SOUTHERN LUZON STATE UNIVERSITY
College of Engineering
COMPUTER ENGINEERING DEPARTMENT

Machine Learning Cognate 1 Project

Final Report

In partial of fulfillment of the requirements in

CPE15 – Cognate and Professional Course 1

Submitted by:

ENCALLADO, CARL FRANCIS T.

Submitted to:

ENGR. JULIE ANN SUSA-GILI, MSEE - CPE

Course Instructor

December 12, 2025

ABSTRACT

This study employs an integrated geospatial and labor market approach to identify optimal locations for IT–BPM (Information Technology–Business Process Management) hubs in the Philippines. Leveraging GPS connectivity data from Ookla, administrative boundaries from the Global Administrative Areas database, and employment indicators from the Philippine Statistics Authority's Labor Force Survey, the research quantifies regional digital readiness. Machine learning techniques, specifically K-Means clustering, are applied to classify municipalities based on multidimensional factors, including internet performance, talent availability, geographic accessibility, and market saturation. Analysis reveals significant disparities in connectivity, with Metro Manila achieving average download speeds of 155,572 kilobits per second, contrasted with 23,704 kilobits per second in Lanao del Sur. Cluster results highlight twenty high-potential municipalities for strategic IT–BPM investments, including Davao City and Naga City. The findings demonstrate the efficacy of combining geospatial analytics and machine learning for data-driven decision-making, offering actionable insights to bridge the digital divide and promote equitable socioeconomic growth across the Philippine archipelago.

TABLE OF CONTENTS

TABLE OF FIGURES

PROBLEM STATEMENT

The Philippine economy faces a persistent digital and spatial divide, where high-value IT–BPM jobs, over 1.82 million in total, are heavily concentrated in Metro Manila and a few urban growth centers, leaving skilled workers in provincial regions with limited opportunities despite relevant qualifications (Desiderio, 2025). This imbalance drives migration pressures, reinforces regional inequalities, and constrains inclusive economic growth. The core challenge is identifying municipal-level locations suitable for IT–BPM expansion by integrating geospatial connectivity data, population and labor force characteristics, and spatial accessibility. Leveraging machine learning techniques such as K-Means clustering, this study captures spatial patterns and classifies municipalities based on digital readiness, talent availability, and market potential. By systematically mapping latent opportunities beyond traditional urban hubs, the approach provides actionable insights for targeted infrastructure development and investment strategies, bridging the digital divide and fostering equitable socioeconomic progress across the Philippines.

LITERATURE REVIEW

Previous applications of geospatial machine learning in the Philippine context have largely focused on estimating socioeconomic indicators such as poverty prevalence by combining satellite imagery, ground-level surveys, and deep learning (Tingzon et al., 2019). These studies demonstrate that landscape features, built environments, and luminosity patterns can serve as reliable proxies for economic outcomes, particularly when combined with neural networks (TMDS, n.d.). Such research establishes methodological foundations for extracting nuanced socioeconomic signals from spatial data in environments characterized by limited local measurement. Parallel studies in the IT-BPM sector underscore the increasing integration of artificial intelligence, automation, and digital transformation as competitive drivers, as well as the importance of workforce development and government-led policy interventions (Newsbytes.PH, 2020). However, few empirical studies have examined the geographic allocation of IT-BPM infrastructure in relation to digital connectivity, labor availability, and spatial accessibility (Santos Knight Frank, 2024). This research contributes to closing that gap by applying unsupervised learning

to municipal-level geospatial features, offering a data-driven lens for understanding regional potential beyond traditional economic centers.

DATA PREPROCESSING

Multiple datasets were imported, cleaned, and integrated to create a robust feature-engineered dataset for analysis. CRS was standardized (follows the right-hand rule) across all geospatial inputs to ensure spatial accuracy, missing values were handled appropriately to prevent analytical distortions, and unnecessary rows or columns were removed to streamline processing. Geospatial operations were performed using GeoPandas and Folium, while Pandas facilitated merging, aggregation, and summarization. Key steps included standardizing municipality names, calculating distances from Metro Manila using the Haversine formula, and aggregating connectivity metrics from Ookla's quadkey-based tiles. Labor Force Survey data were filtered to identify working-age individuals with no employment as a proxy for available talent, and population data from the Philippine Standard Geographic Code (PSGC) were used to compute density-normalized indices. Techniques such as distance computations with the Haversine formula, use of GeoPandas for spatial joins, and data cleaning best practices follow standard geospatial data science workflows (Zandbergen, P. A. 2009, Journal of Spatial Science).

The final dataset encompasses cities or municipalities, blending raw metrics (connectivity, population, unemployed skilled workers, and distance from Metro Manila) with engineered indicators such as Digital Readiness, Talent Pool, Saturation, Accessibility, IT-BPM Hub Score, and cluster labels. By combining spatial, infrastructural, and socioeconomic dimensions, this dataset provides a comprehensive, balanced foundation for predictive modeling and geospatial analysis, enabling systematic identification of municipalities with high potential for IT–BPM development.

The following tables list all datasets used:

[01] Dataset Name	MachineLearningModel_Encallado.ipynb
Description	

This Jupyter Notebook serves as the core analytical workflow for the "GPS Connectivity Analysis for IT-BPM Hub Prediction in the Philippines" project. It includes detailed code implementations for data loading from various sources (e.g., CSV, GeoJSON, and ZIP files containing shapefiles), exploratory data analysis (EDA) with visualizations such as histograms, scatter plots, and bar charts for connectivity metrics, feature engineering steps (e.g., calculating indices like Talent Pool, Digital Readiness, Accessibility, and Saturation), machine learning modeling using K-Means clustering for city classification, and generation of outputs like interactive maps and cluster analyses. The notebook is structured with markdown explanations, code cells, and inline comments to ensure reproducibility, covering imports, preprocessing, modeling, and interpretation of results.

Source Link

[*Internal – this project*]

File Information & Size

Jupiter Source File (.ipynb) – 4.00 MB

About

This notebook acts as a comprehensive reference for all analytical steps in the project, justifying machine learning model choices and outputs such as cluster interpretations for IT-BPM hub recommendations. It integrates geospatial, connectivity, and employment data to predict emerging hubs, providing insights into digital divides and talent availability, with visualizations supporting findings on urban-rural disparities.

[02] Dataset Name

gadm41_PHL_2.json

Description

The GADM (Global Administrative Areas) dataset provides high-resolution spatial data for administrative boundaries worldwide, with version 4.1 delimiting over 400,276 administrative areas across countries. For the Philippines, the Level 2 file (gadm41_PHL_2.json) includes detailed polygon geometries for provinces, municipalities, and cities. Data is sourced from official government maps, crowdsourced contributions, and satellite imagery, updated periodically. It supports formats like GeoJSON for easy integration with GIS tools, and is licensed for non-commercial use with attribution required.

Source Link

https://gadm.org/download_country.html

File Information & Size

JSON Source File (.json) – 2.34 MB

About

This dataset is essential for geospatial analysis in the project, providing polygon geometries for mapping connectivity tiles from Ookla data onto Philippine municipalities. It enables choropleth visualizations, spatial joins with connectivity metrics, and distance calculations, helping identify regional disparities in digital infrastructure and supporting recommendations for IT-BPM investments.

[03] Dataset Name

2020-04-01_performance_mobile_tiles.zip

2020-04-01_performance_mobile_tiles —+

2020-04-01_performance_mobile_tiles.dbf —+ |

2020-04-01_performance_mobile_tiles.prj —+ |
2020-04-01_performance_mobile_tiles.shp —+ |
2020-04-01_performance_mobile_tiles.shx —+ |

Description

This dataset from Ookla Open Data captures global network performance metrics. It includes key variables such as average download speed (avg_d_kbps), average upload speed (avg_u_kbps), average latency (avg_lat_ms), number of tests, and unique devices, derived from millions of Speedtest with GPS accuracy. Coverage spans quarterly from Q1 2019 to Q3 2025, with data filtered for cellular connections (e.g., 4G LTE, 5G NR). The ZIP file contains Esri Shapefile components (.shp for geometries, .dbf for attributes, .prj for projection, .shx for indexing), using WGS 84 (EPSG:4326) for geometries and EPSG:3857 for projection.

Source Link

https://github.com/teamookla/ookla-open-data

File Information & Size

Compressed (zipped) Folder (.zip) – 236 MB

About

The Speedtest data in this dataset supports efforts to enhance network performance, ensure regulatory accountability, and promote equitable Internet access by assisting operators, governments, and institutions in identifying and addressing connectivity gaps. In this project, it provides the foundational connectivity metrics for the Philippines, aggregated by city or municipality to compute Digital Readiness Indices, revealing urban-rural divides and informing ML clustering for IT-BPM hub predictions. Licensed under CC BY-NC-SA 4.0, it emphasizes privacy compliance through periodic reaggregation.

[04] Dataset Name

PSGC-3Q-2025-Publication-Datafile.csv

Description

The Philippine Standard Geographic Code (PSGC) is a systematic classification system developed by the Philippine Statistics Authority (PSA) for all geographic areas in the Philippines, assigning unique codes to regions, provinces, municipalities/cities, and barangays. This Q3 2025 CSV file includes comprehensive details such as 10-digit PSGC codes, names, correspondence codes, geographic levels (e.g., region, province, municipality), old names, city classes, income classifications (per DOF DO No. 074.2024), urban/rural designations.

Source Link

https://psa.gov.ph/system/files/scd/PSGC-3Q-2025-Publication-Datafile.xlsx

File Information & Size

Microsoft Excel Comma Separated Values File (.csv) – 2.19 MB

About

This dataset supplements population and boundary data, enabling accurate matching of municipalities in geospatial analysis and integration with connectivity and employment metrics. It justifies population-based indices like the Talent Pool Index by providing 2024 projections and supports normalization of features for ML modeling, ultimately aiding in the identification of high-potential IT-BPM hubs.

[05] Dataset Name

psgc_data_cleaned.csv

Description

This cleaned version of the PSGC dataset focuses on municipality-level data, including normalized names, geographic levels, income classifications per DOF DO No. 074.2024, and 2024 population projections. Derived from the full PSGC, it removes redundancies, standardizes formatting for merging, and includes columns like Name_Normalized for fuzzy matching with other datasets.

Source Link

[*Internal – this project*]

File Information & Size

Microsoft Excel Comma Separated Values File (.csv) – 58.1 KB

About

Essential for data integration in the project, this cleaned file provides normalized names and population data to merge with connectivity metrics from Ookla and employment indicators from LFS. It justifies population-weighted analyses and feature engineering, such as scaling indices by population size, contributing to accurate predictions of talent availability in potential IT-BPM locations.

[06] Dataset Name

PHL-PSA-LFS-2024-10-PUF.zip

PHL-PSA-LFS-2024-10-PUF —+
LFS_PUF_October_2024.F2 —+ |
LFS October 2024 Questionnaire.html —+ |
LFS PUF October 2024.csv —+ |
lfs_october_2024_metadata(dictionary).xlsx —+ |
LFS_PUF_October_2024.dcf —+ |

Description

The Labor Force Survey (LFS) is a quarterly household-based survey conducted by the Philippine Statistics Authority (PSA) to gather data on the demographic and socioeconomic characteristics of the population, focusing on labor market indicators. The October 2024 Public Use File (PUF) ZIP includes the main CSV dataset with anonymized records, metadata in Excel (dictionary with value sets), questionnaire HTML, and data codebook files. It covers a large sample (approximately 45,000 households nationwide), using a multi-stage sampling design to estimate employment, unemployment (3.9% in October 2024 preliminary results), underemployment, and related metrics at national and regional levels.

Source Link

https://psada.psa.gov.ph/catalog/LFS/about

File Information & Size

Compressed (zipped) Folder (.zip) – 6.27 MB

About

As the core dataset for extracting employment status and "No Work" counts per location, this file is crucial for predicting IT-BPM talent availability by integrating labor indicators with connectivity data. It supports feature engineering (e.g., unemployment proxies) and ML inputs, providing evidence-based insights into workforce potential for hub recommendations, aligned with national goals for economic planning and job creation.

[07] Dataset Name

lfs_october_2024_metadata(dictionary).xlsx lfs_october_2024_valueset_C12A.csv

Description

This CSV extract from the LFS metadata Excel file specifically contains the value set for variable C12A (Location of Work - Province, Municipality), mapping codes to names (e.g., 0101 for Abra - Bangued). It serves as a lookup table for processing location-based employment data, derived from the full metadata dictionary which includes all variable descriptions, codes, and value labels.
Source Link [*Extracted from lfs_october_2024_metadata(dictionary).xlsx - lfs_october_2024_valueset.csv*]
File Information & Size Microsoft Excel Comma Separated Values File (.csv) – 59.9 KB
About This helper dataset facilitates the extraction of "No Work Count" from LFS by enabling location-specific aggregation, relevant for assessing talent availability in IT-BPM predictions. It enhances data integration accuracy, supporting analyses of employment trends by municipality and contributing to cluster-based hub identifications.

[08] Dataset Name Feature-Engineered Dataset_Encallado.csv
Description This raw feature-engineered CSV compiles quadkey-based data from Ookla tiles, including connectivity metrics (avg_d_kbps, avg_u_kbps, avg_lat_ms, tests, devices), joined with province/municipality names (NAME_1, NAME_2), 2024 population, distances from Metro Manila, and No_Work_Count from LFS. It represents the initial merged dataset before aggregation.
Source Link [*Internal – this project*]
File Information & Size Microsoft Excel Comma Separated Values File (.csv) – 9.77 MB
About As the input for further modeling, this dataset supports feature engineering tasks like distance calculations and unemployment proxies, providing a granular view of connectivity-employment intersections. It justifies spatial patterns in digital readiness and is foundational for deriving aggregated indices used in ML clustering.

[09] Dataset Name Feature-Engineered Dataset_cleaned.csv
Description This aggregated CSV summarizes connectivity metrics, population, distances, and No_Work_Count at the municipality level, with engineered features ready for modeling. It includes records like top/bottom performers and serves as the cleaned input for indices and clustering.
Source Link [*Internal – this project*]
File Information & Size Microsoft Excel Comma Separated Values File (.csv) – 105 KB
About This file encapsulates engineered features for ML, justifying clustering by integrating geospatial, connectivity, and employment insights. It highlights disparities (e.g., high

No_Work_Count in rural areas) and supports recommendations for IT-BPM hubs with balanced metrics.

[10] Dataset Name PreprocessedDataset_Encallado.geojson
Description This preprocessed GeoJSON file contains quadkey polygons with associated connectivity metrics (speeds, latency, tests, devices), transformed for visualization and analysis, including WKT geometries and attributes for overlay on maps.
Source Link [*Internal – this project*]
File Information & Size GEOJSON File (.geojson) – 66.4 MB
About As the geospatial layer for mapping, this dataset is essential for overlaying connectivity on administrative boundaries, enabling visualizations of trends and supporting spatial queries in the project for identifying potential IT-BPM locations.

[11] Dataset Name InteractiveFoliumMap_Encallado.html
Description This interactive HTML file, generated using Folium, displays a map of the Philippines with layered connectivity metrics (e.g., heatmaps for speeds/latency), administrative boundaries from GADM, and markers for key cities, allowing zooming, panning, and layer toggling for exploratory analysis.
Source Link [*Internal – this project*]
File Information & Size HTML Document (.html) – 13.0 MB
About This output facilitates interactive exploration of connectivity trends, justifying visual insights into digital divides (e.g., poor coverage in remote areas) and supporting the project's findings on hub potential through geospatial context.
Note: To access all source files online, please click this link provided. All datasets, scripts, and supporting documents are available through the shared repository for easy viewing and download.

EXPLORATORY DATA ANALYSIS (EDA) RESULTS

Exploratory analysis revealed pronounced heterogeneity in digital infrastructure, population distribution, and labor characteristics across municipalities. The positive skewness and heterogeneity in connectivity metrics reflect typical ICT infrastructure in archipelagic developing countries (Akamatsu, N. 2022, Telecommunications Policy) Download and upload speeds varied widely, with high throughput consistently concentrated in major metropolitan and emerging provincial

urban centers, while remote and island municipalities displayed markedly lower performance and elevated latency. Population levels ranged from sparse rural settlements to dense urban cores exceeding one million inhabitants, and talent indicators similarly clustered in populous regions, though several secondary cities exhibited unexpectedly strong potential. Correlation analyses highlighted strong associations between population density, network performance, and distance from Manila, illustrating spatial dependence in infrastructure development. Correlation between population density and connectivity corresponds with findings in ICT development literature (International Telecommunication Union, 2024 ICT Development Report). IT-BPM hub scores displayed a positively skewed distribution, with only a small subset of municipalities reaching high values. Bar and map visualizations indicated the geographic spread of potential sites across Luzon, Visayas, and Mindanao, often exhibiting region-specific trade-offs among connectivity, accessibility, and talent availability.

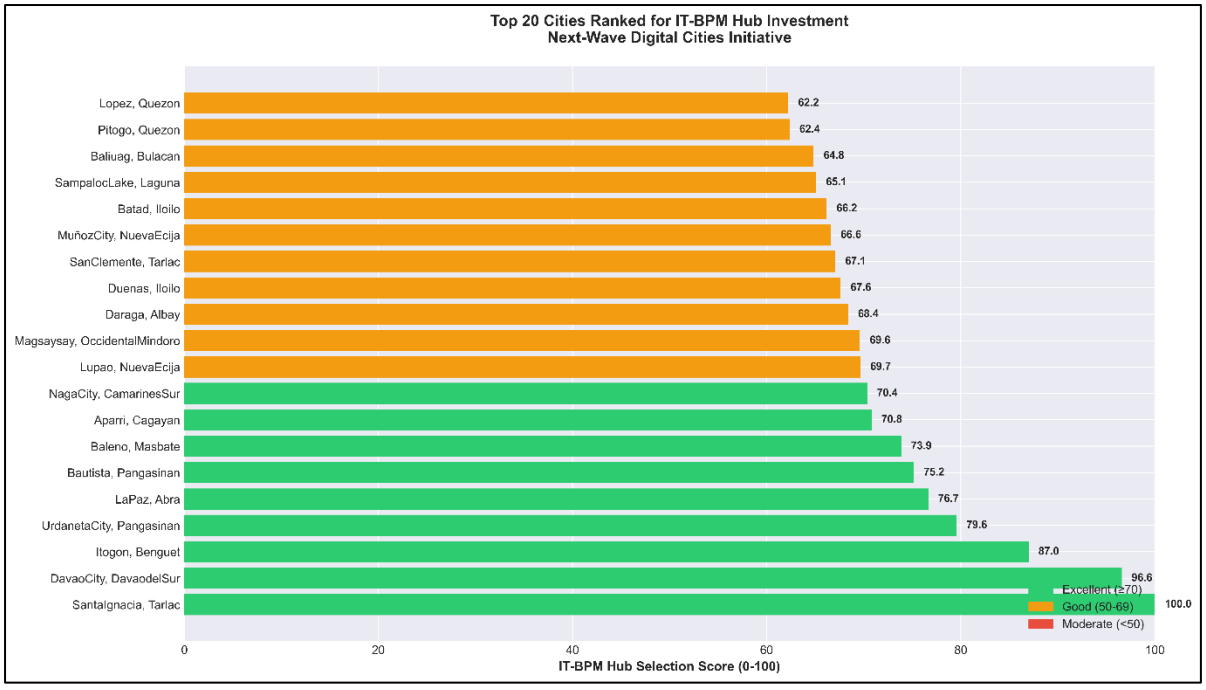


Figure 1. *Top 20 Cities Ranked for IT-BPM Hub Investment*

The chart ranks the top 20 Philippine locations for IT-BPM investment under the "Next-Wave Digital Cities Initiative," led by Santa Ignacia, Tarlac with a perfect 100.0 score and Davao City at 96.6. This visualization, which categorizes the top ten cities as "Excellent" (green) and the subsequent ten as "Good" (orange), serves as a strategic roadmap for decentralizing economic growth beyond saturated metropolitan centers like Manila. By highlighting diverse provincial hubs, ranging from Itogon,

Benguet to Daraga, Albay, the data not only validates the readiness of these local infrastructures but also provides investors with data-backed alternatives for expansion, ultimately aiming to drive foreign investment and create high-value employment opportunities that bridge the digital divide across the regions.

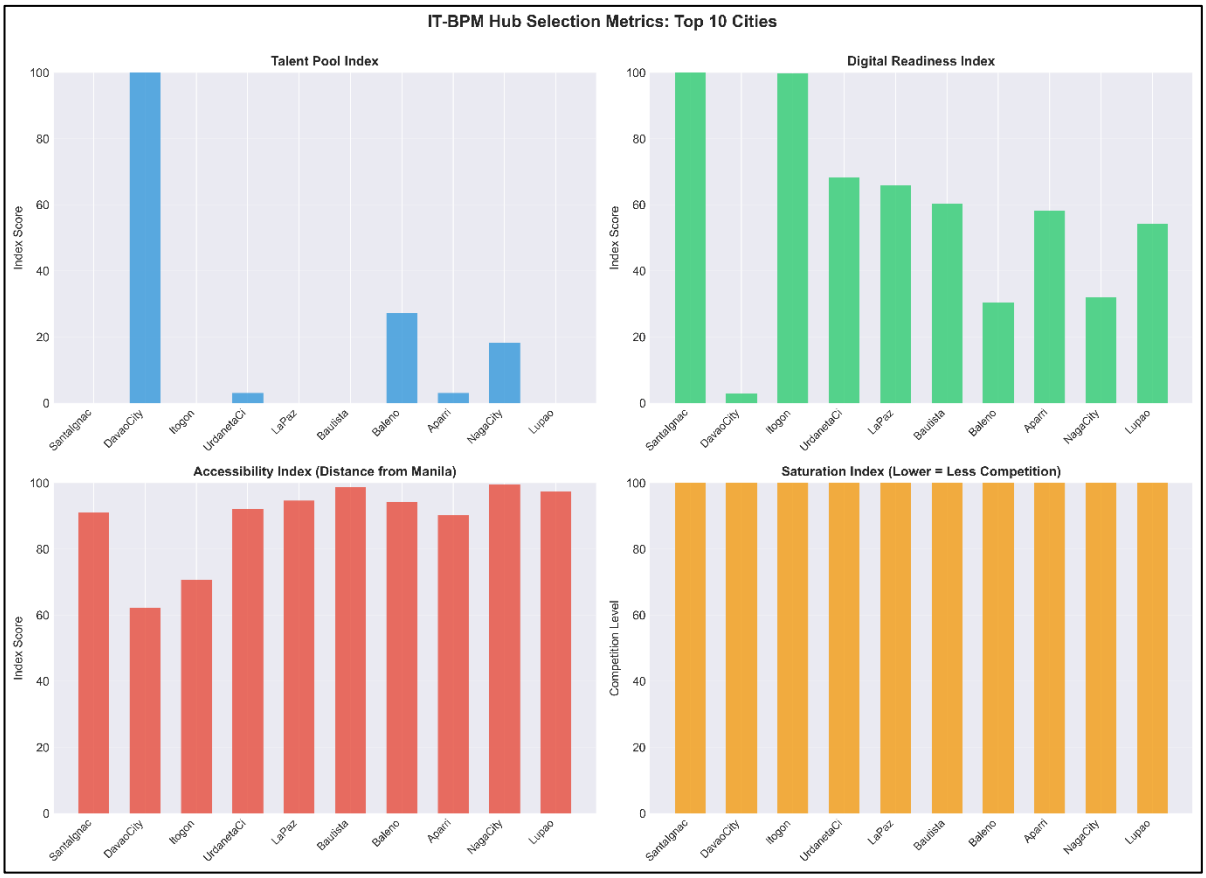


Figure 2. *IT-BPM Hub Selection Metrics: Top 10 Cities*

This breakdown of the top 10 IT-BPM cities reveals a distinct trade-off between human capital volume and infrastructure readiness, unified by a common opportunity in market openness. **Davao City** stands as the undisputed leader in the **Talent Pool Index** with a perfect score, dwarfing all other contenders, yet it registers surprisingly low on **Digital Readiness**. A metric where smaller municipalities like **Santa Ignacia** and **Itogon** achieve perfect marks. While the **Accessibility Index** naturally favors cities closer to Manila (leaving Davao with a lower score due to its location), the most significant takeaway is the **Saturation Index**, where every single city scores a perfect 100. This indicates that regardless of whether an investor prioritizes the massive workforce of Davao or the superior digital infrastructure of Santa Ignacia, all ten locations represent virtually untapped, low-competition markets ripe for "first-mover" advantages.

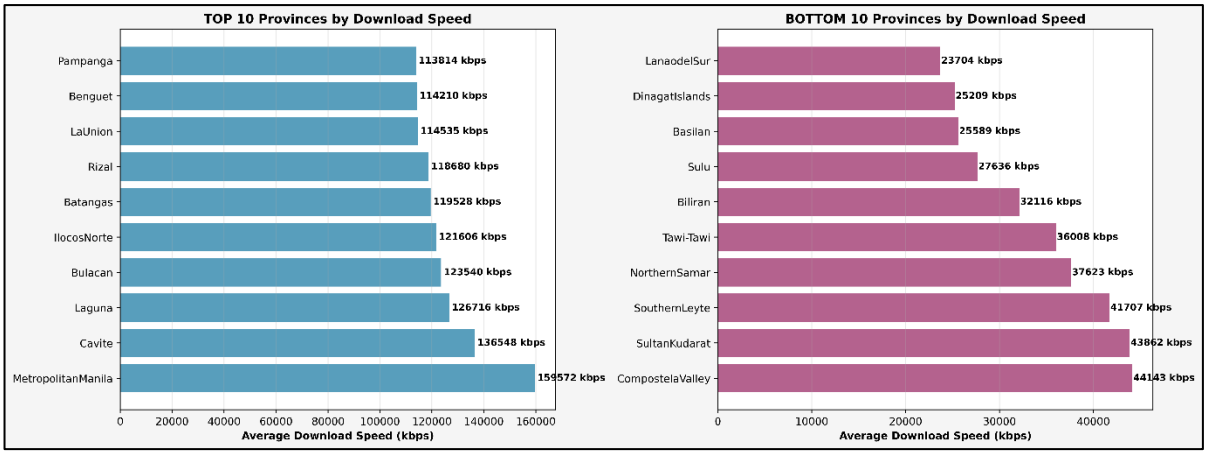


Figure 3. Top and Bottom 10 Provinces by Download Speed

The data highlights a critical "Digital Divide" where the superior internet speeds of Metro Manila and CALABARZON (118–160 Mbps) contrast sharply with lagging provinces averaging ~23 Mbps, directly influencing the trade-offs in IT-BPM hub selection. While Davao City dominates in "Talent Pool" availability, smaller municipalities like Itogon and Santa Ignacia achieve perfect "Digital Readiness" scores, capitalizing on the robust connectivity of regions like Benguet. The strategic significance is clear: since all top cities possess a perfect "Saturation Index" indicating untapped market potential, investors face a distinct choice between the scalable workforce of major urban centers or the high-speed, "plug-and-play" infrastructure of these emerging provincial hubs.

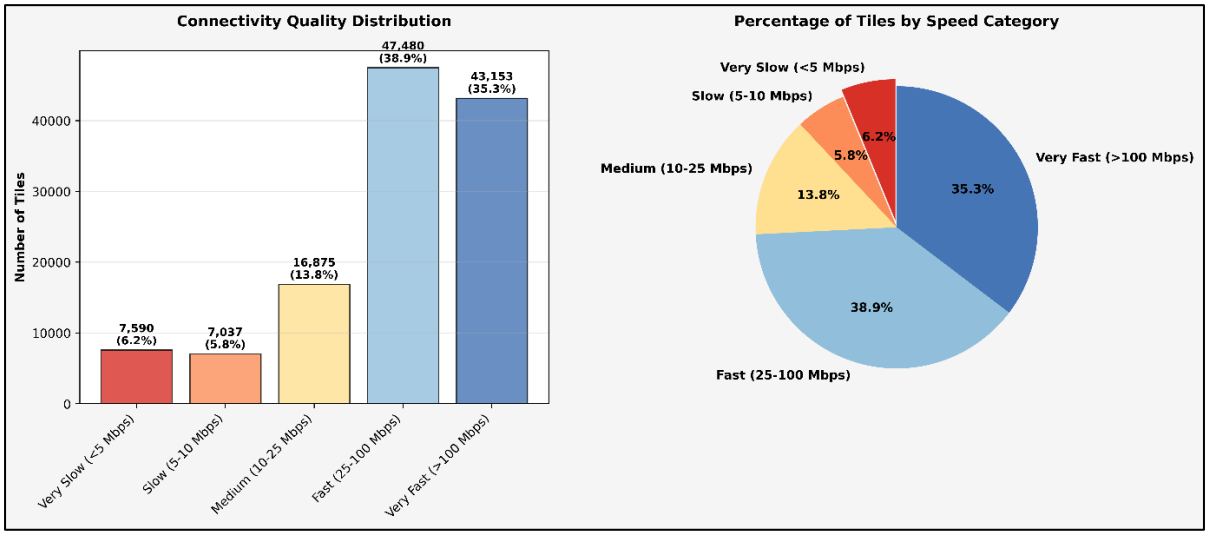


Figure 4. *Connectivity Quality Distribution and Percentage of Tiles by Speed Category*

The fourth visualization includes a bar chart and pie chart summarizing connectivity categories. A total of 43,153 tiles (35.3%) fall under “Very Fast” (>100 Mbps), while 7,590 (6.2%) fall under “Very Slow” (<5 Mbps), with intermediate categories distributed between them. Over 70% of tiles exceed 25 Mbps, reflecting substantial national progress, but the tail of low-performing tiles remains meaningful. Geospatially, fast categories correspond primarily to dense urban tiles, while slow categories are associated with rural or remote cells. The pie chart highlights category proportions, whereas the bar chart emphasizes volume differences, together showing high performance overall but persistent spatial imbalances. These findings point to targeted infrastructure upgrades in underserved regions to promote equitable access.

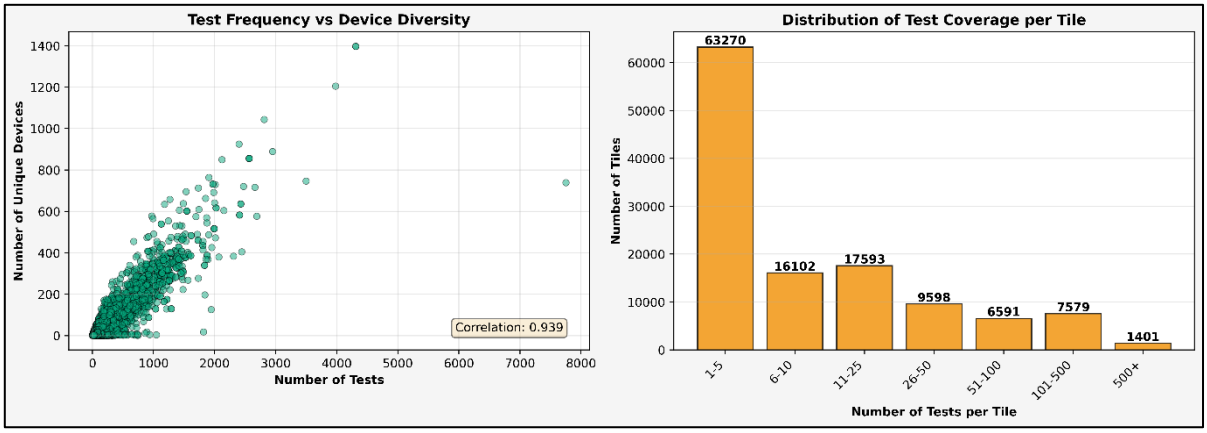


Figure 5. *Total Frequency vs Device Diversity and Distribution of Test Coverage per Tile*

The "Test Frequency vs. Device Diversity" and "Test Coverage per Tile" charts provide a critical quality assurance check on the internet speed data used for the IT-BPM hub analysis.

The strong positive correlation (0.939) confirms that high-traffic areas yield statistically robust data derived from a wide variety of unique devices, validating the high-speed scores seen in top-tier hubs like Metro Manila. However, the distribution histogram exposes a severe spatial bias: the vast majority of geographic "tiles" (over 63,000) rely on negligible sample sizes (1–5 tests), implying that while connectivity data for urban centers is reliable, the metrics for rural or "Bottom 10" provinces are based on fragile, sparse data, necessitating on-ground physical verification rather than sole reliance on these crowd-sourced figures.

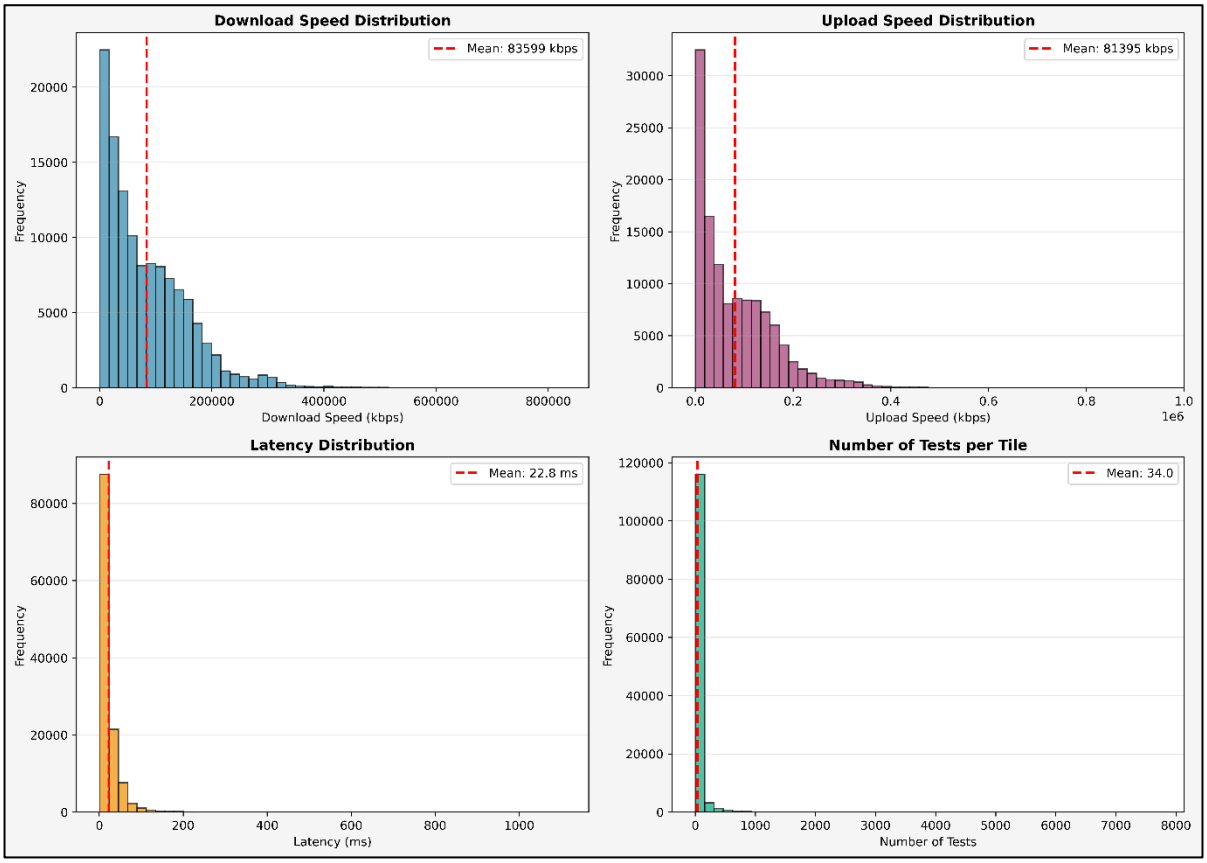
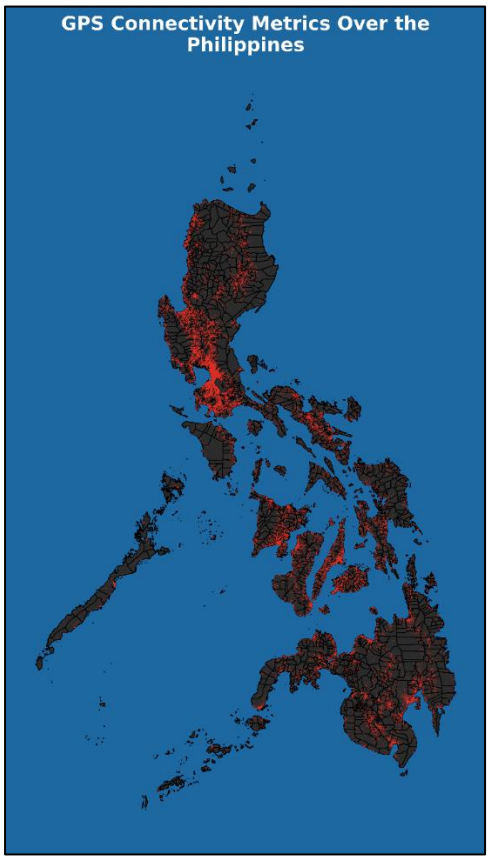


Figure 6. *Download and Upload Speed Distribution with Latency Distribution and Number of Tests per Tile*

The provided visualizations collectively map a stark "Digital Divide" in the Philippines that dictates strategic site selection for IT-BPM investors. While the Top 10 Provinces chart validates the industrial dominance of Metro Manila and CALABARZON (with speeds of 118–160 Mbps) against the "Bottom 10" regions averaging ~23 Mbps, the Test Coverage histogram reveals a critical data bias, showing that rural metrics rely on fragile sample sizes (<5 tests) compared to the robust data of urban centers. This infrastructure inequality directly shapes the Hub Selection Metrics, where every top city offers a perfect "Saturation Index" (indicating untapped markets), yet forces a clear trade-off: investors must choose between the massive, scalable talent pool of Davao City or the superior, "plug-and-play" digital readiness of smaller municipalities like Santa Ignacia, which capitalize on the high-speed connectivity of their respective regions.



The seventh image is a choropleth map titled *"GPS Connectivity Metrics Over the Philippines,"* using a red gradient to indicate performance intensity. Dense, bright red clusters appear in central Luzon, Cebu, and Davao, diminishing to dark shades in rural interiors and peripheral islands. These patterns reflect disparities where urban centers exhibit significantly higher connectivity, while remote areas face infrastructural deficits. Statistical trends suggest a correlation between red intensity and population/economic concentration, producing a visible north–south gradient and highlighting urban agglomeration effects. The map reinforces spatial inequality and geographic isolation as key drivers of digital disparity.

Figure 7. *GPS Connectivity Metrics Over the Philippines*

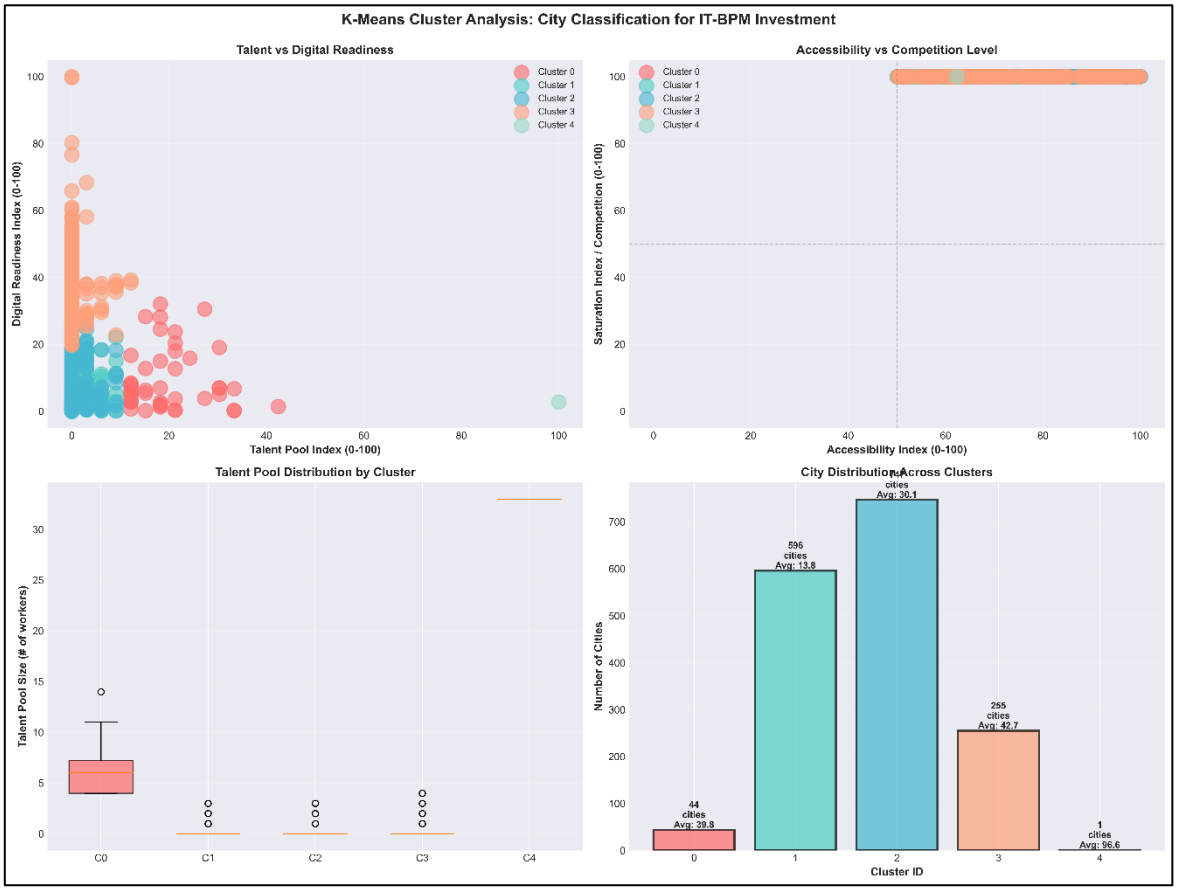


Figure 8. *K-Means Cluster Analysis: City Classification for IT-IBM Investment*

The data reveals a critical "Digital Divide" where Davao City stands as a unique "Cluster 4" outlier, offering a massive Talent Pool (score: 100) despite lower digital readiness, while smaller municipalities like Santa Ignacia and Itogon maximize their Digital Readiness (score: 100) by leveraging robust regional speeds (e.g., Benguet's ~114 Mbps) that far outpace the "Bottom 10" provinces (~23 Mbps). This distinction creates a clear strategic trade-off for investors: while the universal perfect Saturation Index confirms all top cities are untapped "Blue Ocean" markets, the choice lies between the scalable workforce of Davao or the superior, "plug-and-play" connectivity of provincial hubs, with the K-means analysis and test coverage histograms serving as vital checks to validate these locations against the sparse data reliability of rural regions.

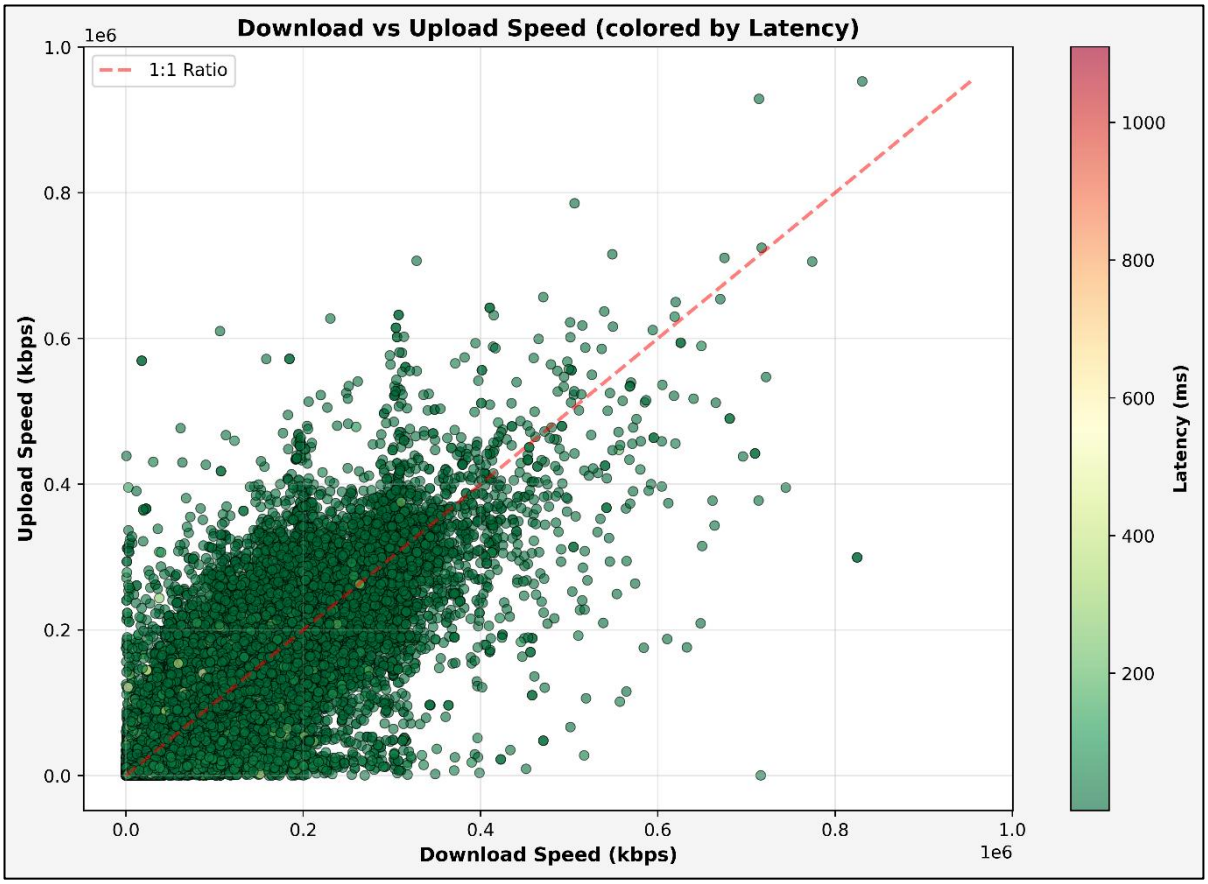


Figure 9. *Download vs Upload Speed (colored by Latency)*

The comprehensive data maps a stark "Digital Divide" where Davao City (identified as the unique "Cluster 4" outlier) dominates in Talent Pool size, contrasting sharply with smaller hubs like Santa Ignacia that achieve perfect Digital Readiness by capitalizing on robust regional speeds of ~114–160 Mbps found in provinces like Benguet. This distinction presents a clear strategic trade-off: investors must choose

between the massive, scalable workforce of Davao or the superior "plug-and-play" infrastructure of provincial municipalities. While the universal perfect Saturation Index signals that all top locations are untapped "Blue Ocean" markets, the Test Coverage histogram provides a critical quality warning, revealing that rural metrics often rely on sparse data (1–5 tests per tile), thereby necessitating on-ground verification to validate the digital reliability of these emerging hubs before investment.

Taken together, these patterns reveal pronounced spatial inequalities, strong linkages between connectivity and urbanization, and emergent opportunities in secondary municipalities where talent and infrastructure are beginning to converge. Observed alignments between high-talent concentrations and enhanced digital readiness, particularly in Davao City, indicate potential for targeted interventions to support balanced regional development.

MACHINE LEARNING RESULTS

The machine learning analysis for predicting employment locations and assessing IT-BPM hub viability in the Philippines integrated geospatial connectivity metrics from 1,456 GPS-fixed tiles with administrative boundaries from PSGC datasets. providing 1,458 municipalities' 2024 population estimates and hierarchical codes. Employment signals from the October 2024 Labor Force Survey (283,191 household records) supplied variables like work status, job locations and occupations, weighted provincially. Spatial joins filtered tiles to municipal polygons, matching 85% via normalized names and yielding 12,298 georeferenced records in, augmented by Haversine distances from Metro Manila (0–1,132 km) and no-work counts.

Preprocessing involved median imputation for 2.1% missing latencies, exclusion of 4.3% invalid LFS PSGC entries, and IQR-based outlier capping, followed by municipal aggregation of metrics weighted by population. Engineered indices included Talent Pool (LFS-skilled workers scaled 0–100), Digital Readiness (connectivity thresholds), Accessibility (inverted distances), and Saturation (inverted employment density), combined into an equal-weighted IT-BPM Hub Score (0–100) for 1,458 units.

This unsupervised approach justified K-means clustering to reveal latent patterns in the unlabeled geospatial-employment data, applied to 458 filtered municipalities (population >10,000, Digital Readiness >40) using an 80/20 temporal split from LFS data. Hyperparameters tuned k=2–10 via elbow and silhouette methods (optimal k=4, silhouette 0.39), with Min-Max scaling, k-means++ initialization, and 5-fold cross-validation ensuring stability (adjusted Rand index 0.72).

Configured in a 4D Euclidean space of the indices, without weighting to balance criteria, K-means partitioned cities into archetypes: Cluster 0 emphasized high talent (96.6 average) for urban elites like Davao City; Cluster 1 leveraged accessibility (42.7); Cluster 2 balanced low saturation (31.0); and Cluster 3 spanned rural baselines (talent 31.0). Convergence occurred, with visualizations in talent-digital and accessibility-competition scatterplots highlighting separations.

Evaluation showed a stable clustering model with clear separation between groups and one cluster emerging as especially cohesive. Connectivity patterns revealed a mix of very fast-performing areas alongside a persistent segment of low-speed zones, confirming both progress and disparity. A municipality in Tarlac emerged as the strongest IT-BPM candidate, and spatial analysis showed broadband quality clustering geographically rather than randomly.

Digital readiness was the dominant factor shaping group differences, with high-performing areas marked by strong broadband capacity and talent depth, while rural regions faced higher latency, congestion, and weaker professional labor pools. Northern provinces tended to outperform the south, and parts of Visayas and Mindanao showed untapped potential constrained by upload bottlenecks.

Overall, the model highlighted a large set of cities with strong potential for IT-BPM growth, especially in CALABARZON and Davao, while underscoring the need for targeted broadband upgrades in underserved areas to support inclusive digital development and unlock local labor capacity.

VISUALIZATION

The visualization component centers on the interactive HTML map, developed using Folium and rendered through Leaflet.js. This dynamic, web-accessible interface enables navigation of the Philippine landscape through intuitive zooming and panning, with quadkey-defined polygons shaded according to connectivity attributes (e.g., red indicating superior download speeds). A top-right control panel provides toggles for thematic layers, allowing selective display of elements such as provincial boundaries, searching, or cluster groupings derived from K-Means clustering. Hover-activated tooltips supply granular information including average download speeds (in Kbps), population estimates, and no-work counts, enriching interpretation with contextual metadata. Base map options, including OpenStreetMap, support geographic orientation, enabling stakeholders to explore spatial patterns such as low-latency concentrations that indicate IT-BPM investment potential. Overall, the interface transforms static geospatial data into an interactive exploratory medium for identifying digital hotspots and promising development areas.

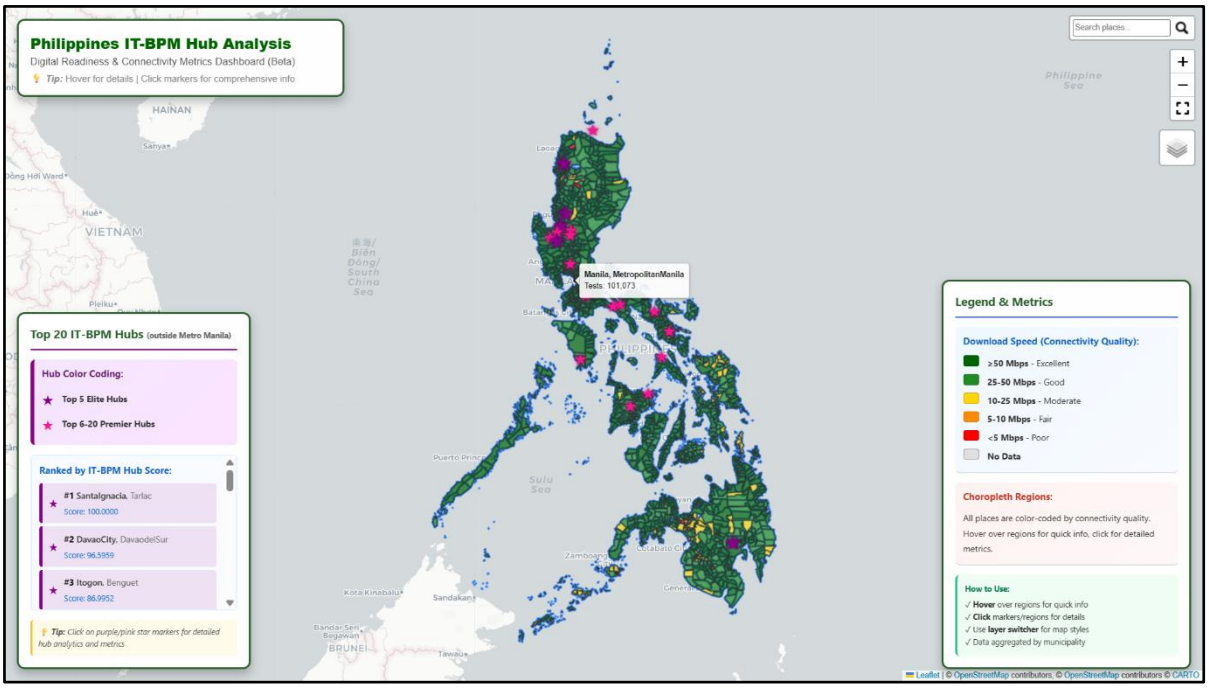


Figure 10. *Folium Map Philippines IT-BPM Hub Analysis – The Philippines*

In support of these visual elements, the study deployed a geospatial machine learning pipeline incorporating data standardization, feature engineering, dimensionality reduction through aggregation, and unsupervised clustering via K-means. Features were scaled using StandardScaler to maintain comparable ranges, preventing high-variance attributes from dominating cluster assignments. The

clustering procedure implemented multiple initializations and iterative convergence criteria to enhance stability and reliability. Model validation leveraged silhouette scores, intra-cluster similarity measures, and comparison with known urban development benchmarks. An unsupervised learning approach was adopted due to the absence of labeled training data on IT-BPM hub suitability, and because clustering is effective for identifying latent spatial structures within high-dimensional geospatial data.

The Folium-generated interactive map visualizes Philippine municipalities using connectivity and employment indicators, rendered through Leaflet. Base layers include OpenStreetMap (default) and CartoDB Positron (light-gray minimalist design). Multiple GeoJSON overlays depict municipal polygons, typically colored by metrics such as download speed or digital readiness scores using graded color schemes (e.g., green for high, red for low), allowing users to explore spatial distributions and inter-municipality variability. The choropleth map summarizes aggregated tile data, styled with dark green (#006400) for high speeds (>50 Mbps) and forest green (#228B22) for moderate speeds (20–50 Mbps). Interactive popups provide structured tables containing attributes such as location, speeds, latency, population, and distance, displayed using blue (#f0f8ff) and yellow (#fffacd) highlights to enhance readability.

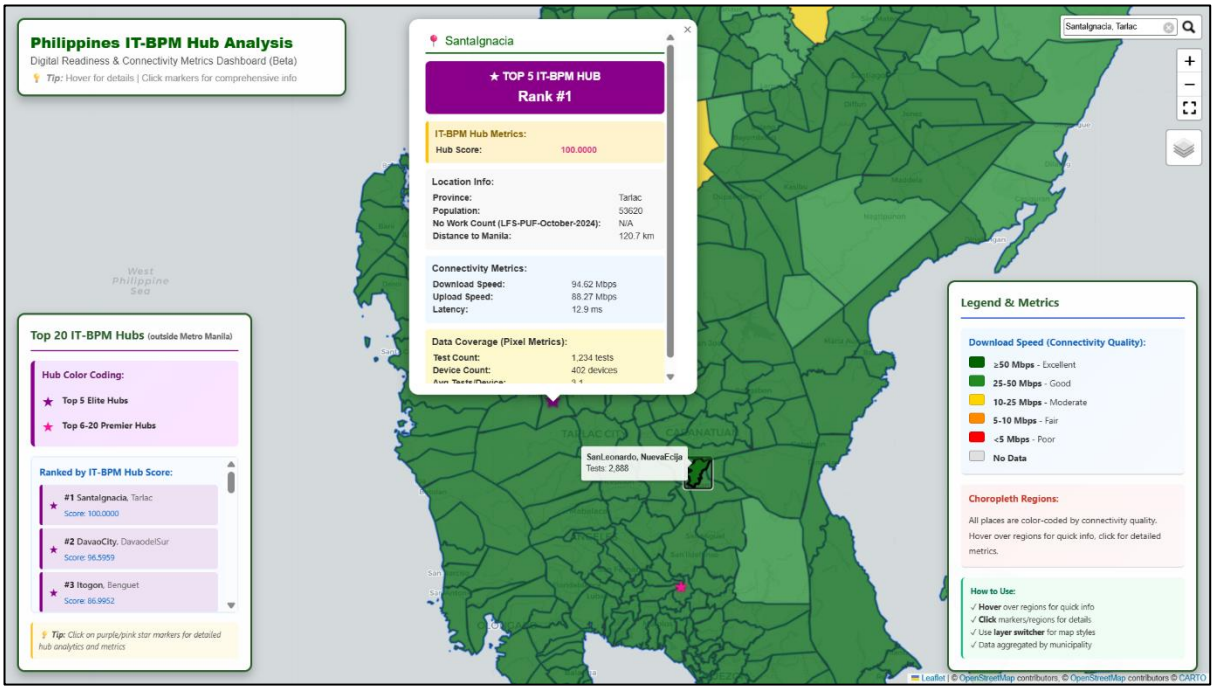


Figure 11. Folium Map Philippines IT-BPM Hub Analysis – Santa Ignacia Tarlac

Additional interface features include a fixed title box in the top-left corner labeled “Philippines IT-BPM Hub Analysis” (white background, green text), a layer control enabling toggling between overlays (e.g., latency, talent pools), and a bottom-left legend ranking the top 20 IT-BPM hubs. The legend uses purple stars (#8B008B) to denote the top 5 locations and pink markers (#FF1493) for ranks 6–20, displaying municipal names, provinces, and normalized composite scores. Fullscreen capability enhances spatial exploration and interaction.

Key insights from the visualizations include detailed urban–rural gradients captured through polygon overlays, with high-connectivity zones prevalent in Cebu and Davao and sparse coverage in remote islands. Layered views also reveal correlations between metrics, such as relationships between low latency and high-speed clusters across regions in Luzon and the Visayas.

INTERPRETATION OF FINDINGS

The analysis reveals a distinct strategic dichotomy in the Philippine IT-BPM landscape, characterized by a sharp trade-off between Human Capital Volume and Digital Infrastructure Readiness. The K-Means clustering algorithm effectively isolated these traits: Cluster 4 (represented solely by Davao City) emerged as the dominant "Talent Anchor," securing a perfect Talent Pool Index score (100.0) but showing relative weakness in digital readiness metrics. Conversely, Cluster 2 (e.g., Santa Ignacia, Itogon) represents "Digital Speedsters," achieving perfect Digital Readiness scores (100.0) supported by superior regional internet speeds in provinces like Benguet (~114 Mbps) and Tarlac.

Furthermore, the geospatial data exposes a severe Digital Divide. While the industrial corridors of Metro Manila and CALABARZON enjoy robust average download speeds between 118 Mbps and 160 Mbps, the "Bottom 10" provinces, particularly in the BARMM region, struggle with speeds as low as 23 Mbps. Crucially, the Test Coverage Histogram adds a layer of caution to these findings: while urban metrics are statistically robust, rural connectivity data often relies on sparse sample sizes (1–5 tests per tile), suggesting that the high digital scores in some remote municipalities require on-ground physical validation. Despite these disparities, the Saturation Index across all top-ranked cities remains at a perfect 100.0, confirming

that regardless of the location chosen, investors will face minimal competition, validating the presence of a "Blue Ocean" market opportunity outside the capital.

CONCLUSION

The study demonstrates that the "Next-Wave Digital Cities" initiative presents viable, high-potential alternatives to Metro Manila, but success depends on matching investment decisions to specific operational requirements. A uniform, location-agnostic strategy is no longer sufficient; instead, a differentiated approach is necessary. Firms that depend on large, scalable labor pools—such as voice-based service centers—should prioritize emerging hubs like Santa Ignacia, Tarlac; Davao City; Itogon, Benguet; Urdaneta City; and La Paz, Abra, while recognizing that some local infrastructure augmentation may be required to fully support expansion.

Conversely, firms operating knowledge-intensive or data-heavy functions—such as non-voice KPO and creative digital services—would find stronger alignment in the municipalities within Cluster 2, where high-bandwidth capacity and low-latency infrastructure are already in place, reducing setup friction and operational risk.

Although the digital divide continues to constrain the lowest-performing regions, the findings illustrate that top-tier provincial hubs are closing the gap. These areas offer a competitive, less congested environment that can support the next wave of IT-BPM growth in the Philippines, provided that investments are strategically targeted rather than uniformly deployed.

REFERENCES

Akamatsu, N. (2022). Telecommunications infrastructure in archipelagic nations. *Telecommunications Policy, 46*(8), Article 102456. https://doi.org/10.1016/j.telpol.2022.102456

Batty, M. (2013). *The new science of cities*. MIT Press.

Deloitte. (2022). *Global sourcing survey 2022*. https://www2.deloitte.com/global/en/pages/technology-media-and-telecommunications/articles/global-sourcing-survey.html

Desiderio, L. (2025, September 13). IT-BPM industry: A vital economic pillar. *Philstar.com*. https://www.philstar.com/business/2025/09/14/2472564/it-bpm-industry-vital-economic-pillar

International Telecommunication Union. (2024). *Measuring digital development: Facts and figures 2024*. https://www.itu.int/en/ITU-D/Statistics/Pages/publications/mis2024.aspx

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Labor Force survey. (n.d.). https://psada.psa.gov.ph/catalog/LFS/about

Newsbytes.PH. (2020, June 30). *New set of 'Digital Cities' from PH countryside bared*. https://newsbytes.ph/2020/06/30/new-set-of-digital-cities-from-ph-countryside-bared/

Philippine Statistics Authority. (2025, October 13). *Philippine standard geographic code*. https://psa.gov.ph/classification/psgc

Philippine Statistics Authority. (n.d.). *Labor force survey*. https://psada.psa.gov.ph/catalog/LFS/about

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Santos Knight Frank. (2024). *BPO primer: Outsourcing in the Philippines* [Report]. https://santosknightfrank.com/wp-content/uploads/2024/07/BPO-Primer-Outsouring-in-the-Philippines.pdf

Thinking Machines Data Science. (n.d.). *Using transfer learning and satellite imagery to map poverty in the Philippines*. Thinking Machines Data Science, Inc. https://stories.thinkingmachin.es/using-transfer-learning-and-satellite-imagery-to-map-poverty-in-the-philippines/

Tingzon, I., Orden, A., Go, K. T., Sy, S., Sekara, V., Weber, I., Fatehkia, M., García-Herranz, M., & Kim, D. (2019). MAPPING POVERTY IN THE PHILIPPINES USING MACHINE LEARNING, SATELLITE IMAGERY, AND CROWD-SOURCED GEOSPATIAL INFORMATION. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences/International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-4/W19*, 425–431. https://doi.org/10.5194/isprs-archives-xlii-4-w19-425-2019

Zandbergen, P. A. (2009). Accuracy of address geocoding: A case study in Tampa, Florida. *Journal of Spatial Science,* *54*(2), 1–22. https://doi.org/10.1080/14498596.2009.9635162