GEOPHY

INSIDE THE GEOPHY AVM

THE EVOLUTION OF COMMERCIAL REAL ESTATE (CRE) VALUATIONS

Version 1.2.6
February 7, 2019


Canadarhq

TABLE OF CONTENTS

CRE VALUATIONS, REDEFINED

In the wake of the 2007 financial crisis — triggered, in part, by a poor understanding of securities pegged to real estate — financial institutions of all stripes now strive to get a clearer picture of the present and future value of real estate assets and portfolios. Yet traditional means of real estate valuation continue to rely on the intuition and proprietary information of brokers that often serve vested interests and obfuscate insights into the forces that impact real estate value.

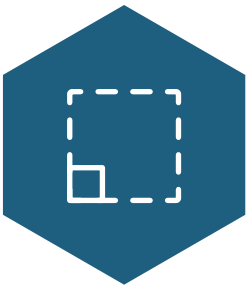
In a world where medical doctors now use data to predict infections before physical symptoms occur, and retailers exploit data generated from digitally connected channels to anticipate buying trends and promote growth, such a limited, 'small data' approach to real estate asset valuation is no longer sustainable.

The weight and impact of commercial real estate investment and management calls for a more efficient means of harnessing today's breadth of information for fast, actionable, data-driven valuations of commercial real estate assets. GeoPhy's automated valuation model (AVM) provides that much-needed innovation with enterprise-grade, AI-powered insight into the value, and value drivers of commercial real estate assets.

CLARITY

Property value drivers explicitly exposed.

ACCURACY



Data-driven values delivered at scale.

VALUE

Enterprise grade market values at affordable rates.

SPEED

CRE data and property values in minutes, not days.

GeoPhy, Amsterdam truck traffic data

Advances in data, analytics and ML continue to enhance the way we work, making us more productive, accurate and fast. In fact, a 2018 global institutional investor study found 62% of institutional investors believed trading algorithms and sophisticated quantitative models would make investment markets more efficient.[1] Yet real estate, the single largest asset class in the world, remains by and large an industry relying on manual processes, subject to human error and biases.

With a workforce of 74,000 appraisers in the U.S. alone, manually assessing assets sometimes worth billions of dollars, real estate has long been the beneficiary of a lucrative operating environment rewarded even amidst functional deficiencies.[2] Now, however, the industry faces a reckoning driven by fiduciary and regulatory scrutiny, and the need to mirror other industries that embrace technology to provide greater efficiency and value for its stakeholders.

To help modernize the industry approach to valuations, GeoPhy has developed the first purpose-built automated valuation model (AVM) for use by institutional investors and lenders in the commercial real estate (CRE) sector. Going beyond traditional property valuation techniques, GeoPhy takes a "big data" approach, in combination with sophisticated machine learning to provide faster, more reliable CRE property values at a fraction of the cost of a traditional property appraisal.

Built on a dynamic, semantic data integration platform, the GeoPhy AVM identifies and evaluates structural value drivers within the market to quickly and efficiently calculate commercial property values. These value drivers include both standard demographic and economic measures, in addition to more modern, "hyperlocal" metrics, such as proximity to music events, green space, local crime rates, and even the tone of reviews for local businesses.

Rather than traditional hedonic models, which are limited both statistically and by an individual's predisposition towards "standard" explanatory variables, GeoPhy's supervised machine learning models rely on stochastic gradient boosting decision trees (GBDT). The performance of these models improve over time, as we add new transaction data and additional contextual data sources to the model.

With GeoPhy's comprehensive, technology-enabled approach, institutional CRE investors and lenders can now base high-stakes investment and lending decisions on unbiased, data-driven valuations less prone to human error and more sensitive to a wide range of market indicators that more accurately assess current fair market values of commercial real estate.

[1] FRI, "The Fidelity Global Institutional Investor Survey: The Future of Investment Management" https://institutional.fidelity.com/app/proxy/content?literatureURL=/9891548.PDF

[2] GeoPhy, "Big Data in Real Estate? From Manual Appraisal to Automated Valuation" https://jpm.iijournals.com/content/43/6/202

AN IDEAL CASE FOR MACHINE LEARNING

Instead of being confined to lab experiments or research papers, the use of ML has accelerated with regard to mainstream industry adoption. There now exist endless business applications for machine learning, ranging from retail, finance, and healthcare, through to education and charity. In general, machine learning is particularly suited to problems where:

- Applicable associations or rules might be intuitive, but are not easily classified or described by simple logical rules;
- Potential outputs or actions are defined but which action to take depends on numerous conditions which cannot be predicted or uniquely identified before an event happens;
- The data is problematic for traditional analytical techniques. Specifically, extensive data (datasets with large numbers of data points or attributes in every record compared to the number of records) and highly correlated data (data with similar or closely related values) can present problems for traditional analytical methods.

THE CRE MACHINE LEARNING PROBLEM

Multiple approaches to commercial real estate valuation exist, each with its nuances and level of specificity required. Three of the most commonly used methods include:

- Cost approach
- Sales comparison
- Income approach (to include discounted cash flow, gross income multipliers, and direct capitalization)

While each approach has its advantages, they all require synthesis of significant amounts of information. Deciphering factors such as rental rates, vacancy rates, employment growth, demographics, new construction supply, zoning regulations, interest rates, capital availability, and more, are not simple propositions. Interpreting the interdependence of these variables, as well as accounting for historical data and knowing how to weight the variables properly is more than a single person can synthesize in a quick, efficient manner.

In contrast, ML models are specifically designed to process such vast amounts of data. They excel in environments with massive amounts of interdependent variables. In such scenarios, ML can determine a property's value in much the same way Amazon decodes the likelihood of users clicking on an advertisement.

GROUNDED IN PROVEN PRINCIPLES

While machine learning-based models already support single-family valuations, most AVMs still rely on highly-manual methods that have trouble keeping up with rapidly evolving industry needs, or use linear models specific to a region, metro area or neighborhood. These localized linear models neither capture nuanced interdependencies, nor capitalize on the rich breadth of data affecting CRE property values.

GeoPhy uses a wider spectrum of data to identify global patterns and uses hyperlocal data to refine analysis specific to a property. This key innovation allows GeoPhy to work with more data and explore non-linear relationships that expose relevant features that drive CRE property value.

Instead of relying on a small set of "comps" or cap rates, the GeoPhy AVM utilizes boosted decision tree models to generate valuations. At its core, the GeoPhy AVM uses a large set of commercial real estate transactions, exploiting structural relationships between observable characteristics, "value drivers" and property prices. The operating income of an asset (and its components) are key inputs into the model, as well as macro and micro characteristics of the property market.[3]

CORE ASSUMPTIONS

- Asset's financial data is key for an accurate valuation.
- Asset's hyperlocal vicinity data often underutilized, though a powerful driver of value.
- Local and regional market and economic trends play a critical role in valuation.
- Robust supervised ML techniques that utilize non-linear relationships in data better approximate value than simpler linear regression models.
- The combination of advanced modelling techniques and large, diverse, quality-grade datasets allows for a generalizable market-specific model that can be used to value assets across heterogenous geographic regions.
- Data features that drive a valuation up or down and the magnitude of their influence is a critical component of the AVM's value.
- Explainable, interpretable results and supporting data is a part of the GeoPhy AVM package.

[3] GeoPhy, "Big Data in Real Estate? From Manual Appraisal to Automated Valuation" https://jpm.iijournals.com/content/43/6/202

A MACHINE LEARNING PRIMER

Learning From Patterns

Machine learning models learn from patterns in the data, which may be, linear, non-linear and more likely a combination of both. If the dataset is labeled, then ML models can correlate the patterns found to the labels through supervised learning techniques. GeoPhy uses supervised methods to learn how to predict the monetary value (e.g., sale price) of an asset by using the sales price as a label. The goal when building a valuation model is to identify and learn from the most relevant patterns in the data, to best predict an asset's price.

Large Scale Trial and Error

The ML process is based on trial, error, and iteration of many computational experiments. Learning from these experiments is expedited by the use of a loss function (a mathematical formula which dictates how best to optimize results) which, in turn, helps guide the model to the most effective 'learning path' in a relatively short period. The power of ML comes from a feedback process that is enhanced by large amounts of data, advances in model techniques, and the availability of ever-increasing computational power. For example, before these advances, certain model functions could only be optimized analytically if the function was in closed form (i.e., it is possible to find an exact answer through a finite number of analytical techniques), whereas most real-world data are not this well-behaved and require vast iterations of possible solutions that eventually converge on a reasonable approximation. This requires advanced methods of inference, a lot of data and computing power.

An Approximation of Real World Interactions

Eventually, the learning process will plateau. Once this happens, the model will have identified and directly related the most important data patterns and interactions with a prediction (i.e., monetary value) to the best of its ability. If there are sufficient data and training, we can reasonably assume that these patterns and interactions approximate true real-world dynamics. This is one of the reasons success of the ML model is very much dependent on the datasets used to train it.

INTENTIONAL DESIGN

The GeoPhy data science team draws upon the expertise and guidance from professionals in the commercial real estate sector, including certified appraisers, economists, and a panel of external experts. We take an intentional approach to selecting contextual data features and considering data sources used to build our models. Where necessary, we remove data or sources observed to:

- Consistently, through cross-validation, show no influence on commercial real estate valuations.
- Directly affect compliance with The Equal Credit Opportunity Act and the Fair Housing Act.
- Fail our source reliability assessment and/or data quality tests.
- Introduce an imbalance or measurable bias to our valuations.
- Poorly represent the commercial real estate market distribution or the AVM's spatial-temporal scope.

ITERATIVE LEARNING

The GeoPhy model development team performs in-depth error analysis as part of the development cycle, which guides the team when optimizing contextual data enrichments and feature engineering. For model performance, detailed analysis also helps us understand the strengths and weaknesses of the GeoPhy AVM.

We prioritize an iterative cycle to ensure our AVMs perform well on properties typical for a given commercial real estate market scope, i.e., properties that trend towards the normal distribution in terms of financial income and property size. We understand one size may not fit all and perform investigative analysis on our testing results to better explain our confidence in the valuation we provide.

We provide additional information on how we measure success of our models in the Model KPI section.

5-STEP DESIGN & BUILD ITERATION

The GeoPhy AVM model is developed and continuously updated using 5 core steps:

GET DATA

When it comes to the effectiveness of machine learning, more relevant data almost always yields better results — and the real estate industry currently sits on wealth of data. But not all data is suitable or structured for immediate use. That is where the strength and value of GeoPhy's Semantic Data Management Platform (GeoPhy DMP) comes into play. Using an ontology developed explicitly by real estate experts for the real estate market, the GeoPhy DMP automatically ingests, merges and processes files from thousands of different sources, ranging from county tax records to hyperlocal contextual data sources to semantically connect data points.

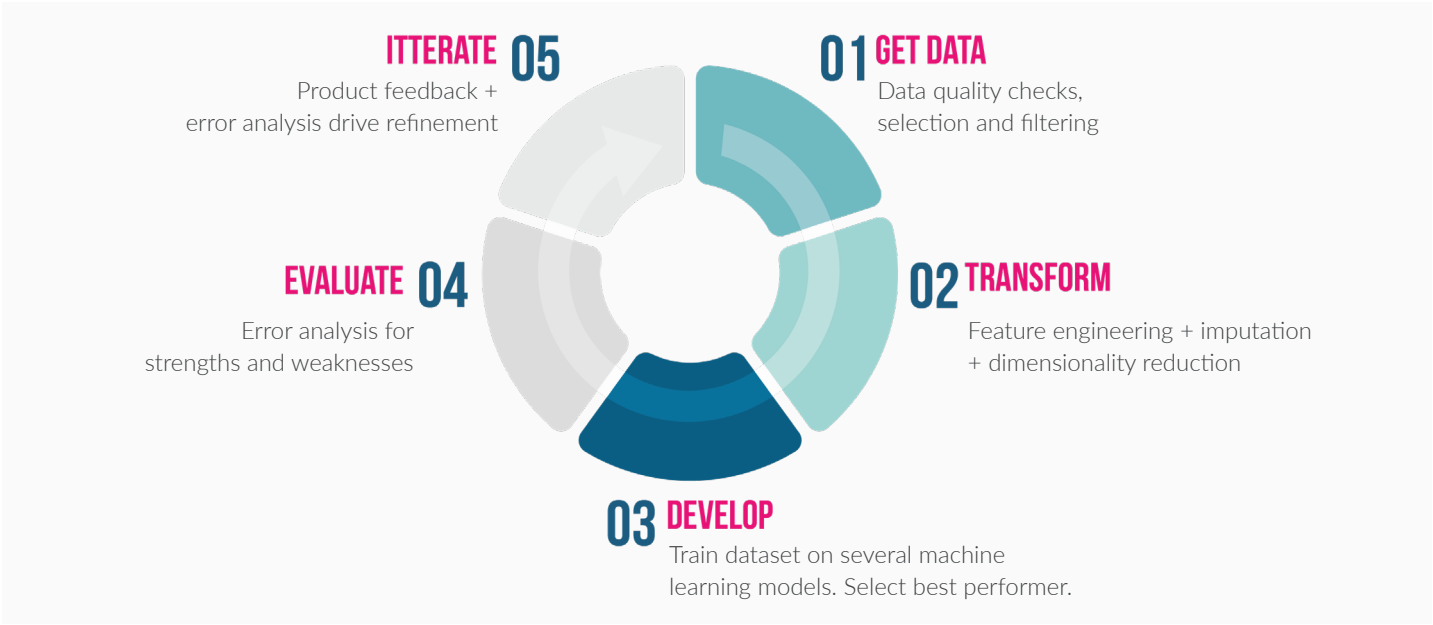


Exhibit 1: *The GeoPhy AVM design cycle*

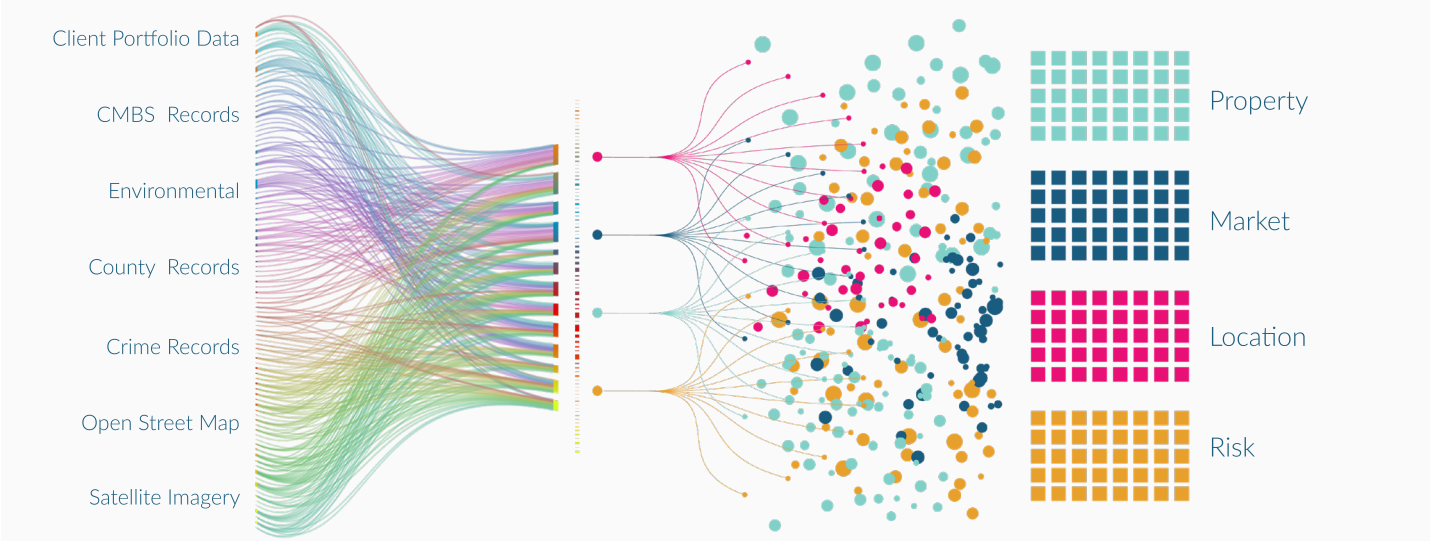


Exhibit 2: *Aggregated data from over 10,000 individual sources*

DATA SOURCES: **PROPERTY DATA**

Property-specific transaction data is key to producing accurate valuations, and we extract asset data from a range of sources, including:

- Public property records sourced from all county assessor and county recorder offices
- CMBS records, sourced from the SEC
- REIT filings, sourced from the SEC
- REIT websites and directly from REITs
- Client data, including lenders, limited partners (LPs) and private equity investors
- Appraisal reports

We utilize asset data on:

- Transaction price
- Transaction date
- Operating statements (including net operating income, expenses, etc.)
- Occupancy rate
- Asset size
- Other property-specific characteristics

We typically refresh data quarterly, with some sources updated at higher frequencies. All data is automatically ingested into the GeoPhy DMP, allowing for accurate and fast refreshment of data that feeds the GeoPhy AVM.

DATA SOURCES: **CONTEXTUAL DATA**

In addition to property data, we collect contextual data sources relevant to the commercial real estate market or sub-market and enrich property datasets. This contextual data includes information that ranges from the hyperlocal to the macro level, reflecting the importance of "location location location" in determining the fair market value of commercial real estate assets.

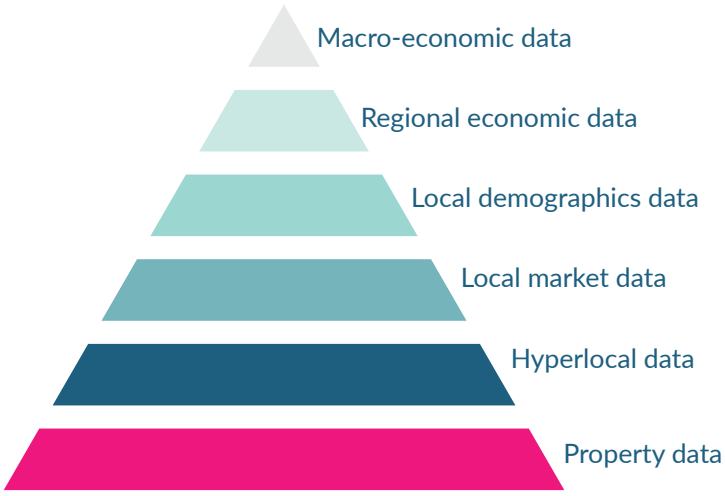


Exhibit 3: *A high-level overview of the contextual data used in the GeoPhy AVM.*

National Level
(e.g. 30-year mortgage rates, stock price index)
- Macroeconomic Indicators

Regional Level
(e.g. mortgage delinquency rates)
- Economic Indicators

Local Level
(e.g. crime reports, school proximity)
- Demographic Statistics
- Employment & Income Statistics
- Market Supply
- Housing Statistics
- Market Indicators
- Industry Statistics
- Education Statistics
- Transportation Statistics
- Crime Statistics

Hyperlocal Level*
(e.g. retail presence, events density)
- Hyperlocal Amenities
- Hyperlocal Social Trends
- Hyperlocal Scores and Indexes

**GeoPhy enrichment data layers*

STRUCTURE DATA

As commercial real estate data becomes more digital, data collection will become more structured and more user-friendly to process. Until that time, GeoPhy's data engineers and semantic data architects wrangle unstructured data into structured, annotated datasets.

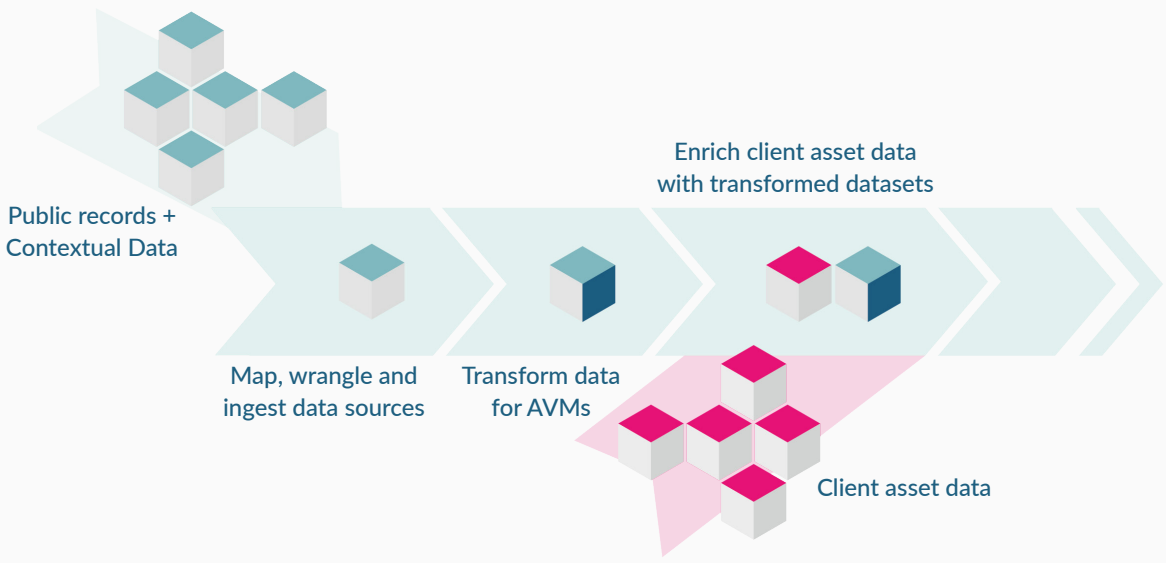


Exhibit 4: *GeoPhy structured data flow*

DATA QUALITY

Data quality has always been a persistent risk to the users of model-based information products. The impact of incorrect, incomplete, missing or untraceable data is a well-known problem. As data consumers and data product developers, data quality touches all parts of GeoPhy and our approach to data quality is proactive by design. To that end, we take the following steps to source, check and select data to ensure data accuracy, relevance, coverage, and granularity.

DATA SOURCE ASSESSMENTS

We address the intake risks of collecting and ingesting external contextual data sources by performing source reliability assessments. We also check for source consistency, and where possible, validate specific values by comparing samples with alternative sources to better understand its accuracy, coverage, and completeness.

ONGOING DATA CONSISTENCY CHECKS

Once a source has passed our initial assessments, updates to the source — whether streamed into our data management platform or arrived in pre-wrangled batches — are value spot-checked and assessed for distribution similarity. If discrepancies are found, our data engineers and data scientists work to resolve the issue before it impacts our products. This is critical for the ongoing performance and maintenance of our AVM products.

DATA SELECTION

Once the property data in our semantic data integration platform has been enriched with contextual data sources, a final exploratory data analysis is performed to evaluate the quality, completeness and expected distribution of the dataset. This includes:

- Assessing expected distribution and statistical moments (mean, variance, skewness, and kurtosis), and value ranges of each contextual data source
- Checking whether data enrichment matching is executed as intended and/or monitoring the level of missing values
- Identifying outliers by taking into consideration the size and income of properties

Once the data has passed quality checks, our data scientists apply filters and remove outliers that do not meet market-based thresholds that we set. These asset filter thresholds are typically developed with internal real estate experts and are not solely based on statistical thresholds.

TRANSFORM

Our data scientists further enhance data ingested into GeoPhy's DMP to ensure data used to build the model appropriate for the CRE market. This enhancement or transformation includes feature engineering and feature selection. Where necessary, our data scientists may choose to fill the gaps of incomplete data with imputed (estimated) values to ensure our models can get the most out of the datasets without biasing results.

FEATURE ENGINEERING

In addition to the data layers and indexes developed by GeoPhy, we put significant effort into feature engineering (the process of creating derived data). We ensure the derived features used in the GeoPhy AVM models are both logical for the valuation and easily interpretable for our clients. As an example, we aggregate reported crime data from all categories in a given county and normalize total instances of crime by the estimated population for a given time period and location.

DATA IMPUTATION

While the majority of our enrichment datasets have complete geospatial-temporal coverage, depending on the dataset, we may need to fill missing values. We use imputation methods based on network propagation, nearest neighbor approaches or time-averaged statistical estimates. As an example, for some zip code tabulation areas (ZCTAs), the percentage of owner-occupied housing data is not available and therefore will be imputed using nearest neighbor methods (selecting and averaging neighborhood zip code values) to best approximate the area's value. As a general rule, we do not impute critical property data such as net operating income. Imputation is therefore limited to spatial-temporal or temporal-based contextual datasets.

FEATURE SELECTION

Since it is not unusual for our enriched datasets to have thousands of features, feature reduction is necessary for optimal model performance. Dimensionality reduction, a commonly used data science method to address this issue, tends to make the GeoPhy AVM model less interpretable for our clients. To address that, we apply internally developed, explainable feature reduction algorithms to the dataset to preserve the features as they were originally calculated. We reduce the number of data features used to train the GeoPhy AVM model (the dataset dimension) to those most diverse from each other and typically, least correlated to each other.

DEVELOP MODEL

Once a dataset has been fully processed, GeoPhy data scientists train, select, optimize and evaluate our models to decipher the relationship between property value drivers and transaction price.

MODEL SELECTION

The goal when building the GeoPhy AVM model is to identify and learn from the most relevant patterns in order to best predict the "true" transaction price. To that end, our data scientists test the performance on a range of trained ML models against each other and its baseline reference using a reserved testing dataset.

We compare predictions made by various models using a battery of standard model key performance indicators (KPIs) tests and charts that demonstrate relative accuracy (precision), overfitting (model bias), distribution of errors and stability of top features. Our data scientists collaborate as a team to select the best performing model for the given problem scope.

MODEL OPTIMIZATION

The models have critical configuration parameters that cannot be directly estimated from data. In machine learning, we call these hidden features hyperparameters (not to be confused with model parameters).

HYPERPARAMETERS AND MODEL PARAMETERS EXPLAINED

We design each machine learning algorithm or flexibility. Hyperparameters allow data scientists to customize a model's learning function. They include configuration limits (e.g., the maximum number of hierarchies in a

decision tree) and the speed at which a model increments its 'learning path' (e.g., alpha or learning rate) among many other settings. Hyperparameters vary depending on the algorithm used, however determining the optimal hyperparameters it is not easily discernible from observing the dataset alone, so tuning needs to be done to optimize them. Hyperparameters are typically tuned by automating thousands of trial and error experiments that iterate through multiple combinations of hyperparameter settings. We then use the hyperparameters that yield the best results (judged on model metrics) in the final model. Once the model learns from the dataset using the optimized hyperparameters, it saves the learned patterns as parameters (e.g., a specific numeric threshold in a decision tree split). These are the model's parameters, also known as model weights. Model parameters are then used make predictions on new data.

GeoPhy data scientists optimize our AVM performance through hyperparameter tuning, with the primary aim of lowering the median absolute percentage error (MdAPE) while avoiding overfitting by reducing the Relative Mean Absolute Error (RelMAE) between training and test datasets. Our ultimate goal is to ensure the largest possible share of valuations is within +/- 10% of the actual asset transaction price.

We then select the best performing model from the experimental batch for further evaluation, testing, development, and refinement. In some cases, we build more than one AVM for a commercial real estate market or region. Alternatively, we use a combination of models in the valuation flow. This allows the GeoPhy AVM products to remain flexible to client needs and varying levels of data availability.

EVALUATE

The best models from the development cycle are those that meet or surpass our performance goals. These models then undergo a detailed error analysis, quality testing and qualitative evaluation phase to understand the models' strengths and weaknesses. The error analysis also includes investigating anomalous, problematic valuations (those that are significantly under- or overvalued), researching and identifying patterns or groups with the assets' subsets that may not be well represented in the current model's data features.

CROSS-VALIDATION & MODEL KPIs

Cross-validation is a method used in machine learning to minimize the effects of sampling bias when developing machine learning models and evaluating their performance. GeoPhy uses this technique in all stages of the process. The exercise aims to quantify the following question:

" How can I trust the ML model will perform consistently on different sets of property data? "

The best way to answer this question is to give the model different datasets to learn from (train), on which it can test and validate results. In the first iteration, the first dataset is used to test the model and the rest are used to train the model. In the next iteration, the second dataset is used as the testing set while the rest serve as the training set. This process is repeated until each of the datasets have been used as the testing set.

Cross-validation systematically does this and averages the performance results so that we have AVM model metrics that better represent its performance in reality.

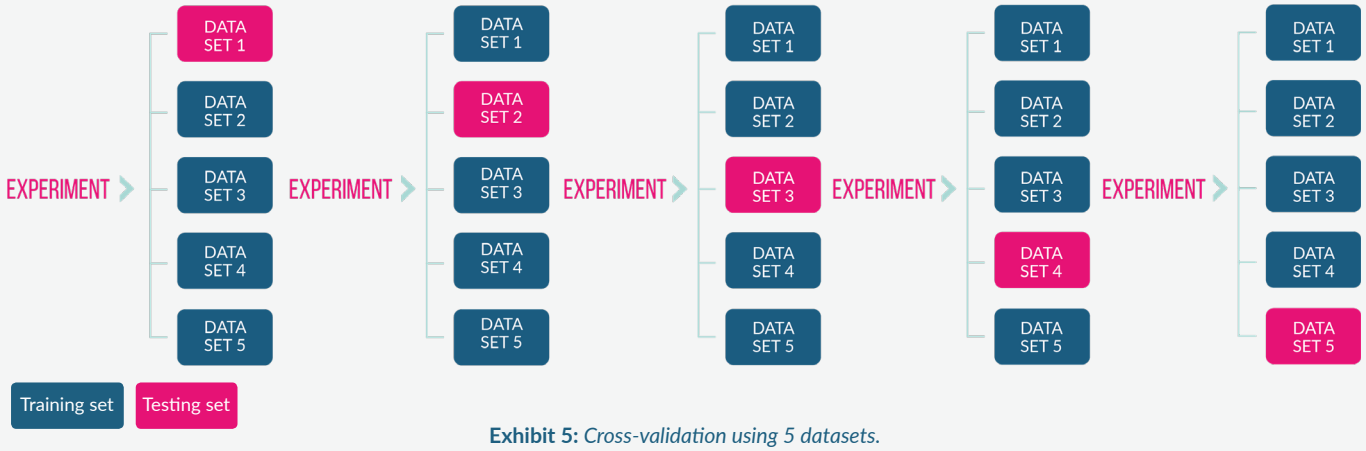


Exhibit 5: *Cross-validation using 5 datasets.*

GeoPhy's goal is to develop models that maximize explained variance while optimizing accuracy. However, understanding bias in our model, robustness, overall reducible error and interpretability are also important considerations for model performance and usability. The following are a selection of KPIs we run for each AVM model.

ERROR METRICS

The measure of model performance by assessing the amount of error in the valuations compared to their actual transaction prices:

MdAPE - Median absolute percentage error: Given a testing dataset, 50% of properties valued by the model will have an error within +/- this value.

PPE < 10% - Percentage point error within +/- 10%: Given a testing dataset, the assets in the dataset were valued within +/- 10% of their actual transaction price.

RelMAE - Relative mean absolute error: The average magnitude of model error when testing new (unknown) transactions compared to training (benchmark) transactions. The lower the value, the better the model is at approximating real-world dynamics.

MdPE – Median percentage error: Given a testing dataset, the 50th percentile of percentage error in percentage terms. Ideally, this would be close to zero, implying the model has no bias in terms of repeatedly overvaluing or undervaluing assets.

ROBUSTNESS & INTERPRETABILITY METRICS

The measure of stability of the model with non-random subsample testing and the ease of which clients can interpret AVM output valuations.

Robustness score - Weighted score of key error metric's standard deviations reciprocals: The higher the value, the more robust the model.

Interpretability score - A subjective score, based on an internal assessment. The score takes into consideration the type of modeling algorithm(s) used, individual feature interpretability and proportion of derived data vs. raw data. The higher the value, the more interpretable the model.

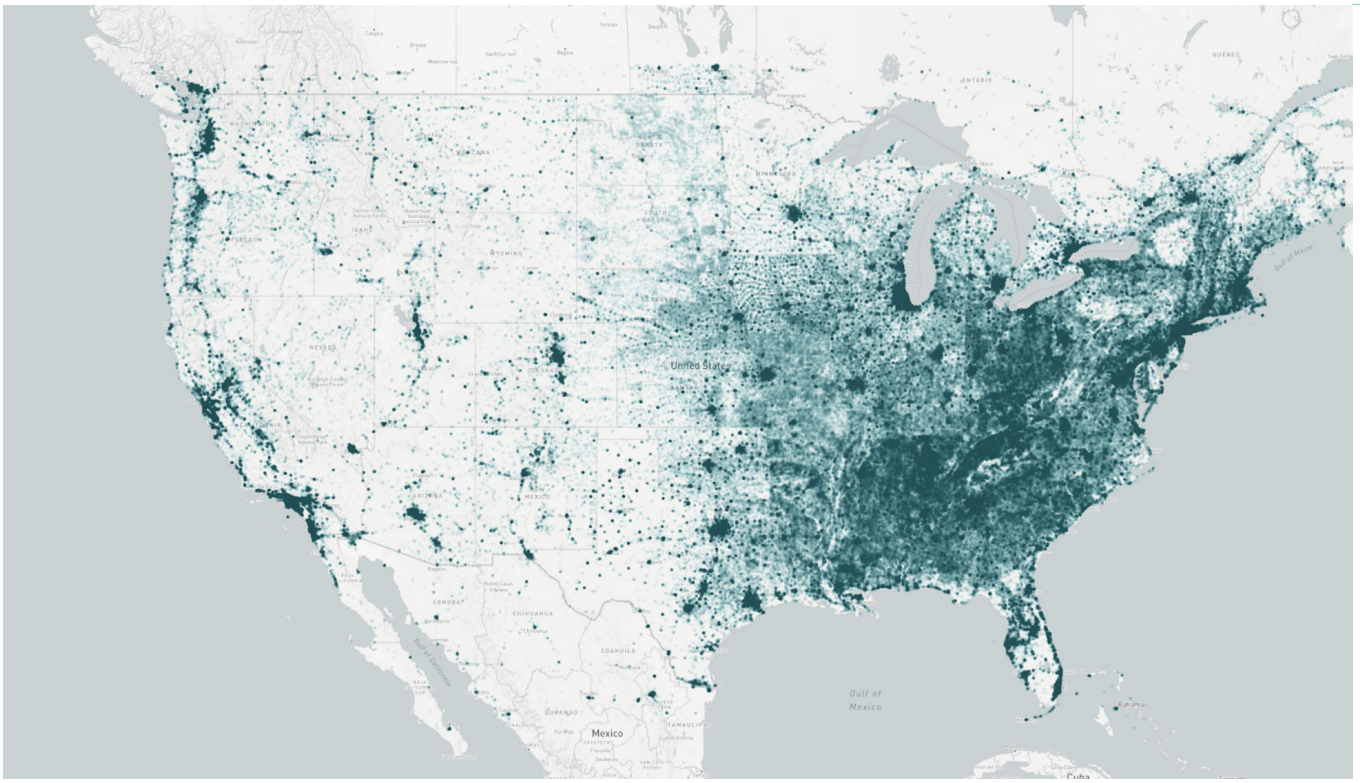
ITERATE

GeoPhy continuously explores new hypothesis for model improvement. Our data engineers and scientists document observations of sufficient statistical significance on error distribution — typically based on a new grouping or pattern manually found during evaluation and error testing — for future development. Where needed, we collect and include additional datasets into the next model development cycle along with internal testing and client feedback on the model.

ONGOING TESTING AND QUALITY CONTROL

The GeoPhy AVM undergoes continuous automated testing to ensure performance of the model and its valuations meet our quality acceptance criteria. Since commercial real estate assets do not transact frequently, we choose to assess the validity of each client valuations by checking:

- Client provided property values in line with historical data in our databases (where available).
- The implied capitalization rate resides within an acceptable range for comparable assets, where the prevailing cap rate is determined by using the GeoPhy TrueCOMP tool.
- Valuations meet a minimum confidence level determined at the model-property-combination level.
- Valuations sit within a percentage point range from the inflation-adjusted and/or housing index-adjusted last sale price.
- Large fluctuations in valuations across time are valid and can be explained by the changes in value drivers.



GeoPhy, US amenities distribution

DATA-DRIVEN VALUES

The GeoPhy AVM exploits the combination of thousands of sources of property data and contextual data, brought together into the GeoPhy Semantic Data Management Platform, with advanced, supervised ML modeling techniques.

Departure from the assessment of commercial real estate values using a small number of comparable local properties, to using all historical transactions of properties of a similar type (e.g., multi-family), may feel a novel approach to the industry. However, there is a common understanding that locational characteristics that determine local property values are not unique to one specific location — presence of public transit, schools, amenities, supply density, etc., all structurally affect real estate values.

Exploiting the systematic relationship between value drivers and real estate prices is made possible by the advent of large pools of historical transaction data, combined with accurate information on the operating performance of commercial real estate assets. While GeoPhy believes in using human expertise to evaluate model inputs — for example, to adjust net operating income (NOI) to reflect one-off idiosyncrasies in income and/or expenses — we use advanced machine learning algorithms to build the most accurate, unbiased, efficient valuation model. To further promote accuracy and client confidence in the reliability of ML-based value assessments, we validate the model both automatically and manually for continuous improvement.

As access to data increases, the GeoPhy AVM continues to increase in accuracy, speed and cost efficiencies. These benefits will ultimately drive better understanding of real estate value, and accelerate judicious commercial real estate performance decisions.

GEOPHY

ABOUT GEOPHY

Founded in 2014, GeoPhy aims to transform antiquated commercial real estate (CRE) processes with data-driven valuations and analytics powered by machine learning. Its AI-powered valuations uncover value drivers that help steer acquisition due diligence, portfolio monitoring, and site selection for institutional lenders and investors in the real estate and financial sectors.

GeoPhy has garnered interest from a wide spectrum of clients including major rating agencies, banks, pension funds, investors, national regulators and large government-backed enterprises in the US and Europe.

Ready to get started?

Get in touch to learn more about GeoPhy's AVM and our integrated valuation solution suite.

info@geophy.com
www.geophy.com

The Netherlands
Stationsplein 10
2611 BV Delft
+31 (0)15 737 0293

United States
530 7th Ave
Suite 1909
NY 10018

@geophy geophy-hq