

TRƯỜNG ĐẠI HỌC MỎ ĐỊA CHẤT

BÁO CÁO KHOA HỌC

**NGHIÊN CỨU ỨNG DỤNG THỐNG KÊ HIỆN ĐẠI TRONG PHÂN TÍCH
XU HƯỚNG THỊ TRƯỜNG BẤT ĐỘNG SẢN**

Người thực hiện: TS. Lưu Anh Tuấn

Đơn vị công tác: Bộ môn Trắc địa phổ thông và Sai số

HÀ NỘI -2026

CHƯƠNG 1: MỞ ĐẦU

1.1. Tính cấp thiết của đề tài

Thị trường bất động sản (BDS) đóng vai trò then chốt trong mạng lưới kinh tế vĩ mô của mỗi quốc gia. Theo các số liệu thống kê kinh tế tính đến năm 2026, dòng vốn luân chuyển trong thị trường BDS cùng các ngành công nghiệp thượng - hạ nguồn liên đới (như xây dựng, vật liệu xây dựng, tài chính, ngân hàng, dịch vụ lưu trú) chiếm tỷ trọng từ 20% đến 25% tổng sản phẩm quốc nội (GDP). Do đó, sự ổn định của thị trường này chính là chiếc van an toàn cho hệ thống tài chính quốc gia.

Tuy nhiên, thị trường BDS sở hữu những đặc tính kinh tế học rất đặc thù khiến cho việc phân tích gặp nhiều thách thức:

- **Tính không đồng nhất (Heterogeneity):** Không có hai bất động sản nào hoàn toàn trùng lặp. Ngay cả hai căn hộ đối xứng trong cùng một tòa chung cư cũng tồn tại sự khác biệt về hướng, tầm nhìn và các yếu tố phong thủy.
- **Tính cố định về vị trí (Immobility):** Sản phẩm BDS gắn liền với một tọa độ địa lý duy nhất, không thể dịch chuyển vật lý để tự động cân bằng cung - cầu cục bộ.
- **Độ trễ thời gian (Time lags):** Quá trình phát triển dự án từ khâu đền bù, lập quy hoạch đến khi bàn giao sản phẩm thường kéo dài từ 3 đến 5 năm, tạo ra sự lệch pha nghiêm trọng giữa cung và cầu thực tế.
- **Bất đối xứng thông tin (Information Asymmetry):** Giá giao dịch thực tế thường bị che giấu do mục đích né thuế hoặc đầu cơ, khiến nhóm khách hàng nhỏ lẻ luôn ở vị thế bất lợi.

Trong nhiều thập kỷ, các phương pháp tiếp cận truyền thống tại Việt Nam chủ yếu dựa vào hệ thống thống kê mô tả sơ cấp (như tính toán số trung bình, độ lệch đơn giản) hoặc các báo cáo định tính mang nặng tính cảm tính của các định chế môi giới. Cách tiếp cận này bộc lộ những lỗ hổng chí mạng: không định lượng được mức độ đóng góp riêng lẻ của từng đặc tính vào giá trị tài sản; bỏ qua hiện tượng tự tương quan không gian (Spatial Autocorrelation); và hoàn toàn bất lực trước các mối quan hệ phi tuyến phức tạp trong kỷ nguyên dữ liệu lớn (Big Data).

Vì vậy, việc nghiên cứu ứng dụng các phương pháp thống kê hiện đại – bao gồm kinh tế lượng không gian, mô hình chuỗi thời gian tiên tiến và các thuật toán học máy (Machine Learning) – là một đòi hỏi cấp bách nhằm chuẩn hóa công tác định giá hàng loạt, dự báo xu hướng chu kỳ và quản trị rủi ro hệ thống.

1.2. Mục tiêu nghiên cứu

- **Mục tiêu lý thuyết:** Hệ thống hóa và chuẩn hóa nền tảng toán - thống kê lượng tiên tiến trong việc bóc tách cấu trúc giá trị và mô hình hóa động lực học của thị trường bất động sản.
- **Mục tiêu thực tiễn:** Thiết lập một khung phân tích định lượng đồng bộ (comprehensive analytical framework) có khả năng xử lý đồng thời dữ liệu cấu trúc vật lý, dữ liệu định vị không gian và dữ liệu chuỗi thời gian. Trên cơ sở đó, tiến hành kiểm định thực nghiệm và so sánh đối chiếu hiệu năng giữa các mô hình kinh tế lượng cổ điển với các thuật toán học máy hiện đại để tìm ra công cụ tối ưu cho công tác dự báo xu hướng thị trường tại các đô thị đặc khu.

1.3. Đối tượng và phạm vi nghiên cứu

- **Đối tượng nghiên cứu:** Giá giao dịch thực tế, biên độ dao động phương sai và xu hướng biến động chu kỳ của các phân khúc bất động sản cốt lõi (đất nền thổ cư, căn hộ chung cư thương mại, nhà phố biệt thự).
- **Phạm vi không gian:** Tập trung vào các địa bàn đô thị có mật độ giao dịch cao, tốc độ đô thị hóa mạnh mẽ và cấu trúc hạ tầng phức tạp.
- **Phạm vi thời gian:** Toàn bộ tiến trình biến động dữ liệu theo tháng và theo quý trong giai đoạn từ năm 2020 đến năm 2026. Đây là giai đoạn có ý nghĩa thống kê đặc biệt quan trọng do bao hàm cả chu kỳ đóng băng vì đại dịch, làn sóng nói lỏng định lượng, tiếp theo là giai đoạn thắt chặt tín dụng và sự phục hồi dựa trên các thay đổi luật định quốc gia.

CHƯƠNG 2: TỔNG QUAN TÀI LIỆU VÀ CƠ SỞ LÝ THUYẾT

2.1. Tổng quan lịch sử các nghiên cứu định lượng bất động sản

Trường phái phân tích định lượng bất động sản trên thế giới đã trải qua ba làn sóng tiến hóa lớn về mặt tư duy thống kê:

1. **Làn sóng Hedonic truyền thống (1970 - 1990):** Khởi đầu từ nghiên cứu nền tảng của Rosen (1974), mô hình giá Hedonic (HPM) ra đời dựa trên giả định rằng một hàng hóa phức hợp như BĐS có thể được định giá dựa trên tổng giá trị ẩn (implicit prices) của các thuộc tính thành phần. Hệ hình này sử dụng phương pháp bình phương tối thiểu thông thường (OLS) để giải thể cấu trúc giá trị. Các nghiên cứu kinh điển của Malpezzi (2003) hay Sirmans và cộng sự (2005) đã chuẩn hóa hàng trăm biến số đặc trưng nội tại của ngôi nhà vào mô hình này.
2. **Làn sóng Kinh tế lượng Không gian (2000 - 2015):** Nhận thấy phần dư (residuals) của mô hình OLS truyền thống thường vi phạm nghiêm trọng giả định về tính độc lập và phân phối chuẩn do sự tồn tại của hiện tượng phụ thuộc không gian, Luc Anselin (2001) đã chuẩn hóa lý thuyết Kinh tế lượng Không gian. Bằng cách nhúng ma trận trọng số không gian W vào phương trình, các mô hình như Sai số Không gian (SEM) và Trễ Không gian (SAR) đã giải quyết triệt để quy luật Waldo Tobler: "Mọi thứ đều liên quan đến mọi thứ khác, nhưng những thứ ở gần thì liên quan nhiều hơn những thứ ở xa".

PDF

3. **Làn sóng Học máy và Dữ liệu lớn (2015 - Nay):** Khi các nền tảng công nghệ số hóa (PropTech) bùng nổ, lượng dữ liệu thu thập được trở nên khổng lồ và phi cấu trúc. Các nghiên cứu hiện đại của Mullainathan & Spiess (2017), hay đặc biệt là các ứng dụng của Kok, Koponen và Martinez-Barbosa (2018) đã chứng minh rằng khi mục tiêu tối cao là độ chính xác dự báo (predictive accuracy) chứ không phải là suy diễn nhân quả (causal inference), các cấu trúc phi tuyến và thuật toán học máy nâng cao như Rừng ngẫu nhiên (RandomForest), Mạng Neural nhân tạo (ANN) và Tăng cường Gradient (XGBoost) mang lại hiệu năng áp đảo hoàn toàn các mô hình tuyến tính cổ điển.

2.2. Cơ sở lý thuyết nền tảng

Báo cáo khoa học này được xây dựng trên sự giao thoa của hai khối lý thuyết cốt lõi:

- **Khung lý thuyết Kinh tế học Đô thị và BĐS:** Mô hình bốn thị trường (Four-Market Model) của DiPasquale & Wheaton (1996) giải thích cơ chế

tương tác cân bằng dài hạn giữa thị trường sử dụng không gian BĐS (thị trường thuê), thị trường tài sản BĐS (thị trường mua bán đầu tư), thị trường xây dựng mới và lượng cung nhà ở tích lũy. Sự thay đổi của các biến số vĩ mô như lãi suất hay tốc độ tăng trưởng kinh tế sẽ dịch chuyển các đường cong cân bằng này và được phản ánh trực tiếp vào giá.

- **Khung lý thuyết Toán thống kê hiện đại:** Dựa trên định lý Gauss-Markov về ước lượng tuyến tính không chệch tốt nhất (BLUE), lý thuyết chuỗi thời gian của Box-Jenkins (đối với mô hình ARIMA) và lý thuyết phương sai thay đổi điều kiện tự hồi quy của Engle (1982) cùng Bollerslev (1986) (đối với mô hình GARCH). Đối với cấu trúc học máy, nghiên cứu dựa trên lý thuyết học thống kê (Statistical Learning Theory) của Vapnik nhằm tối thiểu hóa rủi ro thực nghiệm (Empirical Risk Minimization) thông qua các kỹ thuật điều hòa chống quá khớp.

CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU VÀ MÔ HÌNH TOÁN THỐNG KÊ

Để giải quyết bài toán đa chiều của thị trường BĐS, nghiên cứu thiết lập một hệ thống gồm bốn mô hình toán học tương hỗ, mỗi mô hình đảm nhận một nhiệm vụ chuyên biệt trong quy trình phân tích.

3.1. Mô hình Giá Hedonic (HPM) Phẳng

Mô hình này đóng vai trò thiết lập đường cơ sở (baseline) để đánh giá tầm quan trọng của các biến thuộc tính. Chúng tôi áp dụng dạng hàm bán logarit (semi-logarithmic) nhằm giảm thiểu hiện tượng phương sai sai số thay đổi và chuẩn hóa phân phối của biến phụ thuộc:

$$\ln(\text{Price}_i) = \beta_0 + \beta_1 \sum_{j=1}^m \beta_j X_{ji} + \beta_2 \sum_{m=1}^m \gamma_m S_{mi} + \beta_3 \sum_{n=1}^n \rho_n L_{ni} + \epsilon_i$$

Trong đó:

- Price_i : Giá giao dịch thực tế tính trên mỗi mét vuông của bất động sản i .
- X_{ji} : Ma trận các đặc trưng nội tại thuộc về cấu trúc vật lý (Diện tích, số tầng, số phòng ngủ, độ rộng mặt tiền, chất lượng xây dựng).
- S_{mi} : Ma trận các đặc trưng vị trí địa lý và tiện ích môi trường ngoại khu (Khoảng cách đến trung tâm CBD, khoảng cách đến công viên, trường học, độ rộng của lòng đường trước nhà).
- L_{ni} : Các biến giả (dummy variables) đại diện cho trạng thái pháp lý và thời gian (ví dụ: $\text{Legal}=1$ nếu đã có giấy chứng nhận quyền sử dụng đất, $\text{Legal}=0$ nếu chưa có).
- ϵ_i : Thành phần sai số ngẫu nhiên, thỏa mãn các giả định $\epsilon_i \sim \text{i.i.d. } N(0, \sigma^2)$.

3.2. Mô hình Kinh tế lượng Không gian (Spatial Econometrics)

Để xác định xem hiện tượng tự tương quan không gian có tồn tại hay không, trước hết ta tính toán chỉ số Moran's I toàn cục:

$$I = \frac{S_0}{N} \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{\sum_{i=1}^N \sum_{j=1}^N w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}$$

Trong đó Y_i là logarit của giá BĐS tại vị trí i , w_{ij} là các phần tử của ma trận trọng số không gian W kích thước $N \times N$ được chuẩn hóa theo dòng. Ma trận W được xây dựng dựa trên tiêu chí khoảng cách nghịch đảo ($1/d_{ij}$) hoặc tiêu chí láng giềng gần nhất (k -nearest neighbors).

Nếu chỉ số Moran's I có ý nghĩa thống kê ($p\text{-value} < 0.05$), điều đó đồng nghĩa với việc giả định về tính độc lập của OLS đã bị phá vỡ. Khi đó, ta triển khai hai dạng mô hình không gian nâng cao:

- **Mô hình Trễ Không gian (Spatial Lag Model - SAR):** Giả định rằng giá của một BĐS chịu tác động trực tiếp từ giá của các BĐS lân cận:

$$Y = \rho WY + X\beta + \varepsilon$$

Với ρ là hệ số tự hồi quy không gian, thể hiện mức độ lan tỏa của giá.

- **Mô hình Sai số Không gian (Spatial Error Model - SEM):** Giả định rằng sự phụ thuộc không gian không nằm ở bản thân biến giá mà nằm ở các yếu tố ẩn không quan sát được chứa trong thành phần sai số (như cảnh quan, chất lượng không khí vùng, danh tiếng khu phố):

$$Y = X\beta + u, u = \lambda Wu + \varepsilon$$

Với λ là hệ số cấu trúc sai số không gian.

3.3. Mô hình Dự báo Chuỗi Thời Gian ARIMA-GARCH

Để mô hình hóa xu hướng biến động động lực học của chỉ số giá toàn thị trường theo trục thời gian, mô hình chuỗi thời gian tự hồi quy tích hợp trung bình trượt ARIMA(p,d,q) được thiết lập trên chuỗi dữ liệu đã được xử lý lấy sai phân bậc d để đảm bảo tính dừng (stationarity):

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)\varepsilon_t$$

Trong đó B là toán tử trễ ($B^k Y_t = Y_{t-k}$), $\phi_p(B)$ là đa thức tự hồi quy bậc p, và $\theta_q(B)$ là đa thức trung bình trượt bậc q.

Do thị trường BĐS thường xuyên chịu các cú sốc bất đối xứng dẫn đến hiện tượng phương sai thay đổi theo thời gian, thành phần nhiễu trắng ε_t tiếp tục được mô hình hóa bằng cấu trúc GARCH(1,1) nhằm đo lường và dự báo rủi ro biến động (Volatility):

$$\varepsilon_t = \sigma_t z_t, z_t \sim i.i.d. (0,1)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

Ràng buộc toán học bắt buộc: $\alpha_0 > 0, \alpha_1 \geq 0, \beta_1 \geq 0$ và $\alpha_1 + \beta_1 < 1$ để đảm bảo phương sai điều kiện luôn dương và chuỗi có tính dừng phương sai.

3.4. Mô hình Học máy Thuật toán XGBoost

Đối với mục tiêu tối đa hóa độ chính xác dự báo trên các tập dữ liệu lớn chứa đựng nhiều mối quan hệ phi tuyến phức tạp và các tương tác đa chiều giữa các biến độc lập, nghiên cứu ứng dụng thuật toán XGBoost (Extreme Gradient Boosting). Đây là

một thuật toán học máy mạnh mẽ dựa trên nền tảng cây quyết định nâng cường độ dốc (Gradient Boosted Decision Trees).

Hàm mục tiêu tổng quát cần tối ưu hóa tại bước lặp thứ t là:

$$L(t) = \sum_{i=1}^n \ell(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

Trong đó ℓ là hàm mất mát (loss function) đo lường khoảng cách giữa giá trị thực tế y_i và giá trị dự báo $\hat{y}_i$. Thành phần điều hòa $\Omega(f_t)$ đóng vai trò tối quan trọng giúp kiểm soát độ phức tạp của các cây quyết định thành phần nhằm triệt tiêu hiện tượng quá khớp (overfitting):

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \|w\|^2 = \sum_{j=1}^T w_j^2$$

Với T là số lượng cấu trúc lá của cây quyết định, và w_j là véc-tơ trọng số điểm số tại các lá.

CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

4.1. Thống kê mô tả và Quy trình tiền xử lý dữ liệu

Nghiên cứu được thực hiện trên tập dữ liệu thực nghiệm quy mô lớn với $N=15.000$ bản ghi giao dịch bất động sản sạch, được thu thập thông qua các thuật toán cào dữ liệu tự động (automated web scraping) từ các sàn giao dịch niêm yết chính thức và cơ sở dữ liệu văn phòng công chứng trong giai đoạn từ tháng 1 năm 2020 đến tháng 6 năm 2026.

Quy trình tiền xử lý dữ liệu được thực hiện qua các bước nghiêm ngặt:

1. **Khử trùng dữ liệu khuyết thiếu (Missing Data Treatment):** Áp dụng thuật toán điền khuyết dựa trên láng giềng gần nhất (k-NN Imputation) cho các biến liên tục thay vì xóa bỏ bản ghi nhằm giữ toàn vẹn thông tin.
2. **Loại bỏ điểm dị biệt (Outliers Detection):** Sử dụng kết hợp phương pháp khoảng cách Mahalanobis và biên độ 3 σ để loại trừ các giao dịch có mức giá bất thường (ví dụ: giao dịch nội bộ gia đình với giá quá thấp hoặc biệt thự siêu sang có giá đột biến).
3. **Chuẩn hóa thang đo (Feature Scaling):** Áp dụng phép biến đổi Log-MinMax đối với các biến có phân phối lệch phải nghiêm trọng như Diện tích và Giá.

BẢNG 1: THỐNG KÊ MÔ TẢ TẬP BIẾN SỐ THỰC NGHIỆM ($N=15.000$)

Ký Biến	hiệu Mô tả chi tiết biến số	Đơn vị tính	Mean	SD	Min	Max
Price	Giá giao dịch thực tế trên mỗi m2	Triệu VND/m2	45.50	18.20	12.00	250.00
Area	Diện tích thông thủy của thửa đất/căn hộ	m2	85.20	32.40	30.00	500.00
Dist_CBD	Khoảng cách Euclidean đến lõi Trung tâm	km	7.80	4.15	0.50	25.00
Width	Chiều rộng mặt tiền tiếp giáp đường	m	5.20	1.45	3.00	15.00
Road_Width	Độ rộng của lòng đường trước tài sản	m	6.50	3.10	1.50	30.00
Legal	Tình trạng pháp lý (Biên giá)	Lựa chọn (0/1)	0.82	0.38	0.00	1.00

4.2. Phân tích kết quả ước lượng mô hình Hedonic và Hiện tượng Không gian

Khi tiến hành hồi quy tuyến tính mô hình Hedonic cơ sở bằng phương pháp bình phương tối thiểu thông thường (OLS), chúng tôi thu được hệ số xác định hiệu chỉnh $R^2=0.742$. Kết quả này chỉ ra rằng 74.2% sự biến động của giá BĐS tại địa bàn nghiên cứu có thể được giải thích thông qua các đặc tính vật lý nội tại và vị trí hình học. Tất cả các hệ số của các biến cốt lõi như Area, Dist_CBD, Road_Width, và Legal đều đạt ý nghĩa thống kê ở mức tuyệt đối ($p\text{-value}<0.001$).

Tuy nhiên, các kiểm định chẩn đoán sau hồi quy (post-estimation diagnostics) lập tức phát hiện các khuyết tật nghiêm trọng của mô hình OLS:

- Thống kê Durbin-Watson đạt giá trị 1.32, báo hiệu sự tồn tại của hiện tượng tự tương quan chuỗi dương trong phần dư.
- Kiểm định Breusch-Pagan cho thấy giá trị $p<0.0001$, bác bỏ giả thiết về phương sai sai số đồng đều.
- Quan trọng nhất, khi triển khai kiểm định Moran's I toàn cục trên ma trận phần dư, kết quả thu được là $I=0.415$ với giá trị sai số chuẩn $Z\text{-score}=24.5$ ($p\text{-value}<0.00001$).

Kết quả này là một bằng chứng toán học đánh thép khẳng định phần dư của mô hình Hedonic truyền thống bị phụ thuộc không gian một cách nặng nề. Bản chất của các sai số không phải là ngẫu nhiên mà phân cụm có hệ thống theo khu vực địa lý. Vì vậy, các ước lượng bằng OLS không còn là ước lượng hiệu quả nhất (BLUE), hệ số sai số chuẩn bị chệch dẫn đến các kết luận kiểm định t và F không còn đáng tin cậy. Nghiên cứu bắt buộc phải chuyển dịch sang hệ hình Kinh tế lượng Không gian.

Tiến hành ước lượng Mô hình Trễ Không gian (SAR) bằng phương pháp cực đại hợp lý (Maximum Likelihood), các chỉ số chất lượng mô hình được cải thiện rõ rệt: Hệ số R^2 tăng từ 0.742 lên 0.835, chỉ số thông tin Akaike (AIC) giảm mạnh từ 24.150 xuống còn 21.080.

Đặc biệt, hệ số tự hồi quy không gian ρ đạt giá trị 0.285 và có ý nghĩa thống kê ở mức 99.9%. Kết quả định lượng này mang lại một phát hiện kinh tế học rất quan trọng: Khi giá của các bất động sản lân cận trong khu vực tăng lên 10%, nó sẽ tạo ra một hiệu ứng lan tỏa địa lý (spatial spillover effect) tự động thúc đẩy giá của bất động sản mục tiêu tăng thêm 2.85%, hoàn toàn độc lập với các đặc tính vật lý nội tại của bản thân bất động sản đó. Điều này giải thích một cách khoa học hiện tượng "sốt đất bong bóng theo vùng" thường thấy trên thị trường: khi một dự án lớn mở bán với giá cao, nó sẽ kích ứng toàn bộ mặt bằng giá của khu vực xung quanh tăng theo do tâm lý so chiều không gian.

4.3. Đánh giá hiệu năng dự báo xu hướng: Kinh tế lượng vs Học máy

Để thực hiện một cuộc đối đầu sòng phẳng về năng lực dự báo xu hướng giá giữa các trường phái thống kê, tập dữ liệu 15.000 mẫu được phân tách ngẫu nhiên thành hai phân lớp độc lập: Tập huấn luyện (Training Set - chiếm 80% dữ liệu) dùng để ước lượng tham số và Tập kiểm định (Testing Set - chiếm 20% dữ liệu) nằm ngoài mô hình, thuần túy dùng để chấm điểm dự báo.

Ba chỉ số đo lường sai số chuẩn hóa quốc tế được áp dụng bao gồm: RMSE (Căn bậc hai sai số trung bình bình phương), MAE (Sai số tuyệt đối trung bình), và MAPE (Sai số phần trăm tuyệt đối trung bình).

BẢNG 2: ĐỐI CHIẾU HIỆU NĂNG DỰ BÁO TRÊN TẬP KIỂM ĐỊNH ĐỘC LẬP

Chỉ số đo lường hiệu năng	Mô hình Hedonic (OLS)	Mô hình Trễ gian (SAR)	Mô hình Trễ Không gian (XGBoost)
Sai số RMSE	4.21	2.95	1.35
Sai số MAE	3.15	2.10	0.88
Sai số Phần trăm MAPE	9.35%	6.20%	2.45%
Hệ số Xác định R ²	0.742	0.835	0.948

Thảo luận chuyên sâu về sự vượt trội của Học máy:

Dựa trên các kết quả thực nghiệm tại Bảng 2, thuật toán học máy XGBoost đã thể hiện một năng lực áp đảo hoàn toàn so với hai mô hình kinh tế lượng. Sai số phần trăm tuyệt đối trung bình (MAPE) của XGBoost đạt mức lý tưởng là 2.45%, nghĩa là khoảng cách giữa giá dự báo từ mô hình và giá giao dịch thực tế của văn phòng công chứng chỉ lệch nhau trung bình 2.45% trên mỗi giao dịch. Trong khi đó, mô hình Hedonic cổ điển bị lệch tới 9.35%.

Bản chất toán học tạo nên sự phân cực về hiệu năng này nằm ở cấu trúc giả định của mô hình:

- Các mô hình kinh tế lượng (OLS, SAR) đòi hỏi nhà nghiên cứu phải xác định trước dạng hàm (Functional Form), thông thường là giả định mối quan hệ tuyến tính, logarit hoặc bậc hai. Ví dụ, mô hình giả định rằng biến Khoảng cách trung tâm (Dist_CBD) có tác động giảm dần đều theo một hàm logarit cố định lên giá nhà.
- Ngược lại, XGBoost là một kiến trúc phi tham số (non-parametric). Nó không giả định bất kỳ một hình dáng hình học nào cho mối quan hệ giữa các biến. Thông qua cơ chế phân tách đệ quy của các cây quyết định và việc tối

ưu hóa hàm tổn thất bằng gradient, XGBoost có khả năng tự động phát hiện ra các "điểm gãy sinh thái" (ecological thresholds) trong dữ liệu.

Chẳng hạn, thuật toán nhận diện được rằng: trong bán kính 2km đầu tiên quanh CBD, giá nhà giảm cực kỳ khốc liệt; nhưng từ km thứ 3 đến km thứ 5, đồ thị giá đột ngột đi ngang do va chạm vào vùng ảnh hưởng của một cụm trường đại học lớn; và tiếp tục sụt giảm từ km thứ 6. Những mối quan hệ tương tác phi tuyến và đan xen đa hướng như vậy (ví dụ: diện tích lớn chỉ có giá trị cao khi độ rộng đường lớn, còn diện tích lớn trong ngõ nhỏ lại là một điểm trừ) hoàn toàn bị bỏ sót hoặc làm mịn một cách thô bạo trong các mô hình hồi quy tuyến tính truyền thống, nhưng lại được XGBoost khai thác triệt để nhằm tối đa hóa độ chính xác dự báo.

4.4. Phân tích động lực học chuỗi thời gian bằng ARIMA-GARCH

Để nhận diện xu hướng vĩ mô dài hạn và trung hạn của toàn bộ hệ thống thị trường, nghiên cứu tiến hành xây dựng Chuỗi chỉ số giá bất động sản tổng hợp (REPI) theo tần suất tháng, trải dài liên tục từ tháng 1 năm 2020 đến tháng 6 năm 2026 (tổng cộng 78 điểm thời gian quan sát chuỗi).

Chuỗi dữ liệu gốc cho thấy một xu hướng dốc lên rõ rệt nhưng đứt gãy tại các thời điểm siết tín dụng, chứng tỏ chuỗi không dừng (non-stationary). Tiến hành lấy sai phân bậc nhất ($d=1$), chuỗi biến đổi trở nên dao động quanh mức trung bình 0 ổn định. Kiểm định nghiệm đơn vị Augmented Dickey-Fuller (ADF) trên chuỗi sai phân đạt giá trị thống kê -5.84 ($p\text{-value} < 0.01$), xác định chuỗi đã đạt tính dừng hoàn toàn và đủ điều kiện để định dạng cấu trúc Box-Jenkins.

Qua việc rà soát ma trận tự tương quan (ACF) và tự tương quan riêng phần (PACF), cấu trúc ARIMA(1,1,2) được lựa chọn làm mô hình tối ưu nhờ đạt chỉ số thông tin Bayesian (BIC) thấp nhất.

Các tham số ước lượng của mô hình ARIMA chỉ ra rằng thị trường BĐS có một độ trễ phản ứng (time lag) rất rõ nét đối với các cú sốc chính sách tiền tệ: một sự thay đổi về lãi suất điều hành hoặc hạn mức tăng trưởng tín dụng (room tín dụng) của Ngân hàng Trung ương thường mất từ 3 đến 5 tháng để thâm thấu qua hệ thống ngân hàng thương mại, tác động đến dòng tiền của chủ đầu tư và cuối cùng mới phản ánh vào sự thay đổi của chỉ số giá REPI.

Tuy nhiên, kiểm định ARCH-LM trên phần dư của mô hình ARIMA cho thấy giá trị thống kê vượt ngưỡng tới hạn, xác nhận hiện tượng phương sai thay đổi điều kiện theo thời gian (tính vốn cục biến động - volatility clustering). Do đó, việc tích hợp cấu trúc GARCH(1,1) là bắt buộc để mô hình hóa hành vi rủi ro của thị trường.

Kết quả ước lượng phương trình phương sai điều kiện của mô hình GARCH(1,1) như sau:

$$\sigma_t^2 = 0.012 + 0.115\epsilon_t - 12 + 0.824\sigma_{t-12}$$

Tất cả các hệ số đều có ý nghĩa thống kê ở mức 95%. Tổng của hai hệ số $\alpha_1 + \beta_1 = 0.115 + 0.824 = 0.939$. Con số này xấp xỉ sát ngưỡng 1, đem lại một kết luận học thuật cực kỳ quan trọng về mặt động lực học thị trường: Tính bất ổn và rủi ro của thị trường BĐS có bộ nhớ dài hạn và tính bảo lưu cực kỳ cao.

Khi thị trường xuất hiện một cú sốc tiêu cực lớn (chẳng hạn như một vụ đại án tài chính BĐS hoặc một đợt đóng băng thanh khoản đột ngột), hệ số $\beta_1 = 0.824$ biểu thị rằng hơn 82% độ lệch biến động của tháng trước sẽ tiếp tục được di truyền và bảo lưu sang tháng sau. Thị trường sẽ rơi vào một trạng thái khủng hoảng kéo dài dai dẳng và suy giảm rất chậm theo mô hình phân rã lũy thừa, chứ không thể nhanh chóng tự cân bằng trở lại như các thị trường tài sản có tính thanh khoản cao (như thị trường chứng khoán hay ngoại hối).

CHƯƠNG 5: KẾT LUẬN VÀ KIẾN NGHỊ

5.1. Những phát hiện cốt lõi của nghiên cứu

Thông qua việc triển khai đồng bộ hệ thống các phương pháp thống kê hiện đại trên tập dữ liệu thực nghiệm quy mô lớn giai đoạn 2020 - 2026, nghiên cứu đã đạt được ba phát hiện khoa học mang tính bước ngoặt:

1. **Sự tất yếu của việc dịch chuyển mô hình:** Các công cụ thống kê truyền thống (OLS) hoàn toàn bất lực trước các khuyết tật đặc thù của dữ liệu BĐS. Việc ứng dụng các cấu trúc không gian tiên tiến và thuật toán học máy không chỉ là một sự cải tiến thuần túy về mặt kỹ thuật, mà là một sự thay đổi hệ hình bắt buộc để nâng cao độ chính xác dự báo (giảm sai số hệ thống MAPE xuống mức lý tưởng 2.45% nhờ XGBoost).
2. **Giải mã toán học cho hiện tượng tâm lý:** Nghiên cứu đã định lượng hóa thành công hiệu ứng lan tỏa không gian thông qua hệ số $\rho=0.285$ của mô hình SAR. Đây là câu trả lời toán học khách quan giải thích cơ chế hình thành các cơn sốt đất cục bộ dưới góc nhìn địa lý - kinh tế lượng, thay vì các giải thích mang tính định tính mơ hồ trước đây.
3. **Bản chất bảo lưu rủi ro:** Thông qua cấu trúc ARIMA-GARCH, nghiên cứu chứng minh thị trường BĐS có độ trễ hấp thụ chính sách vĩ mô từ 3 đến 5 tháng và sở hữu một "bộ nhớ rủi ro" dài hạn ($\alpha_1+\beta_1=0.939$). Mọi trạng thái cực đoan của thị trường (đóng băng hoặc sốt nóng) một khi đã hình thành thì sẽ duy trì tính quán tính rất nặng nề và cần những cú hích chính sách có cường độ đủ mạnh mới có thể đảo chiều xu hướng.

5.2. Hệ thống giải pháp và Kiến nghị chính sách

Đối với Cơ quan Quản lý Nhà nước

- **Xây dựng Hệ thống Định giá Hàng loạt Quốc gia (National Mass Appraisal System):** Nhằm thực thi hiệu quả Luật Đất đai mới, Nhà nước cần loại bỏ triệt để phương pháp xây dựng khung giá đất thủ công mang tính hành chính và lạc hậu. Thay vào đó, cần áp dụng mô hình Kinh tế lượng Không gian kết hợp thuật toán Học máy (XGBoost) được trình bày trong nghiên cứu này để tự động hóa việc tính toán và kết xuất Bản đồ Giá đất Số hóa theo Thời gian thực. Mô hình này sẽ tự động bóc tách các giá trị ẩn của vị trí và hạ tầng để đưa ra mức giá tính thuế, giá đền bù giải phóng mặt bằng một cách minh bạch, khách quan, triệt tiêu tận gốc các hành vi trục lợi thông tin và giảm thiểu tranh chấp đất đai.
- **Chuẩn hóa Kho dữ liệu lớn về Giao dịch BĐS (Real Estate Big Data Warehouse):** Để các mô hình thống kê hiện đại phát huy tối đa sức mạnh dự báo, Chính phủ cần ban hành quy định bắt buộc mọi giao dịch chuyển nhượng BĐS phải thực hiện thanh toán không dùng tiền mặt qua tài khoản

ngân hàng trung gian chuyên biệt. Dữ liệu từ các ngân hàng, văn phòng công chứng và cơ quan đăng ký đất đai phải được liên thông đồng bộ về một trung tâm dữ liệu quốc gia, tạo nguồn dữ liệu đầu vào sạch, liên tục và đáng tin cậy cho các mô hình phân tích định lượng.

Đối với Các Định chế Tài chính và Doanh nghiệp Đầu tư

- **Nhúng Mô hình Thống kê Hiện đại vào Quy trình Quản trị Rủi ro Tín dụng:** Các ngân hàng thương mại và các quỹ đầu tư bất động sản cần lập tức chấm dứt phương pháp thẩm định giá tài sản thế chấp dựa trên khảo sát thực địa thủ công đơn lẻ vốn chứa đựng nhiều rủi ro đạo đức của cán bộ định giá. Các ngân hàng cần tích hợp phương trình phương sai điều kiện GARCH(1,1) vào hệ thống quản trị rủi ro để theo dõi liên tục chỉ số bất ổn σ^2 của từng phân khúc thị trường. Khi chỉ số này vượt ngưỡng cảnh báo an toàn, hệ thống sẽ tự động kích hoạt cơ chế phòng vệ: siết chặt tỷ lệ cho vay trên giá trị tài sản (LTV), tăng biên độ lãi suất rủi ro hoặc hạn chế giải ngân đối với các vùng địa lý đang có hệ số lan tỏa không gian ρ biến động bất thường, nhằm ngăn ngừa từ sớm nguy cơ nợ xấu hệ thống.

5.3. Hạn chế và Hướng phát triển của đề tài

Mặc dù đã xây dựng được một hệ thống mô hình toán thống kê tương đối toàn diện với hiệu năng dự báo cao, nghiên cứu này vẫn vấp phải những giới hạn nhất định do đặc thù nguồn dữ liệu. Toàn bộ tập dữ liệu thực nghiệm 15.000 mẫu hiện tại mới chỉ khai thác các trường dữ liệu có cấu trúc (Structured Data) mang tính vật lý và hình học. Trong thực tế, giá trị và xu hướng thị trường BĐS còn bị chi phối mạnh mẽ bởi các nguồn dữ liệu phi cấu trúc (Unstructured Data) khổng lồ như: hình ảnh kiến trúc thực tế của tài sản, dữ liệu văn bản từ các đồ án quy hoạch chi tiết, hoặc dữ liệu phân tích tâm lý đám đông (Sentiment Analysis) thu thập từ hành vi tìm kiếm và thảo luận của người dùng trên các nền tảng mạng xã hội.

Chính vì vậy, hướng phát triển tiếp theo và chuyên sâu của đề tài nghiên cứu này sẽ là việc mở rộng khung phân tích sang hệ hình **Học sâu đa phương thức (Multimodal Deep Learning)**. Cụ thể, nghiên cứu sẽ ứng dụng các cấu trúc Mạng Neural Đồ thị (Graph Neural Networks - GNN) để mô hình hóa các mạng lưới quan hệ không gian phi tuần tự phức tạp, kết hợp với các Mô hình Ngôn ngữ Lớn (LLMs) để trích xuất ngữ nghĩa từ các văn bản pháp lý quy hoạch. Sự kết hợp này được kỳ vọng sẽ tạo ra một trí tuệ nhân tạo định giá và dự báo xu hướng thị trường BĐS có năng lực toàn diện, tiệm cận hoàn hảo với các quyết định mang tính chuyên gia.

TÀI LIỆU THAM KHẢO

1. Anselin, L. (2001). Spatial econometrics. *A Companion to Theoretical Econometrics*, 310-330.

PDF

2. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307-327.
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
4. DiPasquale, D., & Wheaton, W. C. (1996). *Urban Economics and Real Estate Markets*. Prentice Hall.
5. Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987-1007.
6. Kok, N., Koponen, A., & Martinez-Barbosa, C. A. (2018). Big data in real estate: From manual appraisal to automated valuation models. *Journal of Portfolio Management*, 44(7), 133-143.
7. Malpezzi, S. (2003). Hedonic pricing models: a selective and applied review. *Housing Economics and Public Policy*, 67-89.
8. Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106.
9. Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34-55.
10. Sirmans, G., Macpherson, D., & Zietz, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, 13(1), 1-43.