Java desenvolvida não deve possuir qualquer referência a pacotes;
- Deve estar definida uma classe pública com o nome exatamente igual ao nome do ficheiro Java;

[19] *Reflection* é frequentemente utilizado por programas que requerem a capacidade de examinar ou modificar o comportamento da execução das aplicações na JVM. (http://docs.oracle.com/javase/tutorial/reflect/)

- Deve possuir um construtor público com dois argumentos: uma matriz para os critérios e um atributo inteiro para índice da observação (Figura 10);

```
public Euclidean(double[][] crit, int index_row){...}
```

Figura 10 – Construtor da classe "Euclidean" para a distância Euclidiana.

O índice da observação representa a atual observação i, onde é determinada a distância entre essa observação e as restantes observações.

- Deve possuir um método público com o nome "getDistance" que devolva uma matriz n por 1 com as distâncias determinadas entre a observação i atual e as restantes observações (Figura 11).

```
public double[][] getDistance() { return this.distance; }
```

Figura 11 – Método "getDistance".

Seguindo as condições acima descritas, a biblioteca aplica a versão estendida do método GWR, dando a possibilidade ao utilizador de aplicar as suas próprias funções de distância. O excerto de código apresentado na Figura 12 expõe a forma como é obtida a classe e o construtor com os dois argumentos do ficheiro JAR.

```
/* Get supplied JAR file */
File dir = new File(filepath).getAbsoluteFile();
URL[] urls = new URL[] { dir.toURI().toURL() };
ClassLoader cl = new URLClassLoader(urls);
String s = dir.getName();
String[] str = s.split("\\.");
// Name of class must be exactly the same name of the file
Class<?> clazz = Class.forName(str[0], true, cl);
Constructor<?> constructor = clazz.getConstructor(double[][].class, Integer.TYPE);
```

Figura 12 – Excerto de código para capturar o construtor da classe.

Em cada iteração é instanciada a classe fornecida, com os devidos argumentos, e, no final é invocado o método com o nome "getDistance" com os resultados da instância. A Figura 13 apresenta o excerto de código que aplica a funcionalidade descrita.

```
Object obj = constructor.newInstance(crit, iter);
Method method = clazz.getMethod("getDistance");
double[][] d = (double[][]) method.invoke(obj);
```

Figura 13 – Excerto de código para instanciar a função de distância.

5.3 Componente de Visualização - Interface

Na segunda fase da dissertação foi desenvolvida uma interface web, onde se pretende estima e apresentar as habitações com preço justo, isto é, o preço estimado desta em relação ao seu preço de mercado atual, ajustar os valores da procura com os valores da oferta, localizar zonas de potencial satisfação da procura, entre outros. Resumidamente, pretende-se disponibilizar uma oferta de habitações que cumpram os requisitos da procura.

Inicialmente, teve-se que definir o modelo para determinar os preços estimados das habitações, ou seja, determinaram-se os valores estimados. Contudo, teve-se em conta o valor do R^2 dos modelos, escolhendo de entre eles o mais alto valor da correlação.

Na interface web são apresentadas ao utilizador as funcionalidades descritas na secção 4.1.2. Pretende-se que o utilizador ao iniciar uma procura de uma habitação forneça as suas preferências relativas às características da habitação ou determinar zonas onde o preço estimado é abaixo do seu valor de mercado.

Tal como indicado na secção 5.1 referente às tecnologias utilizadas, a interface web foi desenvolvida recorrendo ao Mapbox. Nesta é possível criar diversos projetos (mapas), onde é dada a possibilidade de adicionar marcadores personalizados, alterar cores do mapa, alterar para vista de satélite, entre outros. No contexto desta dissertação foram adicionados marcadores ao mapa, de modo a identificar vários pontos de interesse relevantes para o utilizador que procura uma habitação. Entre eles estão:

1. Escolas;
2. Acessos a autoestradas;

3. Hospitais e Centros de Saúde;
4. Zonas industriais;
5. Centros comerciais.

Estes pontos foram identificados dentro da cobertura dos dados no caso de estudo relativo a Aveiro e Ílhavo, no qual não foram adicionados outros marcadores fora da região definida.

Os dados utilizados para identificação das zonas da região de Aveiro e Ílhavo provêm de um *shapefile*, que contém informação geográfica acerca de cada zona da região identificados de forma única através de um ID. Contudo, a ferramenta Mapbox é possível utilizar os dados geográficos no formato GeoJSON, na qual acabou por ser o formato adotado na interface web. Um dos aspetos que se teve em conta foi o sistema de coordenadas utilizado nas *shapefiles*, no qual se verificou pertencer ao sistema EPSG:3763[20]. A projeção EPSG referida anteriormente corresponde ao sistema global de referência ETRS89[21]/PT-TM06, colocando Portugal Continental no mesmo sistema da rede europeia. Para que os polígonos das zonas sejam apresentados corretamente num mapa teve-se de converter a projeção EPSG:3763 para a projeção EPSG:4326[22], sistema geodético mundial de 1984 utilizado pelos satélites GPS. Assim, criou-se um ficheiro JSON com a informação acerca das zonas no formato GeoJSON com o devido sistema geodético, com o objetivo de facilitar o acesso e manipulação dos dados das zonas.

Para determinar as zonas com habitações de potencial compra, a aplicação web deve ter acesso aos dados de todas as habitações dessa região. Foi decidido desenvolver uma interface web que esteja direcionada à região das zonas de Aveiro e Ílhavo, no qual não haverá possibilidade de avaliar habitações de outras regiões ou zonas. Tal como para as zonas, foi criado um ficheiro JSON com a informação das habitações do caso em estudo.

[20] Mais informações em: http://spatialreference.org/ref/epsg/etrs89-portugal-tm06/

[21] ETRS89 – Sistema de Referência Terrestre Europeu 1989

[22] Mais informações em: http://spatialreference.org/ref/epsg/wgs-84/

5.4 Diagramas de Sequência

Nos diagramas seguintes são apresentados os cenários relativos ao cálculo com os métodos OLS e GWR. A Figura 14 apresenta o diagrama relativo ao cenário para o cálculo com o método OLS.

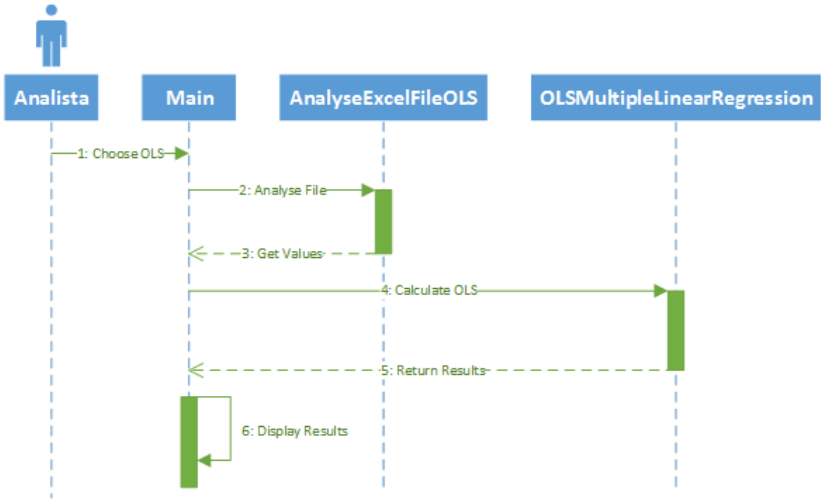


Figura 14 – Diagrama de sequência da biblioteca para o método OLS.

Observa-se na Figura 14 que, para o cálculo com o método OLS, se pretende utilizar três classes: uma classe principal "Main", uma classe para analisar um ficheiro Excel fornecido pela classe principal e uma classe que utiliza uma biblioteca externa que implementa método dos mínimos quadrados, visto que já existem bibliotecas que executem este método e não ser necessário implementar o método OLS.

No método OLS, a classe "AnalyseExcelFileOLS" utiliza como *input* um valor que represente a variável dependente e um vetor de valores de variáveis independentes (1). Estes valores devem ser facultados pelo utilizador e, para isso, pretende-se que seja desenvolvida uma pequena aplicação por linha de comandos onde sejam fornecidas as colunas no ficheiro Excel. No final deste passo (2) devem ser devolvidos um vetor com valores para a variável dependente e um vetor de vetores (matriz) com valores para as variáveis independentes. Os valores obtidos no final do passo anterior são enviados para cálculo para a classe da biblioteca externa (3) e, por fim, são devolvidos os resultados da regressão linear múltipla com o R^2, R^2 ajustado, os coeficientes de regressão, os erros,

entre outros (4). Os resultados devolvidos podem ser tratados como o utilizador pretender ou simplesmente apresentá-los no ecrã (5).

A Figura 15 apresenta o diagrama de sequência quando é utilizado o cálculo com o método GWR.

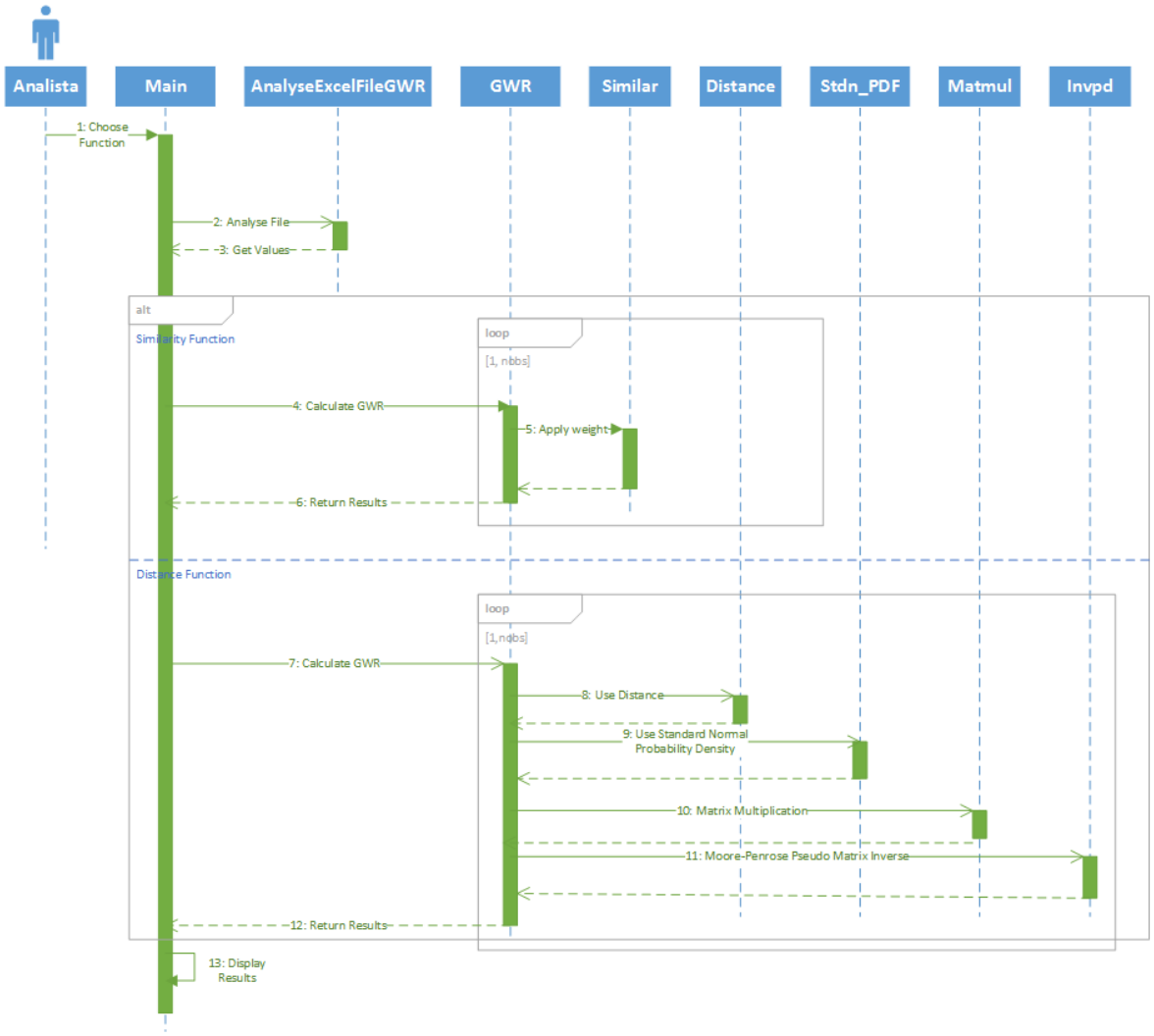


Figura 15 – Diagrama de sequência da biblioteca para o método GWR.

Tal como para o método OLS, pretende-se que o método GWR possua uma classe principal “Main” e uma classe “AnalyseExcelFileGWR”. O analista especifica se pretende utilizar uma função de similaridade ou uma função de distância (1). A classe “AnalyseExcelFileGWR” analisa um ficheiro Excel fornecido pelo utilizador (2-3). Nesta

fase, o utilizador tem a opção de inserir um ficheiro com dados acerca das distâncias entre as habitações seguindo uma métrica específica[23].

Caso o analista tenha escolhido utilizar a função de similaridade, são utilizados como argumentos um vetor para a variável dependente, um vetor de vetores para as variáveis independentes e de critérios (4). A classe GWR executa um *loop* entre as habitações, aplicando ponderações (5) com a classe "Similar" às habitações que se encontram na mesma zona, caso contrário não são aplicadas ponderações. No final são devolvidos os resultados obtidos do método (6).

Caso o analista tenha escolhido utilizar uma função de distância, são utilizados os mesmos argumentos do caso anterior com adição da largura de banda. A classe GWR executa, também, um *loop* entre todas as habitações (8-11), comparando uma com as restantes. Dependendo do número de habitações passadas ao programa são determinados os valores estimados, o R^2 e o R^2 ajustado para cada habitação, utilizando um tipo de distância determinado pelo utilizador (8). Nesta fase, não há razão para a não utilização de outro tipo de distância além da Euclidiana (Charlton, Fotheringham et al. 2009), por isso pode ser facultada uma classe com outro tipo de distância. No passo seguinte (9) é utilizada a classe "Stdn_PDF" que aplica a função densidade de probabilidade para obter a matriz dos pesos. Como os pesos são obtidos com base nas distâncias entre as observações é aplicada a função da equação 10 apresentada na secção 2.4.1. Porém, existem alternativas para construir a função de pesos. Umas delas é a apresentada por LeSage J. na equação 18 na secção 3.1.

A classe "Matmul" tem como intuito realizar multiplicação de duas matrizes que não sejam da mesma dimensão mas que sejam compatíveis em linhas ou em colunas (10). A classe "Invpd" tem o propósito de aplica a pseudo-inversão matricial de Moore-Penrose (11). No final do *loop* são determinados os resultados da aplicação da regressão geográfica ponderada e devolvidos à classe principal "Main" (12). Tal como para o método OLS, o utilizador pode apresentar os resultados no ecrã (13).

[23] Esta métrica pode estar representada em três categorias: nominal (0 para não vizinho e 1 para vizinho), proximidade (1 para perto, 2 para longe e 3 para muito longe) e ordinal (distância real).

6. Resultados

Nesta secção são apresentados os resultados obtidos da implementação das duas ferramentas propostas, realizando uma análise dos mesmos.

6.1 Componente Analítica

A biblioteca Java desenvolvida aceita apenas ficheiros Excel no formato XLSX, contudo não há razão para a biblioteca não suportar outro tipo de ficheiros, por exemplo, XLS (versões anteriores ao Microsoft Office 2003) ou CSV. Na leitura do ficheiro Excel é pedido ao utilizador para indicar quais as colunas que pretende que sejam analisadas para a determinação dos coeficientes de regressão para obter um modelo hedónico, ou seja, as colunas que compõem a variável dependente, as variáveis independentes e/ou os critérios. Aplicando os mesmos atributos que caracterizam a habitação com o método OLS na biblioteca desenvolvida, verificou-se que os resultados foram idênticos, com R^2 de 0.7492 e R^2 ajustado de 0.7488. De notar que para o método OLS, os critérios não são adicionados para o cálculo.

A biblioteca para determinar a regressão geográfica ponderada é mais complexa que o do método dos mínimos quadrados. Esta aplica, como revisto anteriormente na secção 2.4.1, pesos à matriz nas variáveis independentes, com base no cálculo a distância Euclidiana entre as localizações das habitações, e possui a largura de banda. Os valores das larguras de banda foram analisadas, de modo a verificar o comportamento da avaliação do ajuste R^2 e comparar os valores dos coeficientes obtidos na aplicação da mesma regressão em Matlab.

Assim, quanto maior a largura de banda, mais habitações são analisadas na ponderação das distâncias, enquanto, uma largura de banda menor, menos habitações são analisadas. Neste caso particular, foram testados alguns valores para larguras de banda maiores e menores que 1 e analisou-se o R^2. A **Error! Reference source not found.** apresenta os resultados obtidos para as larguras de banda utilizadas na regressão geográfica ponderada.

Tabela 4 – Coeficientes de determinação para diferentes larguras de banda.

h	R^2
3	**0.7493**
2	**0.75**
1	**0.7639**
0.75	**0.7819**
0.5	**0.8173**
0.25	**0.8439**

Como se verifica na **Error! Reference source not found.**, à medida que a largura de banda aumenta, o coeficiente de determinação R^2 diminui. Por outro lado, verifica-se um aumento do R^2 à medida que a largura de banda diminui, obtendo melhores resultados comparativamente à aplicação do método dos mínimos quadrados.

Tendo o R^2 aumentado, analisaram-se os coeficientes de cada variável para as larguras de banda menores que 1, já que foram obtidas melhores correlações entre as variáveis. Na Tabela 5 são apresentados os coeficientes resultantes da aplicação da regressão geográfica ponderada para larguras de banda menores que 1.

Tabela 5 – Valores estimados para larguras de banda menores que 1.

Largura de Banda	$h = \mathbf{1}$	$h = \mathbf{0.75}$	$h = \mathbf{0.5}$	$h = \mathbf{0.25}$
Preço	11.2617	11.2014	10.9328	5.8129
Área	0.0026	0.0027	0.0029	0.0030
Tipo	0.1022	0.0937	0.0811	0.1110
Estado	0.1831	0.1818	0.1729	0.1515
Nº Quartos	0.3113	0.2980	0.2727	0.2764
CBD	−0.0466	−0.0353	0.0111	0.8012
TOM	0.0114	0.0107	0.0091	0.0062
2005	–	–	–	–
2006	0.0301	0.0235	0.0048	0.1926
2007	0.0437	0.0368	0.0365	0.2312
2008	0.0054	0.0018	0.0048	0.2062
2009	−0.0046	−0.0098	−0.0081	0.1832
2010	0.0121	0.0081	0.0101	0.2031
Praias	0.3681	0.0567	0.0302	0.0262
Nº Observações		7288		
$\boldsymbol{R^2}$	**0.7639**	**0.7819**	**0.8173**	**0.8439**

Na Tabela 5 observa-se que a variável independente do ano de 2005 foi excluída, o que mostra que, em relação aos restantes anos, superiores a 2005, o preço de uma determinada habitação pode aumentar ou diminuir conforme o respetivo ano. Alguns valores estimados apresentados na Tabela 5 não estão em concordância. No caso do preço, o valor estimado para este atributo diminuiu consideravelmente, o que não corresponde a um preço base real da habitação. O valor estimado relativo à distância ao centro da cidade (CBD) passou de valor negativo para valor positivo, o que não faz sentido esta variação. Uma habitação mais afastada do centro da cidade tem mais tendência a desvalorizar do que a valorizar. Além disso, uma habitação que se situe perto das praias tem tendência a valorizar significativamente do que ligeiramente o preço da habitação. A razão do coeficiente de determinação (R^2) aumentou à medida que se diminuiu a largura de banda. Este facto revela que efetivamente existe correlação entre as variáveis, no entanto os valores não são fiáveis dentro do contexto da habitação, de acordo com os valores estimados para o preço, distância ao centro da cidade e distância às praias.

Em relação à Tabela 4, valor R^2 diminuiu quando se aumentou a largura de banda. A explicação para este caso deve-se ao facto da largura de banda ser expressa nas mesmas unidades de distância que os critérios inseridos. A regressão geográfica ponderada atribui pesos conforme a largura de banda estipulada, isto é, distâncias superiores à largura de banda são menos ponderadas ou atribuído mesmo valor zero, logo são excluídas da análise.

Na biblioteca foi implementada a função de distância de Manhattan e obtiveram-se os resultados apresentados na Tabela 6.

Tabela 6 – Método GWR com a utilização da distância de Manhattan com largura de banda 1.

	Valor Estimado
Preço (ln)	11.2662
Área	0.0026
Tipo	0.1005
Estado	0.1833
N Quartos (ln)	0.3119
CBD	−0.0479
TOM	0.0113
2005	–
2006	0.0291
2007	0.0418
2008	0.0041
2009	−0.0066
2010	0.0104
Praias	0.3991
Nº Observações	7288
$\boldsymbol{R^2}$	$\mathbf{0.7641}$

Verifica-se na Tabela 6 que com a utilização da distância de Manhattan no método GWR, o R^2 aumentou ligeiramente e os valores estimados para cada variável terem variado ligeiramente em relação à utilização da distância Euclidiana. Esta possibilidade de optar por diferentes funções de distância traz flexibilidade ao método GWR, permitindo ao utilizador poder aplicar uma função de distância que melhor se adeque à sua análise.

No método GWR é possível estimar valores especificando que se pretende que sejam atribuídas ponderações às habitações que são vizinhas, isto é, ponderar as habitações que se encontram no mesmo submercado, todas as restantes são ponderadas com valor

zero. Tal como dito por Grigsby, Baratz et al. (1987) um submercado é definido como "(…) um conjunto de habitações que sejam substitutos razoavelmente próximos entre si mas que sejam substitutos relativamente afastados das habitações dos outros submercados". Os *clusters*, ou submercados, nos concelhos de Aveiro e Ílhavo foram construídos a partir de características territoriais, resultados do projeto DONUT[24]. A Figura 16 apresenta essa segmentação do mercado nos concelhos de Aveiro e Ílhavo.

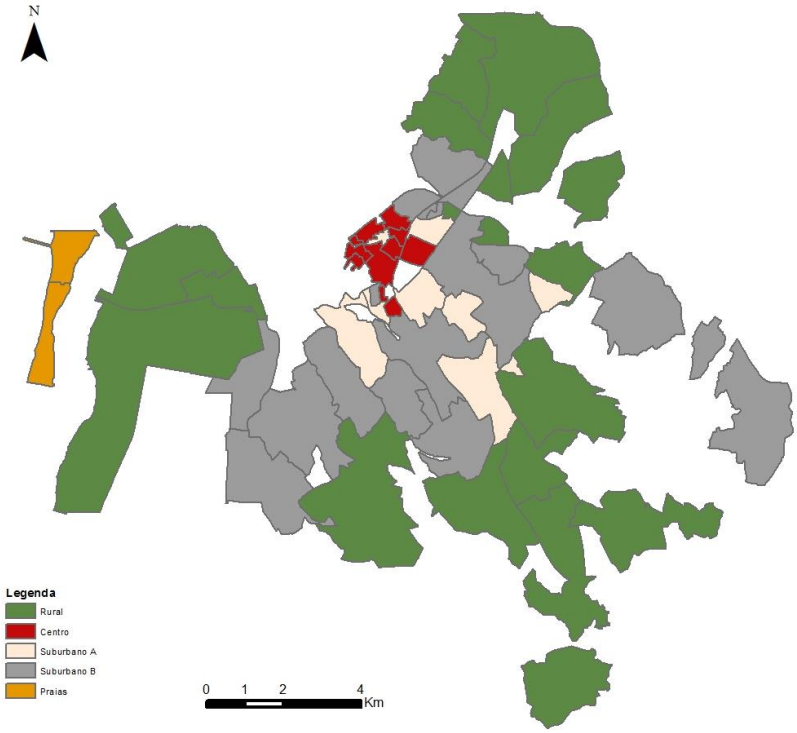


Figura 16 – Definição dos *clusters* dos concelhos de Aveiro e Ílhavo.

Neste caso, em vez de utilizar as coordenadas para os critérios do GWR, foi utilizado como critério o *cluster* de cada habitação. A Tabela 7 apresenta os valores estimados e o R^2 com a utilização do critério dos *clusters*.

[24] Projeto Fatores importantes da procura de habitação em Portugal.

Tabela 7 – Método GWR com métrica nominal de vizinhança.

	Coeficientes
Preço (ln)	10.5210
Área	0.0026
Tipo	0.1226
Estado	0.1661
N Quartos (ln)	0.2740
CBD	0.0183
TOM	0.0090
2005	–
2006	0.0417
2007	0.0507
2008	0.0120
2009	0.0020
2010	0.0184
Praias	0.4394
Nº Observações	7288
$\boldsymbol{R^2}$	$\mathbf{0.7947}$

Verifica-se que na Tabela 7 o valor do R^2 melhorou significativamente, com 0.7947, em relação à utilização de funções de distância. A utilização dos 5 *clusters* no modelo colocou o coeficiente da distância ao centro da cidade com valor positivo, devido ao facto de um dos *clusters* ser o centro e esse capta inteiramente a influência que o centro da cidade tem. A introdução da função de similaridade cobre a parte dos dados não geográficos que inicialmente foi estabelecida nos objetivos. Verificou-se que a largura de banda deixa de ser considerada com recurso à respetiva função, sendo tomadas em conta as ponderações nominais. As ponderações utilizadas na função são de valor 1. Foram utilizados outros valores para a ponderação na função de similaridade, no qual se obtiveram os mesmos resultados apresentados na Tabela 7.

6.2 Componente de Visualização

Na segunda fase do desenvolvimento da interface web, procurou-se utilizar os resultados obtidos da biblioteca Java com as metodologias em estudo nesta dissertação. Tal como referido na secção 5.3, a aplicação web tem como objetivo implementar um caso de

estudo, neste caso para as zonas de Aveiro e Ílhavo, e apresentar os resultados num mapa com recurso às ferramentas referidas na secção 5.1. A aplicação web destina-se a utilizadores que pretendam procurar uma potencial habitação de acordo com as suas preferências ou analisar as zonas onde o preço estimado das habitações está abaixo do preço de mercado atual.

A interface web foi desenhada a pensar na simplicidade e robustez. A ferramenta Mapbox disponibiliza um conjunto de estilos personalizados para o desenho de interfaces web e mapas. Assim, quando o utilizador abre a aplicação web, é apresentada a página inicial como exibe a Figura 17.

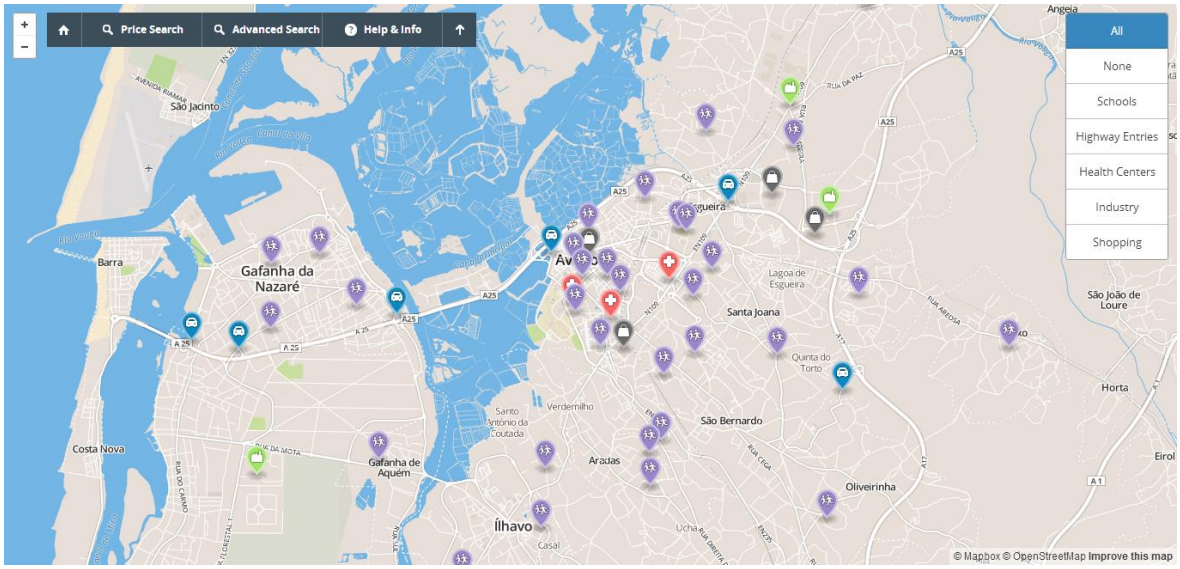


Figura 17 – Página inicial da aplicação web.

Na Figura 17 observa-se que são apresentados ao utilizador três painéis: *zoom*, menu para as pesquisas e menu para os marcadores. O painel do *zoom* encontra-se no canto superior esquerdo, onde é possível realizar as operações para aproximar ou afastar no mapa, caso haja necessidade de visualizar em pormenor algum detalhe específico. À direita do painel de *zoom* encontra-se o painel do menu para as pesquisas e ajuda. Neste painel é possível realizar pesquisas de habitações pelas preferências do utilizador até um preço limite ou pesquisas de zonas com habitações cujo preço é abaixo do preço praticado no mercado. No canto superior direito encontra-se o painel para ativar ou desativar marcadores no mapa. Na Figura 17 observa-se que se encontra selecionada opção para

visualizar todos os marcadores (“All”), no qual são incluídas as localizações de interesse para quem procura uma habitação, por exemplo, escolas ou zonas industriais.

O painel do menu de pesquisas é composto por dois formulários e uma secção para ajuda e informação. Nos dois formulários, pesquisa por preço limite e pesquisa avançada, o utilizador introduz as suas preferências de acordo com os campos que lhe são apresentados. Na pesquisa por preço limite, este deve introduzir, obrigatoriamente, um valor para o preço limite e, para a pesquisa avançada, um valor para a área da habitação que deseja. A Figura 18 e a Figura 19 representam os campos que são disponibilizados ao utilizador para a pesquisa por preço limite e para a pesquisa avançada, respetivamente.

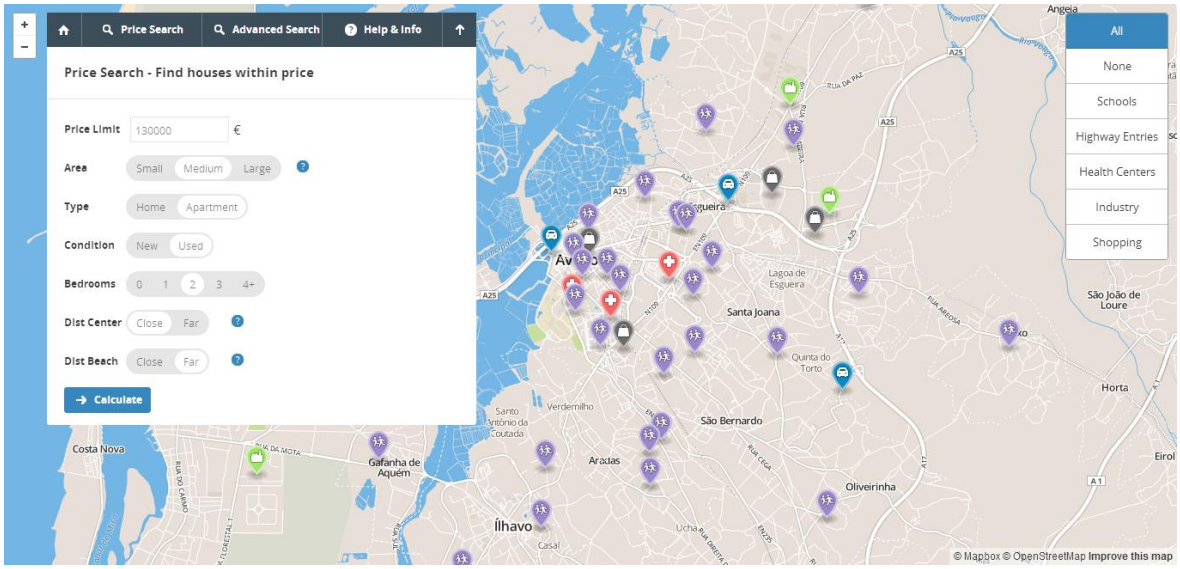


Figura 18 – Menu de pesquisa por limite de preço.

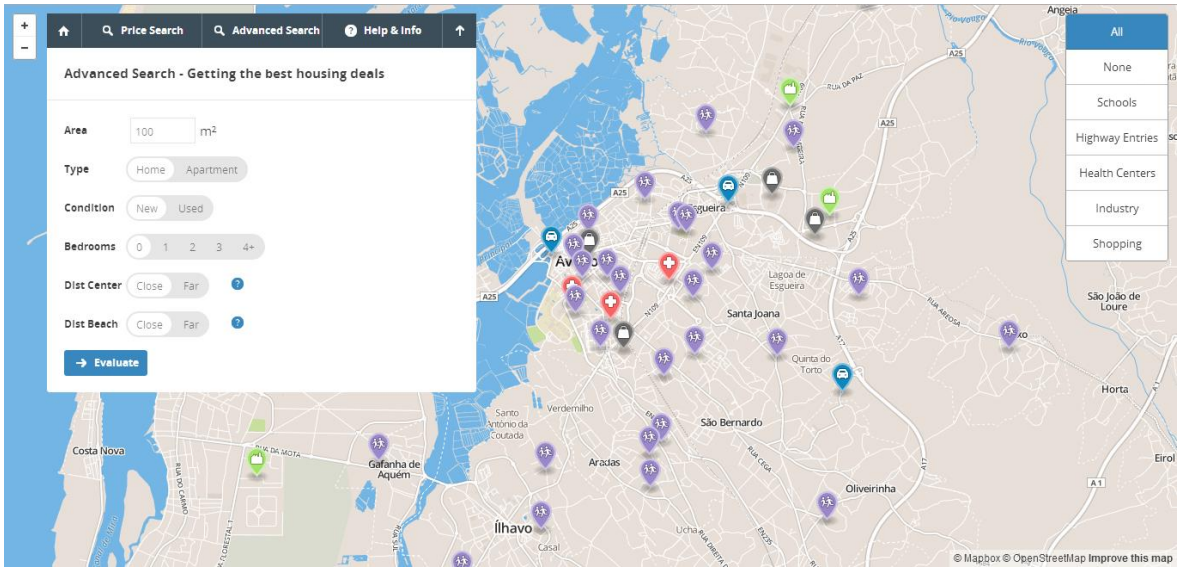


Figura 19 – Menu de pesquisa para as melhores ofertas habitacionais.

Ao serem definidas as preferências por parte do utilizador, é utilizado o modelo hedónico do preço de uma habitação determinado pelo método com melhor correlação entre as variáveis, neste caso, o método GWR estendido com a função de similaridade.

Ao aplicar o caso de uso para determinar habitações no mercado até ao preço limite estabelecido de acordo com as características definidas pelo utilizador, este é sempre informado sobre o preço estimado de uma habitação com essas características. A Figura 20 apresenta um exemplo para o caso da determinação de submercados com habitações com as características a seguir definidas:

- Preço limite: 130.000€;
- Área: média (entre 100 e 150 m^2);
- Tipo de habitação: Apartamento;
- Estado de conservação: Usado;
- Nº de quartos: 2;
- Distância ao centro da cidade: perto (entre 0 e 2 km);
- Distância às praias: longe (mais de 2 km).

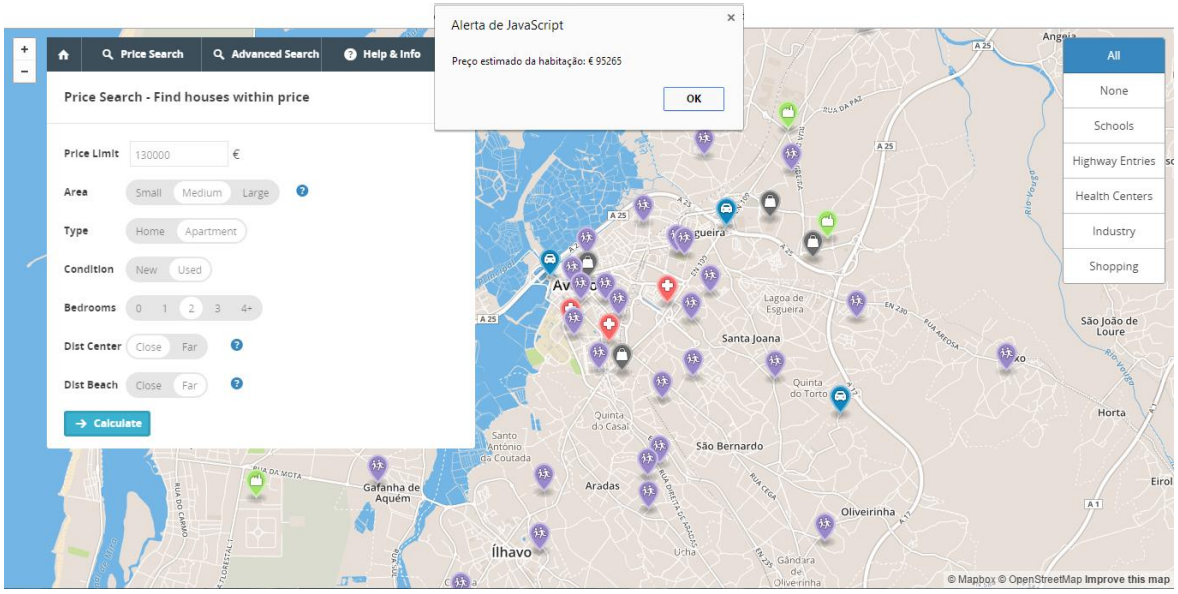


Figura 20 – Preço estimado de uma habitação com as características definidas no formulário.

Verifica-se que para uma habitação com as características definidas, o preço estimado é de 92.265€. De notar que o valor apresentado tem por base o modelo hedónico global determinado pelo método GWR estendido para os submercados dos concelhos de Aveiro e Ílhavo. Ao clicar no botão "OK", são sombreadas os submercados com base em um gradiente de cores desde amarelo e verde, onde amarelo representa poucas habitações com as características pretendidas e a verde as zonas com elevado número de habitações encontradas com preço inferior ao valor estimado pelo modelo. Caso não sejam encontradas quaisquer habitações, as zonas ficam representadas em tom transparente. A Figura 21 apresenta o caso de uma pesquisa pelo preço limite estabelecido.

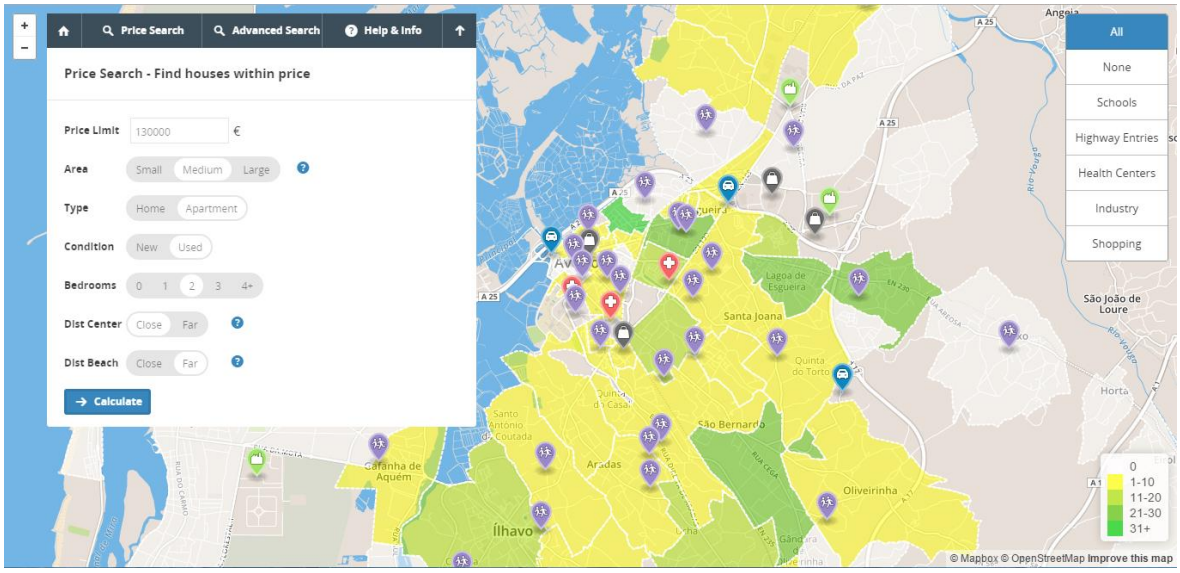


Figura 21 – Resultado da pesquisa por preço limite.

Os submercados apresentados nos resultados têm como base as habitações do ano de 2010. O utilizador pode selecionar um dos submercados com habitações encontradas clicando apenas em cima desta. A Figura 22 apresenta quatro habitações encontradas em um submercado de acordo com as características definidas.

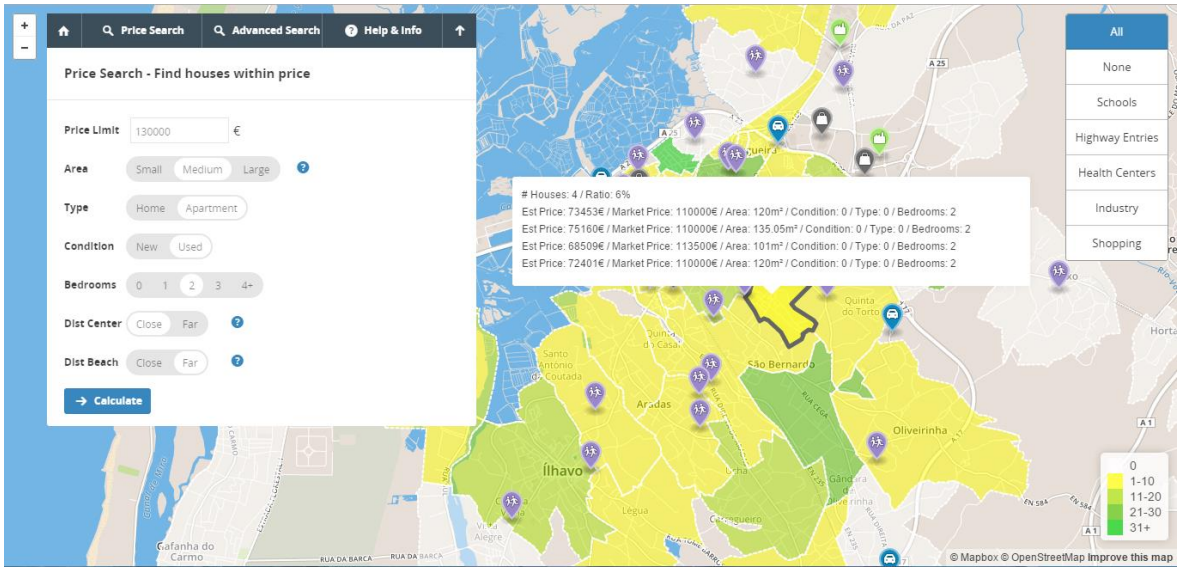


Figura 22 – Habitações de um submercado de acordo com as características definidas.

No menu “Advanced Search” corresponde ao caso de utilização para estimar preço hedónico das habitações de acordo com as características definidas pelo utilizador. A Figura 23 apresenta um exemplo de uma pesquisa com as características definidas:

- Área: 130 m^2;
- Tipo de habitação: Moradia;
- Estado de conservação: Novo;
- Nº de quartos: mais de 4;
- Distância ao centro da cidade: Longe (acima de 2 km);
- Distância às praias: Perto (entre 0 e 2 km).

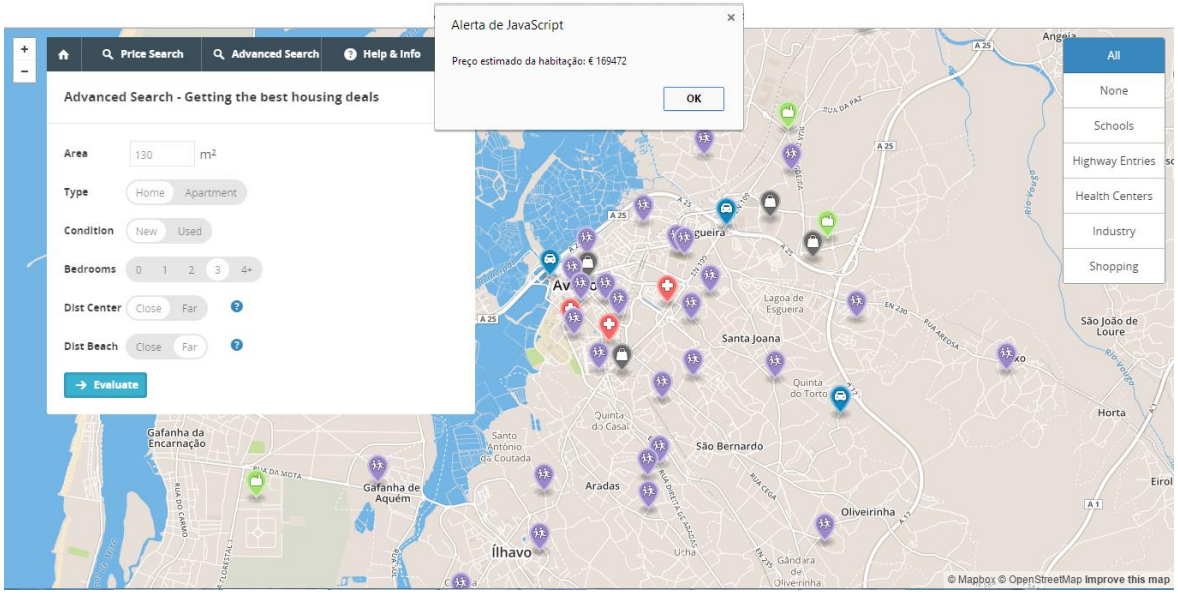


Figura 23 – Preço da habitação de acordo com as características definidas pelo utilizador.

Observa-se na notificação da Figura 23, que o custo estimado de uma habitação com as características definidas é de 358.760€. Uma vez que o número de habitações encontradas com preço de mercado inferior ao valor estimado pode ser reduzido, é indicado ao utilizador o número de habitações e se necessário, este pode prescindir de algumas características para reduzir o custo da habitação. A Figura 24 apresenta a notificação de 7 habitações de acordo com as características definidas.

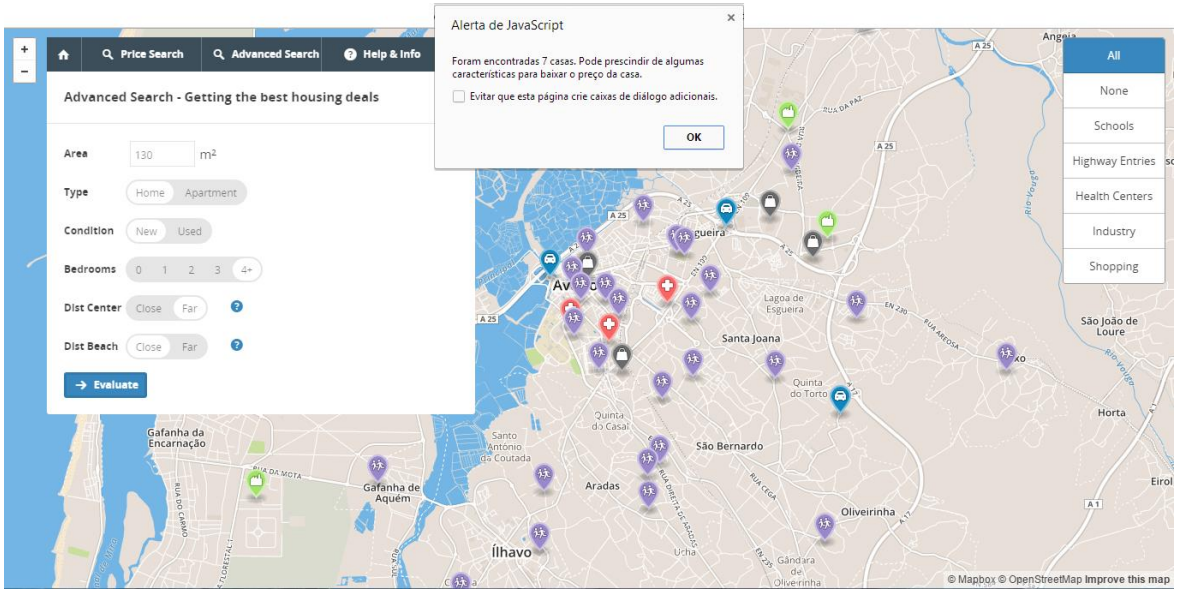


Figura 24 – Notificação do número de habitações encontradas.

Verifica-se que a dispensa de algumas características habitacionais permite reduzir o preço da habitação, algumas mais que outras. Na Tabela 6 e Tabela 7, os valores estimados para cada coeficiente é possível verificar quais as variáveis que mais influenciam o preço da habitação, nomeadamente:

1. Praias;
2. Nº de quartos;
3. Estado de conservação;
4. Tipo de habitação;
5. Área;
6. Distância ao centro da cidade.

A Figura 25 apresenta o submercado sombreado e as habitações encontradas num submercado com potencial negócio, ou seja, abaixo do preço praticado no mercado.

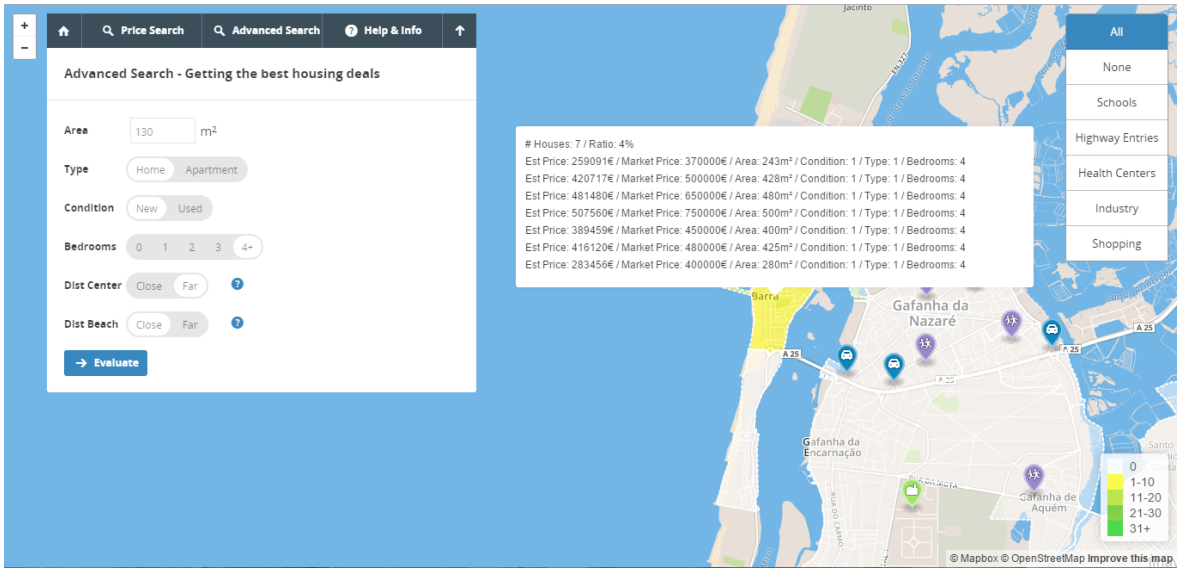


Figura 25 – Apresentação do resultado da pesquisa avançada.

6.3 Discussão

Os resultados obtidos da componente analítica foram positivos, onde se observa a aplicação dos métodos OLS, GWR tradicional e GWR estendido. Neste último, verifica-se que não existem limitações na aplicação de outros tipos de funções de distância aplicados tanto a dados geográficos e não geográficos. Além disso, existe a possibilidade de aplicar uma função de similaridade aos mesmos dados.

O modelo hedónico utilizado nesta dissertação explica, com o método OLS, 74.92% da variação do preço da habitação, contudo, o respetivo método não pondera com base nas distâncias entre as observações. Com a utilização do método GWR tradicional, ou seja, utilizando a distância Euclidiana como meio de ponderação, obteve-se uma razão de 76.39%, onde se verifica que as características intrínsecas da habitação e a distância entre explicam convenientemente o valor de uma habitação. Ao aplicar a distância de Manhattan obteve-se uma razão de 76.41%. Apesar da ligeira melhoria da explicação do modelo no preço da habitação, verifica-se que existe uma maior liberdade de utilização do método GWR para dados geográficos. A utilização do método GWR com recurso a uma função de similaridade verificou-se uma melhoria significativa do R^2 para 79.47%. Na execução foram utilizados como critérios dados não geográficos determinados a partir do projeto DONUT denominados por *clusters*. Esta abordagem traz uma nova noção do espaço, onde esta pode ser tratada a partir de dados que não geográficos e distâncias não Euclidianas.

A componente de visualização apresenta uma prova de conceito onde são apenas utilizados os valores estimados resultantes da componente analítica.

7. Conclusão e perspetivas de trabalho futuro

7.1 Conclusão

O trabalho apresentado nesta dissertação foca-se no desenvolvimento de sistemas de apoio ao estudo e análise do mercado habitacional. Nos concelhos de Aveiro e Ílhavo, caso de estudo utilizado nesta dissertação, verifica-se que existe um mercado habitacional heterogéneo e por isso mesmo procurou-se ajustar os modelos de regressão para a essa realidade. Além disso, esta dissertação mostra a dinâmica do método GWR com utilização de distâncias Euclidianas, distâncias não Euclidianas e métricas de similaridade. A análise com base nesta última referida sustenta que é possível relacionar o mercado habitacional não só ao espaço.

A biblioteca apresentada nesta dissertação aplica as metodologias mais comuns utilizadas na área do imobiliário. O método dos mínimos quadrados é a forma mais simples de determinar o preço de uma habitação de acordo com os seus atributos (área, número de quartos, etc.) ou até características socioeconómicas. Comparativamente ao método dos mínimos quadrados, a regressão geográfica ponderada revelou melhores resultados, já que pondera as distâncias entre as habitações, como foi analisado na secção 2.4.1. Tendo por base as metodologias anteriormente referidas, o objetivo principal desta dissertação foi estender a regressão geográfica ponderada dando a possibilidade de, em vez de ser aplicada a distância Euclidiana (GWR tradicional), aplicar outras distâncias conhecidas. A aplicação da distância de Manhattan obteve um R^2 de 74.41%, aumentado em 0.02% em relação à aplicação do GWR com a distância Euclidiana, utilizando a mesma largura de banda. Verifica-se que ao estender o método tradicional do GWR traz melhores resultados aplicando outro tipo de funções para determinar a vizinhança, ou similaridade, entre as habitações. A utilização da métrica nominal da similaridade apresentou resultados positivos, obtendo um R^2 de 79.47% de correlação das variáveis.

A aplicação web implementada é uma ferramenta que permite visualizar as zonas com habitações de potencial compra e analisar as que se encontram a um preço mais justo. A determinação destas passa por aplicar o modelo hedónico que obteve um R^2 mais elevado de entre os modelos avaliados em cada metodologia, neste caso, o GWR com largura de banda 1 e distância de Manhattan. As ferramentas utilizadas para a construção

da aplicação web foram determinantes e permitiram implementar uma interface relativamente simples e apelativa para qualquer utilizador. A ferramenta web implementa apenas pesquisas, como definido na secção anterior, no qual a definição dos pesos dos coeficientes foram implementados diretamente no código.

A dissertação contribuiu para com a extensão do método GWR para a noção do espaço não Euclidiano multidimensional, tendo uma ferramenta explanatória para estimar preços e preferências habitacionais (Moreira, Moreira et al. 2014). Além disso, a ferramenta contribui para a refinação dos modelos para avaliar as habitações, onde no modelo devem ser identificadas quais as características e como podem estas ser medidas.

7.2 Trabalho Futuro

Após uma revisão dos casos de uso na secção 4.1.2, tendo em conta as prioridades estabelecidas, e uma análise dos resultados na secção 6, verificaram-se que existem aspetos que necessitam de melhorias, tanto na biblioteca como na aplicação web.

A versão estendida do GWR apresentada nesta dissertação distingue-se da versão GWR tradicional pela utilização de outros tipos de distâncias. Adicionalmente, acredita-se que incorporar ou facultar uma matriz de similaridade à versão estendida do GWR, obter-se-ia um R^2 melhorado.

A biblioteca encontra-se a aplicar a similaridade com base numa métrica nominal e seria uma mais-valia aplicar outros tipos de métricas, como por exemplo, de proximidade em 4 níveis: 1 para perto, 2 para médio, 3 para longe e 4 para muito longe. A implementação do *kernel* adaptado traz benefícios para a análise de cada uma das observações. Incluir as médias dos coeficientes de cada zona (submercado).

Importar dados com informações de habitações de outras zonas que não as do caso de estudo. A aplicação web encontra-se centrada para a região das zonas de Aveiro e Ílhavo, já que os dados a analisar são da mesma região.

8. Referências

Anselin, L. (1988). "Spatial Econometrics: Methods and Models", Springer.

Anselin, L. (1998). "GIS Research Infrastructure for Spatial Analysis of Real Estate Markets." Journal of Housing Research **9**(1): 113-133.

Baltagi, B. (2008). "Econometric Analysis of Panel Data", John Wiley & Sons.

Batista, P. (2010). "Data Mining na Identificação de Atributos Valorativos da Habitação". Tese de Mestrado em Planeamento Regional e Urbano, Universidade de Aveiro.

Bennoit, K. (2010). "Ordinary Least Squares Regression." Consultado em 5 de fevereiro, 2014, de http://www.kenbenoit.net/courses/quant1/Quant1_Week8_OLS.pdf.

Bruna, F. and D. Yu (2013). "Geographically Weighted Panel Regression". XI Congreso Galego de Estatística e Investigación de Operacións, A Coruña, Espanha.

Brunsdon, C., A. S. Fotheringham and M. Charlton (1999). "Some notes on parametric significance tests for geographically weighted regression." Journal of Regional Science **39**(3): 497-524.

Brunsdon, C., A. S. Fotheringham and M. E. Charlton (1996). "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity." Geographical Analysis **28**(4): 281-298.

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

Can, A. (1992). "Specification and estimation of hedonic housing price models." Regional Science and Urban Economics **22**(3): 453-474.

Casetti, E. (1972). "Generating Models by the Expansion Method: Applications to Geographical Research*." Geographical Analysis **4**(1): 81-91.

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

Cheshire, P. and S. Sheppard (1995). "On the price of land and the value of amenities." Economica: 247-267.

Crespo R., F. S., Charlton M. (2007). "Application of Geographically Weighted Regression to a 19-year set of house price data in London to calibrate local hedonic price models". 9th International Conference on GeoComputation, Maynooth, Ireland.

Croissant, Y. and G. Millo (2008). "Panel data econometrics in R: The plm package." Journal of Statistical Software **27**(2): 1-43.

Deitel, P. J. and H. M. Deitel (2011). "Java: how to program", Pearson Prentice Hall.

Demšar, U., A. S. Fotheringham and M. E. Charlton (2008). "Exploring the Spatio-Temporal Dynamics of Geographical Processes with Geographically Weighted Regression and Geovisual Analytics." Information Visualization **7**(3-4): 181-197.

Dubin, R. A. (1992). "Spatial autocorrelation and neighborhood quality." Regional Science and Urban Economics **22**(3): 433-452.

Foster, S. A. and W. L. Gorr (1986). "An adaptive filter for estimating spatially-varying parameters: Application to modeling police hours spent in response to calls for service." Management Science **32**(7): 878-889.

Fotheringham, A. S., C. Brunsdon and M. Charlton (2003). "Geographically Weighted Regression: The Analysis of Spatially Varying Relationships", Wiley.

Fotheringham, A. S., M. Charlton and C. Brunsdon (1996). "The geography of parameter space: an investigation of spatial non-stationarity." International Journal of Geographical Information Systems **10**(5): 605-627.

Gerber, A., O. Molefe and A. v. d. Merwe (2010). Documenting open source migration processes for re-use. Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists. Bela Bela, South Africa, ACM**:** 75-85.

Goodman, A. C. (1978). "Hedonic prices, price indices and housing markets." Journal of Urban Economics **5**(4): 471-484.

Grigsby, W., M. Baratz, G. Galster and D. Maclennan (1987). "The Dynamics of Neighborhood Change and Decline". Oxford.

Hsiao, C. (2003). "Analysis of Panel Data", Cambridge University Press.

Huang, B., B. Wu and M. Barry (2010). "Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices." International Journal of Geographical Information Science **24**(3): 383-401.

Lakhani, K. R. and E. von Hippel (2003). "How open source software works: “free” user-to-user assistance." Research Policy **32**(6): 923-943.

LeSage, J. (1998). "Spatial Econometrics".

Marques, J. (2012). "The Notion of Space in Urban Housing Markets". Tese de Doutoramento em Ciências Sociais, Universidade de Aveiro.

Marques, J., E. Castro and A. Bhattacharjee (2012). "A heterogeneidade territorial na compreensão de submercados habitacionais."

Marques, J., E. Castro and A. Bhattacharjee (2012). "Methods and Models for Analysis of the Urban Housing Market". In Emerging Challenges for Regional Development and Evolving Infrastructure Networks and Space. R. Capello and T. P. Dentinho, Edward Elgar.

Mathworks, I. (2013). "Least-Squares Fitting." Consultado, de http://www.mathworks.com/help/curvefit/least-squares-fitting.html.

McMillen, D. P. (1996). "One hundred fifty years of land values in Chicago: a nonparametric approach." Journal of Urban Economics **40**(1): 100-124.

Meen, G. P., M. Andrew, U. o. R. C. f. Spatial, R. E. Economics, T. Great Britain. Dept. of the Environment and t. Regions (1998). "Modelling Regional House Prices: A Review of the Literature", Centre for Spatial & Real Estate Economics, University of Reading.

Mennis, J. (2006). "Mapping the results of geographically weighted regression." The Cartographic Journal **43**(2): 171-179.

Moreira, J., J. Moreira, P. Batista and L. Carta (2014). "Extending Geographically Weighted Regression (GWR) to multi-dimensional non-Euclidean distances - an explanatory tool to estimate housing prices and preferences". IALE - Europe Thematic Workshop, Lisboa.

Pavlov, A. D. (2000). "Space-Varying Regression Coefficients: A Semi-parametric Approach Applied to Real Estate Markets." Real Estate Economics **28**(2): 249-283.

Rosen, S. (1974). "Hedonic prices and implicit markets: product differentiation in pure competition." The journal of political economy: 34-55.

Silva, A. and C. Videira (2001). UML, Metodologias e Ferramentas CASE, Centro Atlântico.

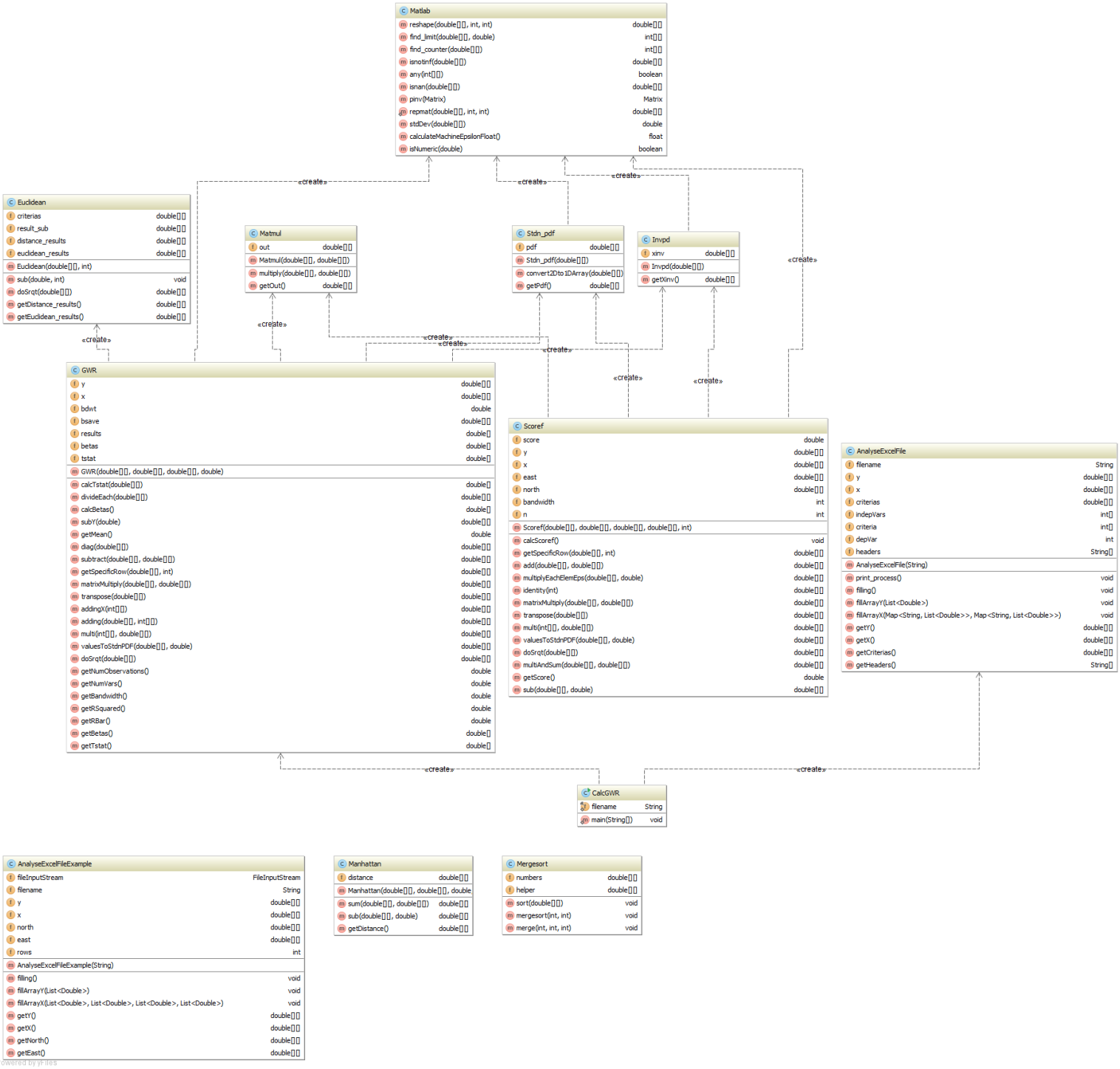
Wang, P. (2006). "Exploring spatial effects on housing price: the case study of the city of Calgary" Master Dissertation, University of Calgary.

Wiegers, K. E. (2003). "Software requirements", Microsoft press.

Yu, D.-L. (2006). "Spatially varying development mechanisms in the Greater Beijing Area: a geographically weighted regression investigation." The Annals of Regional Science **40**(1): 173-190.

Yu, D. (2010). "Exploring spatiotemporally varying regressed relationships: The geographically weighted panel regression analysis."

Anexo A – Diagrama de classes



Anexo B – Criar Função de Distância

O analista tem a possibilidade de desenvolver as suas próprias funções de distância e incorporá-las na biblioteca. A Figura 26 apresenta o exemplo da função da distância Euclidiana.

```
public class Euclidean {

    private double[][] distance;

    public Euclidean(double[][] crit, int index_row){
        double[][] criterias = crit;
        distance = new double[criterias.length][1];
        double temp;

        for (int row = 0; row < criterias.length; row++) {
            temp = 0;
            for (int cols = 0; cols < criterias[0].length; cols++) {
                temp += Math.pow(criterias[row][cols] - criterias[index_row][cols], 2);
            }
            distance[row][0] = temp;
        }

        doSrqt();
    }

    private void doSrqt() {
        for (int i = 0; i < distance.length; i++) {
            distance[i][0] = Math.sqrt(distance[i][0]);
        }
    }

    public double[][] getDistance() {
        return this.distance;
    }
}
```

Figura 26 – Função da distância Euclidiana em Java.

Este quando concluir o desenvolvimento de uma função deve depois converter a classe para um ficheiro JAR. Com recurso à linha de comandos, o primeiro passo é compilar o ficheiro JAVA, o qual vai produzir um ficheiro CLASS através do comando:

```
javac MyClass.java
```

De seguida, deve ser criado o ficheiro *manifest* com extensão TXT com o seguinte conteúdo:

```
Main-Class: MyClass
```

Finalmente, o comando para criar um ficheiro JAR para qualquer uma das plataformas é:

```
jar cfm “MyJar.jar” Manifest.txt MyClass.class
```

Anexo C – Executar Biblioteca de Funções Analíticas

A biblioteca incorpora uma pequena aplicação executável por linha de comandos[25]. O analista pode utilizar a biblioteca em diferentes sistemas operativos, no entanto é exibido o funcionamento em ambiente Linux. Primeiramente, é executada a aplicação com a seguinte sintaxe:

```
java –jar GWRLib.jar
```

É pedido ao analista qual o método que pretende utilizar, qual o ficheiro Excel que pretende analisar e a pasta onde se encontram as funções de distância, como é apresentado na Figura 27.

Figura 27 – Escolha do método, ficheiro Excel e pasta das funções.

No ficheiro Excel são lidos os cabeçalhos de cada coluna, juntamente com a posição da mesma, para o analista especificar a variável dependente, a(s) variável(eis) independentes e os critérios (Figura 28).

[25] Para executar o programa deve estar instalada no computador a versão Java SE Runtime Environment 7+.

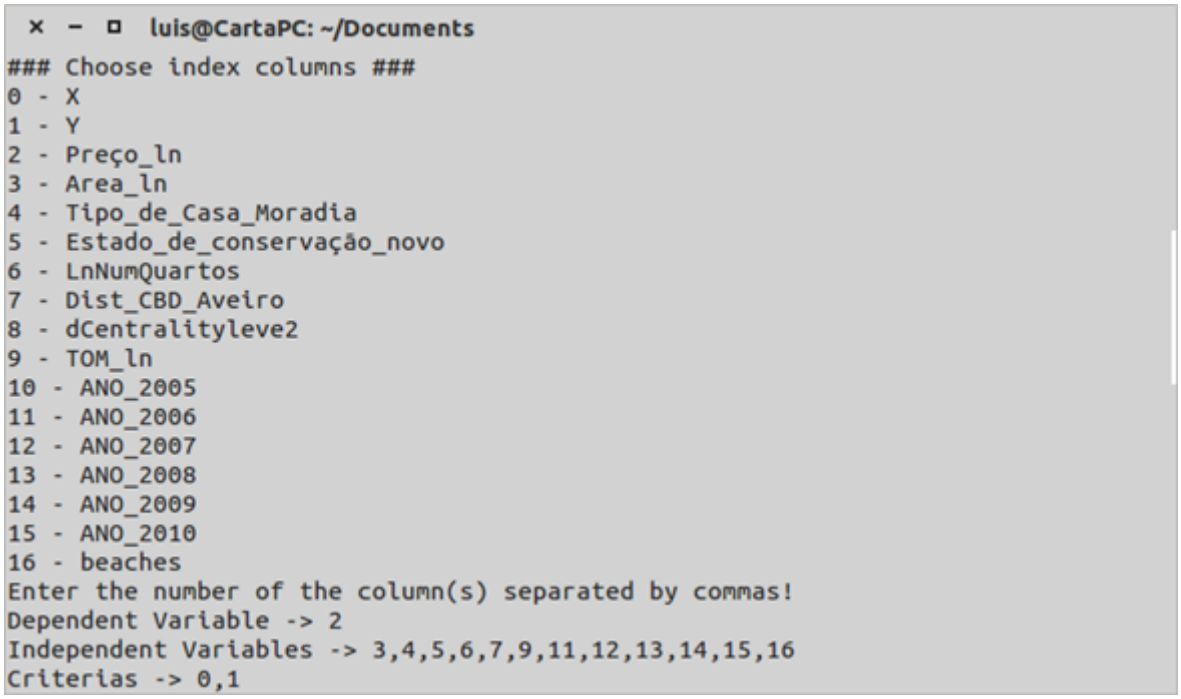


Figura 28 – Seleção das colunas no ficheiro Excel.

Na fase seguinte são adicionados em memória os dados de cada uma das colunas especificadas e é pedido ao analista para definir a largura de banda (Figura 29).

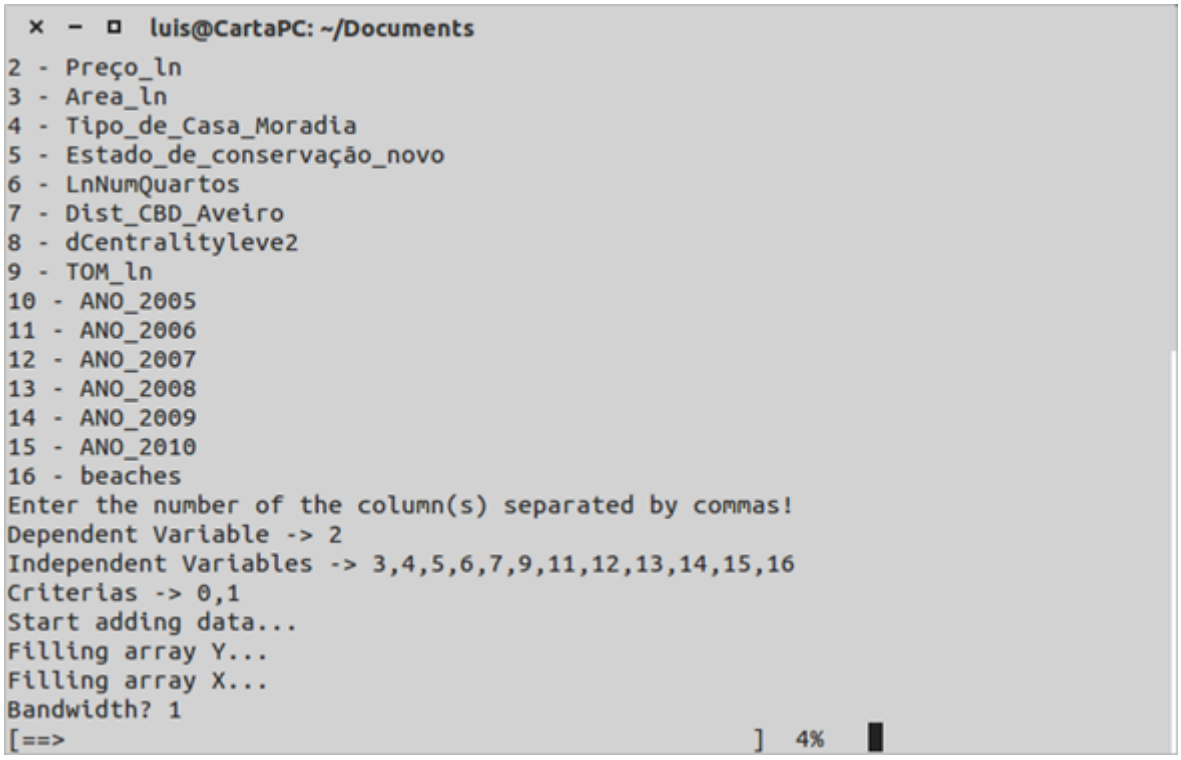


Figura 29 – Definição da largura de banda.

Quando o programa finaliza os cálculos, este imprime na linha de comandos os resultados obtidos, com os valores estimados, o número de observações, o número de variáveis independentes, a largura de banda utilizada, o R^2, o R^2 ajustado, o tempo decorrido dos cálculos e os valores do t de *student* de cada variável (Figura 30 e Figura 31).

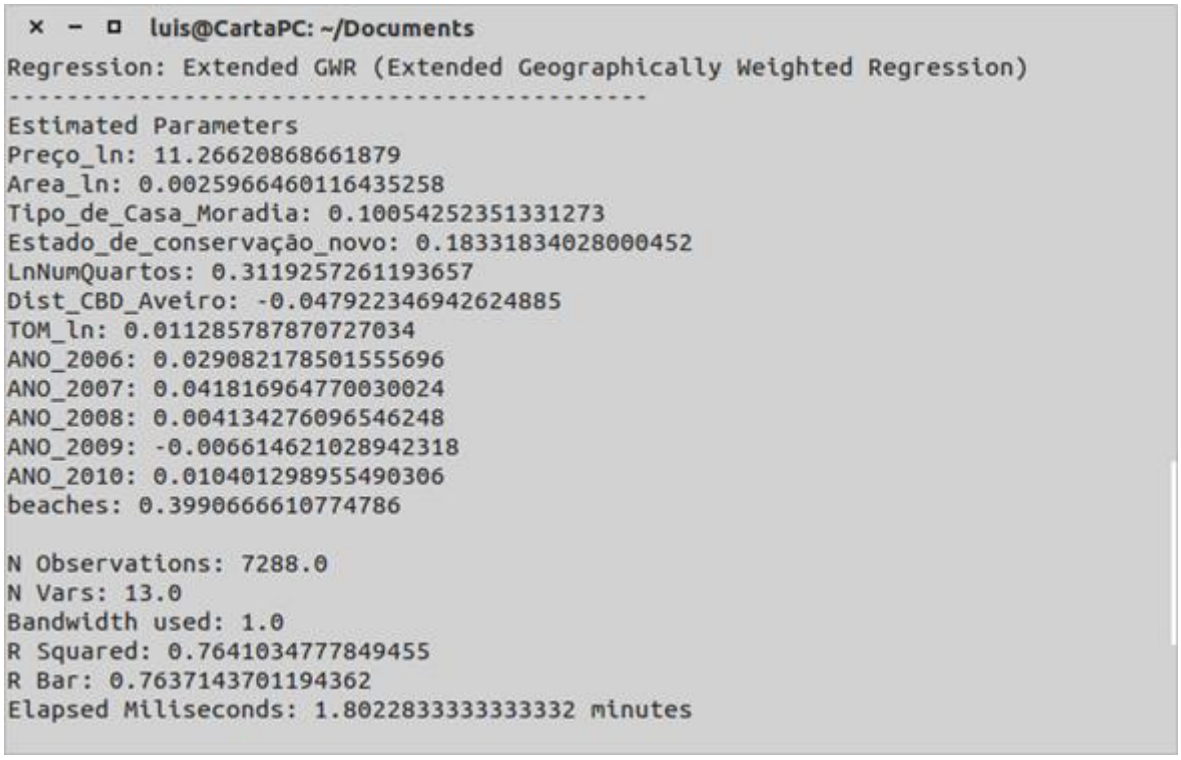


Figura 30 – Resultados do método GWR (1).

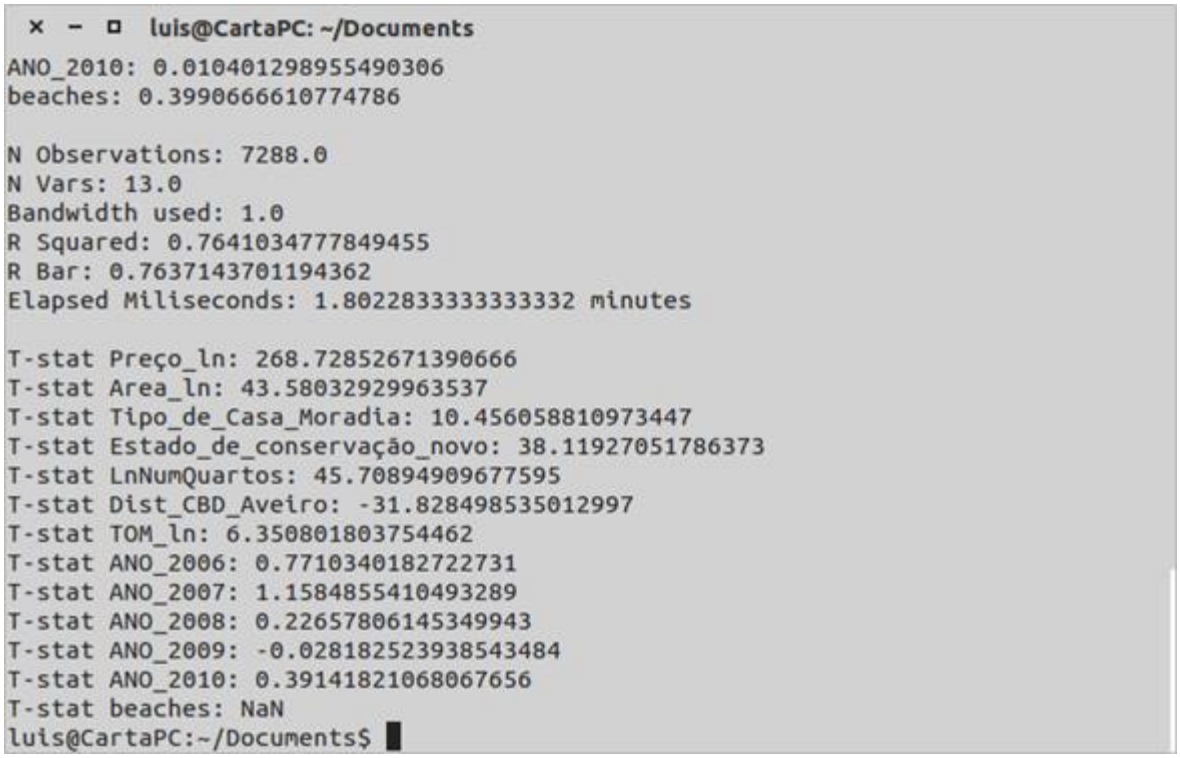


Figura 31 – Resultados do método GWR (2).