

MÉTODO DE ESTIMATIVA DO DÉFICIT HABITACIONAL POR APRENDIZADO DE MÁQUINA PARA UMA CIDADE DE PORTE MÉDIO: RIBEIRÃO PRETO-SP, BRASIL

T. A. da Cunha, J. C. da Silva, A. M. de Almeida, P. K. B. Forcel, F. A. Moreno, R. de
P. Garcia, E. L. Miyasaka, G. A. Cuerva

RESUMO

Atualmente, o déficit habitacional brasileiro gira a casa das 6 milhões de moradias. Além da magnitude do problema, há uma outra lacuna ao estudá-lo: sua localização na escala intraurbana, especialmente segundo seus 4 componentes essenciais. Diante disso, a presente pesquisa é eminentemente prático-objetiva, ou seja, estimar o déficit habitacional para as áreas de ponderação de Ribeirão Preto, valendo-se de Redes Neurais. Uma vez treinada, a Rede Neural pode ser usada para prever ou tomar decisões com base em novos dados de entrada. Antes, porém, foi necessário comparar ao menos dois modelos para reduzir o banco de dados: 1) através do Índice de Correlação de Pearson e 2) a partir da Análise de Componentes Principais. E, calibrá-la aos valores observados de déficit segundo estas mesmas áreas de ponderação. Os dados permitem refletir a acurácia dos procedimentos como um primeiro passo para estimar o déficit para recortes territoriais ainda menores.

1 INTRODUÇÃO

Estimar o déficit habitacional é uma tarefa crucial para o planejamento urbano e políticas públicas direcionadas ao desenvolvimento sustentável das cidades. Neste contexto, o uso de técnicas, mais contemporâneas, de análise de dados, como Redes Neurais em *machine learning*, tem demonstrado grande potencial em investigá-lo segundo diferentes porções do espaço intraurbano. Desse modo, o presente artigo propõe uma abordagem baseada em Redes Neurais para estimar o déficit habitacional inicialmente para as áreas de ponderação de Ribeirão Preto como uma primeira etapa para, em seguida, inferi-lo para seus setores censitários, utilizando para tanto os dados agregados disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) de acordo com o Censo Demográfico de 2010. A metodologia proposta integra técnicas de pré-processamento de dados, construção e treinamento de modelos de Redes Neurais e avaliação da precisão das estimativas obtidas, comparando-as aos dados observados do déficit habitacional ribeirão-pretano.

A qualificação do déficit habitacional é um dos principais desafios enfrentados pelos poderes públicos locais para planejar tanto o modo de ocupação de seus territórios, quanto mensurar a demanda habitacional ao longo do tempo e segundo o perfil de seus residentes. Certamente, essa “cegueira” em nada contribui para a elaboração de Planos Locais de Habitação de Interesse Social (PLHIS) precisos, logo para se compreender a localização e a natureza da defasagem de moradias. Em regiões urbanas como Ribeirão Preto, município localizado no interior do estado de São Paulo, Brasil, o déficit habitacional é uma preocupação que demanda abordagens inovadoras para sua identificação e mitigação. Aliás, a presente pesquisa surgiu de uma demanda específica da prefeitura ribeirão-pretana.

O avanço da tecnologia da informação e o acesso a grandes volumes de dados têm aberto novas possibilidades para abordar problemas complexos por meio de técnicas de análise de dados. Nesse contexto, o *machine learning*, uma subárea da inteligência artificial, tem contribuído com novas perspectivas sobre fenômenos e processos urbanos clássicos a partir de conjuntos de dados heterogêneos e complexos. As Redes Neurais são modelos computacionais inspirados no funcionamento do cérebro humano, capazes de aprender padrões e relações complexas a partir dos dados de entrada. A utilização de Redes Neurais permite a criação de modelos preditivos para lidar com a complexidade e variabilidade dos dados relacionados ao déficit habitacional.

A abordagem proposta neste artigo consiste em três etapas principais: pré-processamento de dados, redução dimensional do banco de dados, construção e treinamento de modelos de Redes Neurais e avaliação da precisão das estimativas. Utilizou-se os dados agregados por áreas de ponderação e a documentação das variáveis segundo setores censitários para tanto. Partiu-se, inicialmente, de um banco de dados com 84 variáveis. Reduzindo-o, em seguida, para um total de 19 variáveis que se encontram tanto no banco de dados de áreas de ponderação, como no de setores censitários, o que permitiria a perfeita comparação entre ambos. Tratam-se, pois, de um conjunto de variáveis muito básicas, como: total de residentes, população masculina, feminina, total de domicílios particulares permanentes, etc..

Espera-se que os resultados deste estudo contribuam para o avanço do conhecimento sobre o déficit habitacional em contextos urbanos brasileiros específicos e inspirem novas pesquisas e iniciativas na área de planejamento urbano e habitação.

2 REFERENCIAL TEÓRICO

2.1 Reflexão crítica sobre a definição brasileira de déficit habitacional e sua operacionalização

Cabe ponderar que há inúmeras definições objetivas de déficit habitacional. Operacionalizá-lo, no entanto, não é sinônimo de abarcar a real necessidade habitacional de uma determinada população. Nesse sentido, o déficit se trata de um recorte prático que não exprime as nuances de conteúdo da realidade habitacional brasileira. É o resultado de uma construção social onde conflitos, interesses, disputas e concórdias ao longo do processo histórico contribuíram para moldá-lo. É dizer, definitivamente a acepção de déficit habitacional atual não é gravada em pedra. Ela, portanto, não é inequívoca, sobretudo, se analisada a partir da experiência de diferentes países ou enquadramentos teóricos. Talvez sequer seja tangível, posto que as condições sociais que o orientam estão em permanente transformação.

Destarte, parece necessário lembrar que o presente artigo não pretende estabelecer um arco que contemple todas as maneiras de calcular o déficit habitacional, senão limita-se a aplicar a metodologia consolidada da Fundação João Pinheiro para calculá-lo (de Miranda-Ribeiro, Viana e Azevedo, 2015). Afinal, o objetivo principal aqui é investigar a aplicação das Redes Neurais na estimativa do déficit à escalas intraurbanas de análise do fenômeno.

Ainda assim, e feita essa pequena digressão, alguma descrição sobre o método da FJP se faz necessária. Discussão essa que objetiva esclarecer as potencialidades da estratégia, bem como suas limitações. Embarços esses que ambicionamos superar ou, no mínimo,

minimizar. Na verdade, que constituem o objeto que estimulou a realização do presente estudo.

Dentro da trajetória histórica de construção do indicador, 2007 é um ano de inflexão para entendê-lo. Antes desse momento, dois dos atuais quatro componentes do déficit habitacional eram, até então, encarados como inadequação habitacional e, logo, não significavam a construção de novas unidades habitacionais, senão sua readequação. Desse modo, tanto o Ônus Excessivo com o Aluguel, famílias com renda de até 3 salários mínimos, residentes em áreas urbanas e que despendiam mais de 30% de sua renda com o pagamento do aluguel, quanto o Adensamento Excessivo de Moradores por Cômodos Considerados Dormitórios em Domicílios Alugados eram, como mencionado, características de domicílios inadequados (de Miranda-Ribeiro, Viana e Azevedo, 2015).

Aliás, é preciso ter em mente que, até essa data, toda forma de coabitação era encarada como déficit habitacional. A partir daquele momento foi possível distinguir entre as famílias conviventes a intencionalidade da coabitação, isto é, se elas realmente desejavam dividir o mesmo teto ou não. Essa mudança foi possível ao se incorporar um quesito específico à Pesquisa Nacional por Amostra de Domicílios.

Nesse sentido, as necessidades habitacionais podem ser separadas em dois grandes domínios; de um lado a necessidade quantitativa de novas unidades habitacionais à incorporação do estoque de moradias brasileiro; de outro, a necessidade qualitativa de aperfeiçoá-lo. À necessidade quantitativa refere-se o déficit habitacional em si. Enquanto que a inadequação habitacional concerne à adequação das moradias existentes sem a obrigatoriedade em se construir imóvel novo.

Assim, atualmente, o déficit habitacional é caracterizado por 4 componentes fundamentais: 1) Habitações Precárias, obtidas pela soma de Domicílios Rústicos, aqueles cujas paredes são de madeira aproveitada, taipa não revestida, palha ou outro material, e Domicílios Improvisados, imóveis e lugares não destinados à habitação (imóveis comerciais, barcos, pontes, vagões de trem, viadutos, barracas, etc.), 2) Coabitação Familiar, compreendida pela soma das famílias conviventes secundárias com intenção de constituir um domicílio exclusivo e das que vivem em cômodo (em geral, cortiços), 3) Ônus Excessivo com Aluguel em Domicílios Urbanos e, por último, 4) Adensamento Excessivo de Moradores por Cômodos considerados Dormitórios em Domicílios Alugados.

Faz-se necessário uma análise crítica especialmente do Ônus Excessivo com Aluguel em Domicílios Urbanos. Como mencionam Miranda-Ribeiro, de Mattos Viana e Salis (2013):

“A partir de diversas análises e reivindicações dos movimentos dos sem casa, percebeu-se que, para a parcela mais pobre da sociedade, o aluguel não é uma opção, mas a melhor alternativa possível.”

Urge ponderar, também, que parcela significativa dos domicílios inadequados é muito precária, basta dizer que há entre eles aqueles sem banheiro, sem energia elétrica, água ou esgotamento adequado. Decerto, são situações mais ou menos reversíveis e remediáveis, porém, talvez parte da inadequação devesse ser encarada como déficit de fato. Não parece ousado afirmar que o déficit é, então, subestimado. Refleti-lo passa, logo, pela análise crítica da combinação entre urgência e necessidade socialmente construída do habitar para um determinado momento e espaço.

Ademais, há um limite inerente à principal fonte de dados utilizada: a Pesquisa Nacional por Amostra de Domicílios (PNAD-IBGE). É verdade que a PNAD é realizada anualmente; sua periodicidade é, portanto, uma grande vantagem, pois permite acompanhar de maneira mais fidedigna a acumulação dos passivos habitacionais ao longo do tempo. Além disso, os quesitos que retratam o déficit habitacional sempre se fazem presentes nas informações básicas da PNAD, diferentemente do que ocorre com outros quesitos de seus suplementos temáticos que são aplicados segundo intervalos de tempo pré-definidos, porém superiores a um ano. Uma qualidade, e ao mesmo tempo uma desvantagem, é o próprio caráter amostral da PNAD. Numa primeira etapa, constrói-se uma amostra de municípios estratificados segundo seus portes populacionais. Uma vez tendo escolhido-os, seleciona-se os setores censitários onde os questionários serão aplicados. Assim, a partir de 2003 a PNAD representa todo o território nacional, entretanto, a partir dos seus critérios de seleção, acaba por enfatizar o caráter urbano da população brasileira. O grande empecilho da PNAD é, então, até onde é possível desagregá-la segundo o tamanho de sua amostra. É possível analisar as informações contidas na PNAD até o nível de Unidades da Federação; no máximo, segundo regiões metropolitanas. Em suma, é impossível calcular o déficit habitacional segundo municípios brasileiros a partir dela. Uma alternativa foi o Censo Demográfico de 2010.

Entretanto, cabe lembrar que a elaboração do Plano Habitacional de Interesse Social (PLHIS) é obrigatória para os municípios que assinaram Termo de Adesão ao Sistema Nacional de Habitação de Interesse Social, uma vez que no ato de adesão eles se comprometeram, entre outras obrigações, a elaborar o PLHIS, considerando as especificidades do local e da demanda. Mas como fazê-lo de modo controlado se não é possível esmiuçar o déficit a partir da escala intraurbana de análise? Por essa razão, o presente estudo se debruça em formas de contornar esse inconveniente.

2.2 Breve recapitulação do conceito de Redes Neurais e sua aplicação aos estudos urbanos

As Redes Neurais têm suas raízes na tentativa de simular o funcionamento do cérebro humano por meio de estruturas artificiais. O termo "Rede Neural" foi introduzido pela primeira vez por Warren McCulloch e Walter Pitts em 1943 (Boreland, Kunze e Levere, 2023), em seu artigo seminal sobre neurônios artificiais. Eles propuseram um modelo matemático de um neurônio simples que poderia realizar operações lógicas.

Os conceitos evoluíram significativamente com o desenvolvimento da computação e da teoria da informação, sobretudo, a partir das décadas de 1950 e 1960. No entanto, foi na década de 80 que as redes se popularizaram com o desenvolvimento de algoritmos de aprendizado como o *backpropagation*, que permitiam treiná-las em várias tarefas.

Desde então, os avanços na capacidade de processamento, conjuntamente com a disponibilidade de grandes conjuntos de dados, têm impulsionado o uso de Redes Neurais em uma variedade de domínios, incluindo reconhecimento de padrões, processamento de linguagem natural, visão computacional, etc..

Por sua vez, a aplicação de Redes Neurais nos estudos urbanos tem se configurado como campo de investigação em expansão, com potencial a oferecer revelações valiosas sobre uma série de questões relacionadas ao planejamento urbano, mobilidade, desenvolvimento

sustentável, qualidade de vida e, por que não, projeções de demanda habitacional (Goh, 1998).

Um dos principais usos das Redes Neurais nos estudos urbanos é na previsão de padrões de crescimento urbano e na modelagem da expansão das cidades ao se analisar dados históricos de desenvolvimento urbano, como densidade populacional, uso da terra e padrões de migração e mobilidade residencial.

Para além da investigação sobre o território, outra aplicação importante das Redes Neurais nos estudos urbanos é interpretar dados sociais e demográficos para entender melhor as necessidades e preferências de uma determinada população. Ao analisar grandes conjuntos de dados de redes sociais, pesquisas demográficas e padrões de consumo, as Redes Neurais podem ajudar os planejadores urbanos a identificar áreas de maior demanda por serviços e recursos, possibilitando um planejamento mais eficiente e direcionado.

Particularmente quanto à questão habitacional, ao analisar grandes conjuntos de dados que incluem informações sobre renda, perfil demográfico, padrões e tendências de mobilidade residencial, características da habitação e preços imobiliários, as Redes Neurais podem identificar padrões complexos e, principalmente, não lineares que influenciam a demanda por habitação em diferentes áreas urbanas. Por exemplo, pode-se utilizá-las para prever onde a demanda por habitação provavelmente concentrar-se-á no futuro, permitindo um planejamento mais direcionado de novos empreendimentos imobiliários e investimentos em infraestrutura. Ao simular diferentes cenários, as Redes Neurais podem ajudar a prever como políticas e intervenções específicas podem afetar a oferta e a demanda por habitação (Zainun, Rahman e Eftekhari, 2010).

Em resumo, as Redes Neurais representam uma ferramenta valiosa para analisar e prever a localização intraurbana do déficit habitacional, orientando políticas e investimentos públicos de maneira mais eficaz e acurada.

3 MATERIAIS E MÉTODOS

No que concerne a abordagem, a pesquisa é eminentemente aplicada e experimental quanto ao método de estimativa do déficit habitacional para pequenas áreas. Isto é, trata-se de uma etapa para lograr estimá-lo para setores censitários. Antes, porém, foi necessário aplicar a técnica de estimativa às próprias áreas de ponderação de maneira a esclarecer suas potencialidades e limitações.

Neste primeiro momento, o procedimento limitou-se à análise das 19 áreas de ponderação do município de Ribeirão Preto, no interior paulista. Tal exercício surgiu, inclusive, de uma demanda da própria prefeitura do município, onde eles desejavam compreender a dimensão do déficit habitacional local de acordo com ao menos três de seus quatro componentes: a) adensamento excessivo de moradores por dormitórios em domicílios alugados, b) ônus excessivo com o pagamento do aluguel e c) coabitação familiar. Aliás, o quarto componente, domicílios precários (compreendido pela soma de domicílios rústicos e improvisados) não lhes era fundamental, pois o poder público local contava com levantamentos de campo que os mapeavam com precisão.

Nesse sentido, cabe discorrer sobre o próprio banco de dados. Inicialmente ele era composto por 84 variáveis, algumas delas, aliás, muito úteis para o cálculo do déficit habitacional; por

exemplo, quantidade de domicílios improvisados por área de ponderação. Infelizmente, nem todas elas estavam presentes na tabela de dados agregados por setores censitários divulgada pelo IBGE. Ou seja, esse seria um imenso obstáculo no futuro à pesquisa do déficit segundo setores censitários, pois as variáveis independentes utilizadas na construção do modelo simplesmente não existiriam no novo banco de dados. A comparabilidade seria também seriamente comprometida.

Por essa razão, reduziu-se o banco de dados a um conjunto de 19 variáveis, sendo 8 dependentes e 1 variável identificadora. Essas 19 variáveis descrevem, por sua vez, perfeitamente tanto áreas de ponderação, quanto setores censitários. A sincronia entre os bancos de dados e ambos os recortes territoriais seria perfeito, permitindo a comparabilidade entre as estimativas de déficit, já que o modelo poderia ser aplicado sem quaisquer modificações. Tratam-se, porém, de variáveis muito básicas como: total de residentes segundo sexo, total de domicílios particulares permanente, raça e renda. Isto é, variáveis que muito tangencialmente mantinham alguma relação com o déficit habitacional. Ou melhor, não eram diretamente utilizadas para calculá-lo. Como discute-se mais à frente, esse será um grande obstáculo para a redução do erro absoluto.

Ainda assim, houve um esforço em reduzir ainda mais o conjunto de variáveis. Desejava-se, ao sintetizá-lo, igualmente diminuir a redundância descritiva em torno do fenômeno e, assim, aumentar a confiabilidade estatística. Primeiro, é preciso entender que quando há muitas variáveis descritivas a quantidade de dados na Rede Neural passa a ser improdutiva, complexificando a arquitetura do modelo de aprendizagem. Nesse sentido, optou-se por reduzir a dimensionalidade do fenômeno ao agrupar variáveis que mantêm maior relação entre si. O fenômeno é caracterizado não mais pelas variáveis, senão pelas dimensões que as aglutinam.

Além disso, o agrupamento das variáveis permite remover a redundância e o ruído do banco de dados, visto que muitas das variáveis podem estar correlacionadas entre si. Ademais, pode haver variáveis que contenham pouca informação útil. Com um conjunto menor e mais significativo de variáveis, as estimativas estatísticas tendem a ser mais estáveis e menos sensíveis a pequenas variações nos dados.

Num primeiro exercício experimental, sintetizou-se o conjunto de dados a partir de duas técnicas estatísticas: a) Índice de Correlação de Pearson (ICP) e b) Análise de Componentes Principais (ACP).

O Índice de Correlação de Pearson é uma medida estatística que avalia a relação linear entre duas variáveis contínuas. Ele varia de -1 a +1, onde: +1 indica a perfeita correlação positiva, ou seja, ambas aumentam juntas na mesma direção em, também, perfeita proporcionalidade; -1 indica uma correlação negativa perfeita, ou seja, as variáveis variam em direções opostas, isto é, são inversamente proporcionais, enquanto uma cresce, a outra diminui e quanto mais próximo de 0 (zero) rondar o índice, menor é a correlação linear entre as variáveis.

O ICP permite esclarecer a redundância, posto que se duas variáveis têm uma correlação alta (próxima de +1 ou -1), elas, em realidade, estão fornecendo informações semelhantes sobre o fenômeno. Portanto, uma delas pode ser removida sem grandes prejuízos para interpretá-lo. Em resumo, valer-se do ICP pode prevenir a multicolinearidade. Ela ocorre quando duas ou mais variáveis independentes em um modelo de regressão estão altamente correlacionadas.

Por sua vez, a Análise de Componentes Principais (ACP) objetiva principalmente reduzir a dimensionalidade do conjunto de dados. Ela realiza essa redução transformando as variáveis originais em um novo conjunto de variáveis não necessariamente correlacionadas chamadas de componentes principais. Primeiro, os dados são centrados subtraindo-se a média de cada variável. Isso garante que a média de cada variável seja zero. Em seguida, calcula-se a matriz de covariância ou correlação entre as variáveis, a depender da situação. Os componentes principais são obtidos a partir dos autovetores e autovalores da matriz de covariância ou correlação. Os autovetores representam a direção dos novos eixos nos quais os dados serão projetados, enquanto os autovalores representam a quantidade de variação explicada por cada componente principal. Os componentes principais são, então, ordenados de acordo com a quantidade de variação explicada. Normalmente, os primeiros componentes principais explicam a maior parte da variação dos dados, algo como 85% dos casos (Bharadiya, 2023).

Ambas as técnicas apresentam vantagens e desvantagens. A aplicação do ICP é simples; tanto para calculá-lo, quanto para interpretá-lo. Porém, ele ignora relações não lineares entre variáveis, bem como desconsidera variáveis latentes, isto é, despreza a estrutura subjacente dos dados. Por fim, ao não reduzir a dimensionalidade dos dados, o critério de seleção das variáveis é, em última instância, subjetivo.

Por outro lado, a ACP avança justamente sobre as limitações do ICP: redução da dimensionalidade, identificação de padrões complexos e remoção da multicolinearidade. Entretanto, interpretar os componentes principais pode ser desafiador, especialmente quando há muitos componentes ou quando as variáveis originais têm naturezas e significados totalmente diferentes entre si. Aliás, a ACP ao transformar as variáveis originais em componentes principais, que podem ser combinações lineares das variáveis originais, complexifica a interpretação direta dos resultados, uma vez que as dimensões passam a descrever o fenômeno, não mais as variáveis originais.

O próximo passo foi treinar a Rede Neural (Boreland, Kunze e Levere, 2023) utilizando o novo banco de dados, agora, com um conjunto de variáveis muito mais reduzido. Porém, antes se faz necessário esclarecer algumas características do aprendizado de máquina. Pode-se resumi-lo a um conjunto de técnicas para “ensinar” o computador a como compreender e interpretar dados. Seu uso justifica-se na crescente dificuldade em interpretar e buscar padrões, de maneira manual, para um conjunto de dados cada vez maior e desafiador. De maneira geral, estas técnicas podem ser classificadas em supervisionadas ou não supervisionadas, a depender da disponibilidade de uma ou mais variáveis de saída. O objetivo das abordagens supervisionadas é construir um modelo matemático-estatístico capaz de prever variáveis de saída a partir das variáveis de entrada. As abordagens não supervisionadas contam apenas com as variáveis de entrada.

4 RESULTADOS E DISCUSSÕES

Antes de qualquer análise sobre a eficácia do uso de Redes Neurais na estimativa do déficit habitacional ribeirão-pretano, considera-se fundamental discorrer justamente sobre a manifestação espacial dele, déficit (**Figuras 1, 2, 3 e 4**).

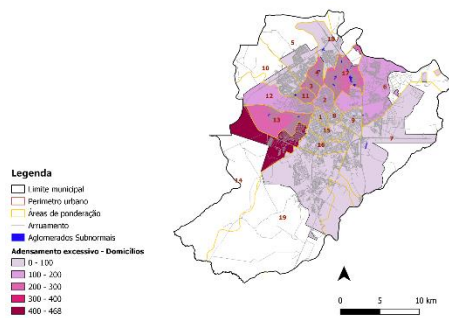


Fig. 1 Adensamento excessivo de Moradores por Dormitório em Domicílios Alugados segundo áreas de ponderação, Ribeirão Preto, 2010.

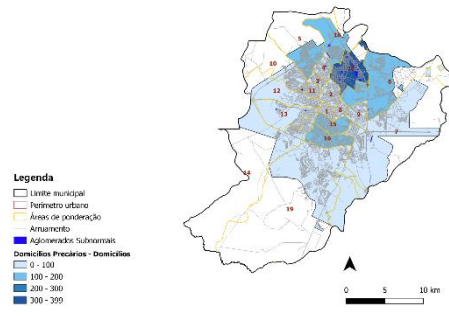


Fig. 2 Domicílios Precários segundo áreas de ponderação, Ribeirão Preto, 2010.

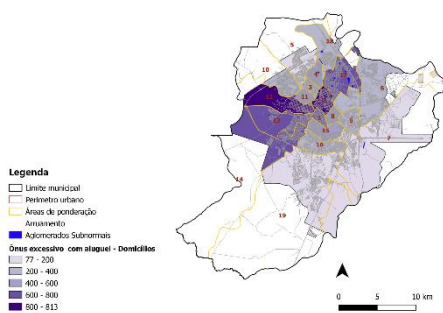


Fig. 3 Ônus Excessivo com Pagamento de Aluguel segundo áreas de ponderação, Ribeirão Preto, 2010.

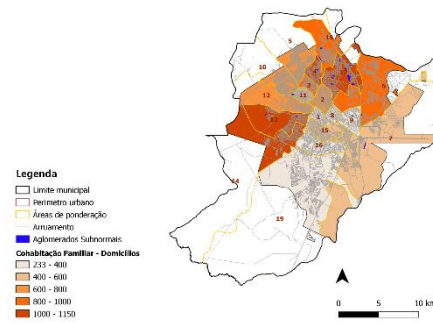


Fig. 4 Coabitação Familiar segundo áreas de ponderação, Ribeirão Preto, 2010.

Verifica-se uma clara demarcação espacial de Ribeirão Preto. É nos setores Norte e Sudoeste que se concentram os maiores valores absolutos e relativos da defasagem habitacional ribeirão-pretana. Conforme asseverado por Villaça (2012), o espaço urbano é um território de disputas. A localização dos grupos não é, portanto, casual; assim como não é fortuita sua periferização, no caso. Afinal, resta aos mais carentes as “sobras”, as áreas menos valorizadas do espaço intraurbano, reorientando, portanto, a distribuição espacial da população segundo vetores de expansão capitaneados pelos quesitos renda e valor da terra. Não se trata, pois, de opção, senão de constrangimento que direciona os mais pobres às áreas mais marginais e carentes. Assim, todo o arco Norte do município certamente é aquele que demanda maior empenho dos planejadores a fim de solucionar ou mitigar o déficit habitacional municipal

Como dito anteriormente, optou-se por uma redução dimensional do banco de dados. Antevia-se que a quantidade de variáveis seria um obstáculo ao cálculo da Rede Neural e seus diversos ciclos computacionais.

Realizou-se, portanto, um exercício experimental para tanto. Num primeiro momento, o banco foi reduzido a partir do Índice de Correlação de Pearson. A **Tabela 1** e a **Figura 5** ilustram tanto as variáveis selecionadas para permanecer no banco, como a correlação entre variáveis e, logo, a redundância delas no banco de dados.

Tabela 1 Seleção de variáveis segundo Índice de Correlação de Pearson

| Variáveis Perfeitas | Variáveis Excluídas |
|---------------------|---------------------|
| AP0020 | AP0001 |
| AP0012 | AP0002 |
| AP0009 | AP0003 |
| AP0004 | AP0004 |
| AP0007 | AP0005 |
| AP0014 | AP0006 |
| | AP0008 |
| | AP0010 |
| | AP0011 |
| | AP0013 |
| | AP0015 |
| | AP0016 |
| | AP0019 |
| | AP0062 |
| | VAR0004 |

Fonte: Elaboração própria.

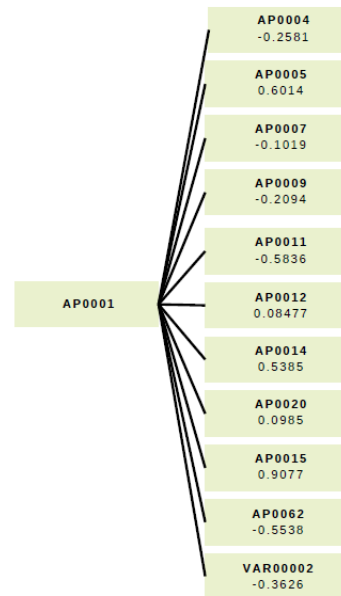


Figura 5 – Exemplo de correlações entre variáveis segundo Índice de Correlação de Pearson. Fonte: Elaboração própria

Porém, em última instância, a seleção através do ICP era uma tarefa subjetiva, além de ignorar relações não lineares entre os pares de variáveis. Nesse sentido, tendo em vista o mesmo objetivo, qual seja, redução dimensional do banco de dados, aplicou-se a técnica de Análise de Componentes Principais (APC). A **Figura 6** permite compreender a correlação entre as variáveis através das semelhanças ou diferenças em suas direções, já a **Figura 7**, a quantidade de dimensões com as quais é possível descrever o mesmo conjunto de dados.

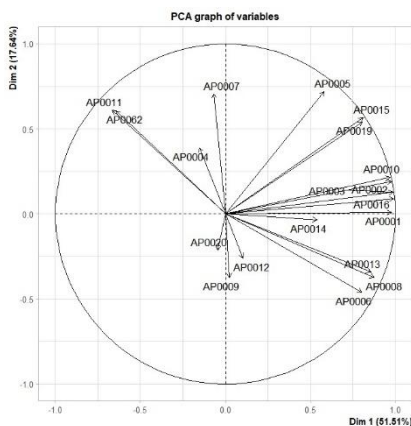


Fig. 6 Direcionalidade e correspondência entre variáveis segundo Análise de Componentes Principais. Fonte: Elaboração própria.

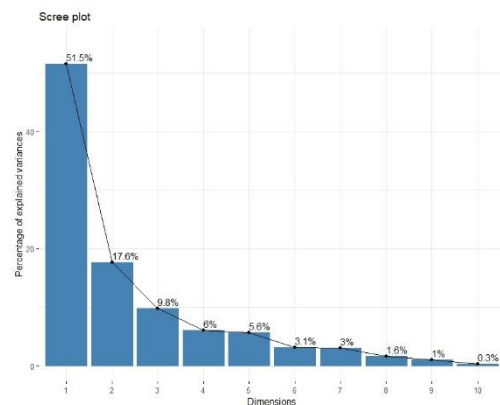


Fig. 7 Representatividade do banco de dados segundo a quantidade de dimensões aplicando Análise de Componentes Principais. Fonte: Elaboração própria.

Pode-se interpretar **Figura 6** da seguinte maneira: quando dois vetores estão próximos (formando um pequeno ângulo), as duas variáveis que eles representam estão positivamente correlacionadas. Quando eles se encontram a 90°, é improvável que se correlacionem. Quando divergem (formando um grande ângulo próximo de 180°), são, por sua vez, inversamente correlacionadas. Em contrapartida, a magnitude indica a importância de uma variável específica, dentre as outras, para a dimensão em questão. Ou seja, quanto maior a magnitude da seta, maior é sua relevância.

Optou-se então pelo uso de 5 dimensões. Elas explicam aproximadamente 85% (85,2%) dos casos segundo suas variâncias, uma parcela significativamente maior do que através do ICP. Nota-se, a partir da quinta dimensão, uma certa estabilização na tendência de cobertura de novos casos. É dizer, adicionar uma nova dimensão ao modelo não resulta necessariamente em grandes ganhos analíticos. Ao contrário, incorporá-las ao modelo apenas dificultaria a aplicação das Redes Neurais às estimativas do déficit por áreas de ponderação. O gráfico também permite constatar a importância de cada dimensão isoladamente. Assim, a primeira dimensão responde por aproximadamente 51% dos casos, enquanto a quinta não ultrapassa 4,5% deles.

Da mesma maneira como cada dimensão corresponde a um peso específico para explicar o conjunto de casos, cada variável apresenta uma importância relativa para se compreender a própria dimensão (**Tabela 2**).

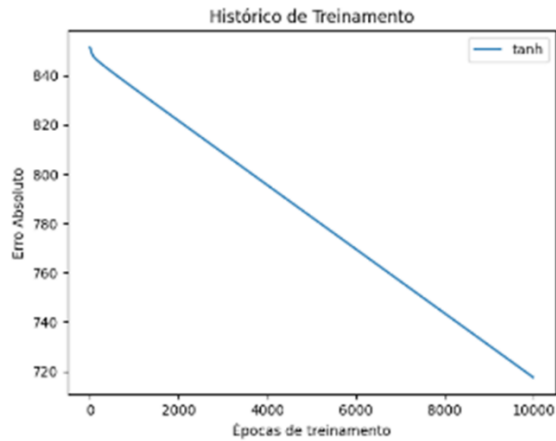
Tabela 2 Ponderações das variáveis segundo dimensões do banco de dados, Análise de Componentes Principais.

| | DIM 1 | DIM 2 | DIM 3 | DIM 4 | DIM 5 |
|--------|--------------------|------------|------------|------------|-------|
| AP0001 | 9,7807989 | | | | |
| AP0002 | 9,7742469 | | | | |
| AP0003 | 9,9556998 | | | | |
| AP0004 | | | 29,7798000 | | |
| AP0005 | | 15,4323406 | | | |
| AP0006 | | 6,3518419 | | | |
| AP0007 | | 14,8596135 | | | |
| AP0008 | | 4,1610059 | | | |
| AP0009 | | | | 28,7444812 | |
| AP0010 | 9591814459,0000000 | | | | |
| AP0011 | | 11,1284465 | | | |
| AP0012 | | | | 39,7259711 | |
| AP0013 | 7,4903769 | | | | |
| AP0014 | | | 13,2070599 | | |
| AP0015 | | 9,5081271 | | | |
| AP0016 | 10,0185619 | | | | |
| AP0019 | | 8,7527861 | | | |
| AP0020 | | | 29,8837165 | | |
| AP0062 | | 11,0227405 | | | |

Fonte: Elaboração própria.

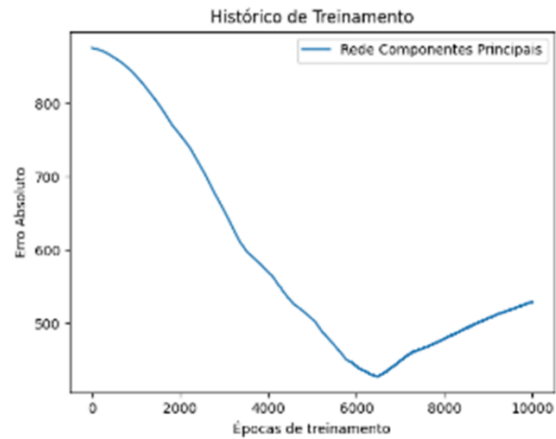
Assim, é possível compreender a importância relativa de cada variável em explicar a dimensão da qual faz parte. Por exemplo, o total de rendimentos nominal mensal dos domicílios improvisados corresponde a quase 40% do valor explicativo da dimensão 4.

Já as **Figuras 8 e 9** elucidam os ciclos de treinamento necessários, à Rede Neural, para atingir os menores níveis de erros nas estimativas do déficit habitacional.



Menor Erro Absoluto Médio: 710

Fig. 8 Erro absoluto segundo épocas de treinamento da Rede Neural, 19 variáveis. Fonte: Elaboração própria.



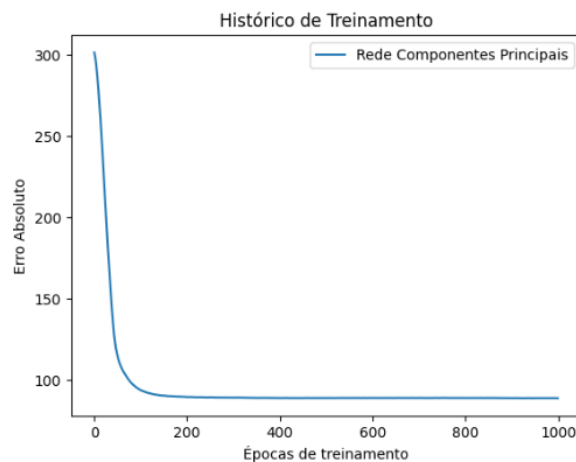
Menor Erro Absoluto Médio: 433

Overfitting após 6500 épocas

Fig. 9 Erro absoluto segundo épocas de treinamento da Rede Neural, 5 dimensões obtidas pela Análise de Componentes Principais. Fonte: Elaboração própria.

No primeiro banco de dados, com 19 variáveis, foram necessários quase 10.000 ciclos e, ainda assim, o menor erro médio absoluto foi de 710 casos, a mais ou a menos, por área de ponderação. Em outras palavras, uma área de ponderação poderia ter 0 domicílios precários ou 710 deles, uma variação considerável. Ao reduzir-se o banco de dados a 5 dimensões através da ACP, a quantidade de ciclos necessários é consideravelmente menor, 6.500. Embora o erro médio absoluto tenha diminuído, 433, ele persiste elevado. Ainda assim, a decisão em sintetizar o banco de dados por meio da Análise de Componentes Principais definitivamente se mostrou acertada.

Apenas como teste experimental, aplicou-se os mesmos procedimentos agora não mais apenas às áreas de ponderação ribeirão-pretanas (19), senão a todas às áreas de ponderação do Estado de São Paulo, 903, no caso (**Figura 10**).



Menor Erro Absoluto Médio: 89

Figura 10 – Erro absoluto segundo épocas de treinamento da Rede Neural, 5 dimensões, 903 áreas de ponderação, Estado de São Paulo. Fonte. Elaboração própria.

Presumia-se que, ao aumentar o número N de casos do banco de dados, a Rede Neural aperfeiçoaria o cálculo das estimativas do déficit habitacional. E, de fato, é o que constata-se. Ao se analisar 903 casos (áreas de ponderação), contra os 19 casos de Ribeirão Preto, em pouco menos de 100 épocas de treinamento a Rede Neural atinge um erro médio absoluto de 89 casos. Muito mais acurada do que anteriormente.

5 CONCLUSÕES

O presente artigo se esforçou por estimar o déficit habitacional segundo áreas de ponderação de Ribeirão Preto como uma primeira etapa para tentar inferi-lo para recortes territoriais ainda menores, como é o caso dos setores censitários. Buscou-se, portanto, esclarecer todas as estratégias empregadas para tanto, bem como as frustrações em valer-se de algumas delas. Embora o erro absoluto ainda seja considerável, não é de se menosprezar que a Rede Neural foi capaz de projetar o déficit habitacional segundo informações que lhe mantêm pouca relação, tais como: total de residentes, de domicílios particulares permanente, de residentes segundo sexo, raça e cor. Nenhuma delas, a priori, é base para calculá-lo de modo direto. Parece, então, essa uma alternativa à estimativa do déficit habitacional em localidades, ou mesmo países, onde dados sobre o assunto inexistem. Por fim, é preciso ponderar que o déficit, por mais que infelizmente seja uma manifestação recorrente da nossa desigualdade, trata-se de um evento com um N diminuto de casos. Os 89 casos de erro segundo as áreas de ponderação paulistas soam razoavelmente aceitáveis segundo essa perspectiva. Acredita-se que seja necessário otimizar a qualidade das próprias variáveis para aperfeiçoar as estimativas ainda mais. Talvez uma solução fosse a microssimulação espacial de dados.

6 BIBLIOGRAFIA

Bharadiya, J. P. (2023). A tutorial on principal component analysis for dimensionality reduction in machine learning. **International Journal of Innovative Science and Research Technology**, 8(5), 2028-2032.

Boreland, B., Kunze, H., e Levere, K. M. (2023). 10 Artificial Neural Networks. **Engineering Mathematics and Artificial Intelligence: Foundations, Methods, and Applications**, 227.

Goh, Bee-Hua. (1998). Forecasting residential construction demand in Singapore: a comparative study of the accuracy of time series, regression and artificial neural network techniques. **Engineering, Construction and Architectural Management**, 5(3), 261-275.

de Miranda-Ribeiro, A., Viana, Raquel de Matos, e de Azevedo, S. (2015). Déficit habitacional municipal em Minas Gerais. **Caderno de Geografia**, 25(43), 144-162.

_____; Viana, Raquel de Matos; Salis, Raíza Maciel (2013). Déficit Habitacional no Brasil em 2007 e 2008: notas metodológicas e principais resultados. **Revista Geografias**, v. 9, n. 1, p. 97-115.

Zainun, N. Y. B., Rahman, I. A., e Eftekhari, M. (2010). Forecasting low-cost housing demand in Johor Bahru, Malaysia using artificial neural networks (ANN). **Journal of Mathematics Research**, 2(1), 14.