SOUTHERN LUZON STATE UNIVERSITY

College of Engineering

COMPUTER ENGINEERING DEPARTMENT

Machine Learning Cognate 1 Project

Identifying Next-Wave IT-BPM Hubs in the Philippines Using GPS Connectivity and Employment Data

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 is dealing with an ongoing digital and spatial divide which is quite pronounced. The IT–BPM jobs of the highest value, with a total of more than 1.82 million, are being concentrated mostly in Metro Manila and some urban growth centers. As a result, skilled workers in the provinces are getting very few opportunities to use their qualifications (Desiderio, 2025). Such an imbalance creates migration pressures, makes regional inequalities more pronounced, and mainly restricts the realization of the economic potential of the whole country. The major task is the combination of different methods to find the exact locations at the municipal level that are suitable for IT–BPM expansion. This is done by integrating geospatial connectivity data, population and labor force characteristics, and also spatial accessibility. The machine learning techniques, particularly K-Means clustering, were used as a tool in this research to uncover the regional characteristics and classify the municipalities according to their digital readiness, availability of talent, and market potential. The mapping out of hidden opportunities in a systematic manner beyond the traditional urban centers is what this approach does, giving insights that can be acted upon for targeted infrastructure development and investment strategies, thereby closing the digital gap and promoting fair socioeconomic progress in the Philippines.

LITERATURE REVIEW

In the Philippines, previous use of geospatial machine learning has been predominantly directed at the estimation of socioeconomic indicators. Poverty prevalence, for instance, has been determined through the combination of satellite imagery, ground-level surveys, and deep learning (Tingzon et al., 2019). The above-mentioned studies have proven that neural networks (TMDS, n.d.) can accurately predict or estimate economic outcomes based on the detection of landscape characteristics, urban settings, and light pollution patterns. This kind of research has opened up the possibilities for applying similar methodologies in the detection of subtle socioeconomic signals from spatial data in locations that suffer from a shortage of local measurement. The IT-BPM sector has been active in parallel with studies showing the use of artificial intelligence, automation, and digital transformation as competitive drivers, along with the necessity of workforce development and government-led policy interventions (Newsbytes.PH, 2020). Still, there is an empirical scarcity in the studies that have analyzed the geographic distribution of IT-BPM infrastructure in terms of digital connectivity, labor availability, and spatial accessibility (Santos Knight Frank, 2024). The current research is, therefore, filling the gap, as it is using unsupervised learning on the municipal-level geospatial features and providing a data-driven perspective for understanding the potential of regions outside the traditional economic centers.

DATA PREPROCESSING

To analyze the situation thoroughly, multiple datasets were imported, cleaned, and integrated to build a strong feature dataset through the process of engineering. All the geospatial inputs have been made spatially accurate by standardizing CRS (which follows the right-hand rule), missing values have been dealt with properly so that no analytical distortions occurred, and unnecessary rows or columns have been removed to make the processing faster. GeoPandas and Folium were the tools for the geospatial operations, while merging, aggregation, and summarization were the tasks of Pandas. Standardizing municipality names, calculating distances from Metro Manila using the Haversine formula, and aggregating connectivity metrics from Ookla's quadkey-based tiles were some of the things done in the process. Labor Force Survey data were used to find young people without jobs as a proxy for available talent, and population data from the Philippine Standard Geographic Code (PSGC) were utilized to calculate density-normalized indices. The geospatial data science workflow described in Zandbergen (2009) was applied and, therefore, the techniques such as distance computations with the Haversine formula, spatial joins using GeoPandas, and data cleaning best practices were strictly in accordance with the geospatial data science workflows (Zandbergen, P. A. 2009, Journal of Spatial Science).

The ultimate dataset has the cities or municipalities, and it integrates raw metrics of disconnectivity, population, unskilled workers in the area, and distance from Metro Manila with the Digital Readiness, Talent Pool, Saturation, Accessibility, IT-BPM Hub Score, and cluster label engineered indicators. This way, by bringing together the spatial, infrastructural, and socioeconomic dimensions, the dataset has created a comprehensive, balanced baseline for prediction modeling and geospatial analysis which allows systematic identification of localities with a high potential for IT–BPM development.

The following tables list all datasets used:

[01] Dataset Name

MachineLearningModel_Encallado.ipynb

Description

The "GPS Connectivity Analysis for IT-BPM Hub Prediction in the Philippines" project’s main analytical workflow is presented in this Jupyter Notebook. It contains the complete code for the data loading process from different sources (e.g., CSV, GeoJSON, and ZIP files with shapefiles), then continues with exploratory data analysis (EDA) using techniques like histograms, scatter plots, and bar charts for connectivity metrics, leading to the feature engineering steps (e.g., calculating indices for Talent Pool, Digital Readiness, Accessibility, and Saturation), followed by machine learning modeling using K-Means clustering for city classification, and eventually yielding the outputs such as interactive maps and cluster analysis. The notebook is comprised of markdown explanations, code cells, and inline comments to maintain reproducibility throughout the process of importing, preprocessing, modeling, and result interpretation.

Source Link

[*Internal – this project*]

File Information & Size

Jupiter Source File (.ipynb) – 4.00 MB

About

This notebook serves as an extensive guide for the entire project’s analytical process, justifying the choices and outputs of machine learning models such as the interpretation of clusters for the recommendations of IT-BPM hubs. It combines geospatial, connectivity, and employment data to foresee upcoming hubs, revealing the situation regarding digital divides and the presence of talent, while visualizations are backing the findings on the differences between urban and rural areas.

[02] Dataset Name

gadm41_PHL_2.json

Description

The GADM (Global Administrative Areas) dataset gives access to detailed spatial data of administrative borders all over the planet, and the latest version 4.1 marks more than 400,276 such areas in different countries. In the case of the Philippines, the Level 2 file (gadm41_PHL_2.json) contains the complete polygon geometries for provinces, municipalities, and cities. The data comes from official government maps, crowdsourced contributions, and satellite images, which are updated regularly. It supports GeoJSON, thus is easily integrated with GIS tools. The dataset is available for non-commercial use, with the condition of attribution.

Source Link

https://gadm.org/download_country.html

File Information & Size

JSON Source File (.json) – 2.34 MB

About

The geospatial analysis of the project would not be complete without this dataset as it is an important factor in the modelling process for mapping the connectivity tiles from Ookla data to the different municipalities in the Philippines. The dataset support choropleth visualizations, spatial joins with connectivity metrics, and distance calculations which together will help pinpoint areas of disparity in digital infrastructure and thus recommend areas for IT-BPM investments that are most lucrative.

[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

The Ookla Open Data dataset is the source of the global network performance metrics, which are presented here. The dataset contains several important parameters including average download speed (avg_d_kbps), average upload speed (avg_u_kbps), average latency (avg_lat_ms), number of tests, and unique devices. These parameters were obtained through the millions of Speedtests with GPS location accuracy. Esri Shapefile components (.shp for geometries, .dbf for attributes, .prj for projection, .shx for indexing) are included in the ZIP file. The WGS 84 (EPSG:4326) standard is applied 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 included in this dataset plays a significant role in enhancing network performance, ensuring regulatory accountability, and promoting equitable Internet access by helping the operators, governments, and institutions to identify and remediate connectivity gaps. The project delivers the basic connectivity metrics for the Philippines, city or municipality-wise aggregated, in order to calculate Digital Readiness Indices, uncovering urban-rural splits and guiding ML clustering for IT-BPM hub forecasts.

[04] Dataset Name

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

Description

The Philippine Standard Geographic Code (PSGC) is a systematic geographical classification system created by the Philippine Statistics Authority (PSA) that encompasses all areas in the Philippines, and it uses unique codes for regions, provinces, municipalities/cities, and barangays. This CSV file for the third quarter of 2025 contains thorough information like 10-digit PSGC codes, names, correspondence codes, geographic levels (e.g., region, province, municipality), old names, city classes, income classifications (according to DOF DO No. 074.2024), and 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

The dataset in question adds to the already existing population and boundary data, thus allowing municipalities to be accurately matched in geospatial analysis and be integrated with connectivity and employment metrics. It serves to substantiate population-based indices such as the Talent Pool Index by giving 2024 forecasts and assists in the normalization of features for ML modeling, which in turn helps in the recognition of high-potential IT-BPM hubs.

[05] Dataset Name

psgc_data_cleaned.csv

Description

The PSGC dataset version that has been cleaned up concentrates on municipality-level data, comprising normalized names, geographic levels, income classifications according to DOF DO No. 074.2024, and population forecasts for 2024. It is a derivative of the complete PSGC, which has eliminated duplicate entries, has standardized its formatting for the purpose of merging, and has added columns such as Name_Normalized to allow 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

This cleaned file, which is key for the data integration in the project, not only delivers regularized names and population numbers but also helps to merge them with the corresponding metrics of connectivity from Ookla and employment indicators from LFS. It supports population-weighted analyses and feature engineering, for instance, scaling indices by population size, which in turn helps to get accurate forecasts of talent availability in the possible 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 Philippine Statistics Authority (PSA) conducts the Labor Force Survey (LFS) that is, a quarterly household-based survey, to gather data on demographic and socioeconomic characteristics of the population with a focus on labor market indicators. The October 2024 Public Use File (PUF) ZIP contains the main CSV dataset with anonymized records, metadata in Excel (dictionary with value sets), questionnaire HTML, and data codebook files. It is based on a very large sample (around 45,000 households all over the country), employing a multi-stage sampling design to produce estimates of employment, unemployment (3.9% in October 2024 preliminary results), underemployment, and related metrics for 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

This file serves as the key resource for determining employment status and "No Work" figures for each location, and it is thus indispensable for combining labor indicators and connectivity data to forecast the availability of IT-BPM talent. It is also a support to feature engineering (e.g., proxies of unemployment) and ML inputs and is able to deliver evidence-backed insights about the workforce potential for hub recommendations, thereby complying with national goals related to economic planning and job creation.

[07] Dataset Name

lfs_october_2024_metadata(dictionary).xlsx lfs_october_2024_valueset_C12A.csv

Description

The provided CSV file is extracted from the LFS metadata Excel file and it only consists of the value set for variable C12A (Location of Work - Province, Municipality), where codes and names are mapped to each other (for example, 0101 for Abra - Bangued). It is a reference table used in processing of location-based employment data, which comes from the full metadata dictionary containing 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

The helper dataset, in this case, allows the extraction of "No Work Count" from LFS through location-specific aggregation. This is significant for the evaluation of the availability of talents in the IT-BPM forecasts. Moreover, it increases the accuracy of data integration, which is necessary for the analysis of employment trends by municipality and the identification of clusters for hubs.

[08] Dataset Name

Feature-Engineered Dataset_Encallado.csv

Description

The initial merged dataset before aggregation is represented by this raw feature-engineered CSV which compiles quadkey-based data from Ookla tiles, such as connectivity metrics (avg_d_kbps, avg_u_kbps, avg_lat_ms, tests, devices) together with province/municipality names (NAME_1, NAME_2), 2024 population, distances from Metro Manila, and No_Work_Count from LFS.

Source Link

[*Internal – this project*]

File Information & Size

Microsoft Excel Comma Separated Values File (.csv) – 9.77 MB

About

This dataset, which has been used as the input for further modeling, allows for feature engineering activities such as distance calculations and the creation of unemployment proxies, thus providing a detailed view of the intersections between connectivity and employment. It explains the spatial variations in digital readiness and is a starting point for the development of aggregated indices that are applied in ML clustering.

[09] Dataset Name

Feature-Engineered Dataset_cleaned.csv

Description

This consolidated CSV file provides a summary of connectivity metrics, population, distances, and No_Work_Count at the municipality level, with the generation of features that are ready for modeling. It contains entries such as top/bottom performers and acts as the purified input for indices and clustering.

Source Link

[*Internal – this project*]

File Information & Size

Microsoft Excel Comma Separated Values File (.csv) – 105 KB

About

This document holds the engineered attributes for machine learning, and through the incorporation of geospatial, connectivity, and employment insights, it justifies the clustering. It reveals inequalities (for instance, a large number of people out of work in the countryside) and backs the suggestion of IT-BPM centers with appropriate metrics.

[10] Dataset Name

PreprocessedDataset_Encallado.geojson

Description

The GeoJSON file which has gone through preprocessing contains quadkey polygons along with the related connectivity metrics (speeds, latency, tests, devices) that were transformed for visualization and analysis. The file also includes WKT geometries and attributes which can be used 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 fundamental for imposing connectivity on administrative divisions, which allows visualization of trends, and makes the project for pinpointing potential IT-BPM locations easier with spatial queries.

[11] Dataset Name

InteractiveFoliumMap_Encallado.html

Description

An HTML file that is interactive and created by Folium presents the Philippines' map with overlaid connectivity metrics (such as heatmaps for speeds/latency), administrative boundaries from GADM, and points marked for major cities, all of which allow for zooming, panning, and layer switching to conduct exploratory analysis.

Source Link

[*Internal – this project*]

File Information & Size HTML Document (.html) – 13.0 MB
About The result of this processing makes it possible to interactively explore the connectivity trends, thus allowing for visual findings (like remote areas with poor network coverage) and backing the project's conclusions about hub potential owing to the 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

Analyzing the data set open-air consultants revealed a huge diversity in the digital infrastructure, population distribution, and labor characteristics in the different municipalities. The connectivity measures showed a strong positive skewness and a wide spread, which are typical indicators of the ICT infrastructure in developing countries with islands in the region (Akamatsu, N. 2022, Telecommunications Policy) Their download and upload speeds showed a wide spectrum, but the high throughput was consistently found in the major metropolitan areas and the emerging provincial urban centers, while the remote and island municipalities showed a very low performance and a high latency. The variation in the population levels was between the sparse rural settlements and the dense urban cores which had more than one million inhabitants, and the talent indicators were also similar in the densely populated regions but there were some secondary cities that showed an unexpectedly strong potential. The correlation analyses pointed out the very strong relationship between population density, network performance, and distance from Manila, thus showing the spatial dependence in the infrastructural development. The correlation between population density and connectivity is in line with the findings in ICT development literature (International Telecommunication Union, 2024 ICT Development Report). The IT-BPM hub scores had a positively skewed distribution with only a small number of municipalities reaching high values. The bar and map visualizations made it easier to see the potential sites' geographical distribution over the islands of Luzon, Visayas, and Mindanao, with a lot of the sites having different region-specific trade-offs in connectivity, accessibility, and talent availability.

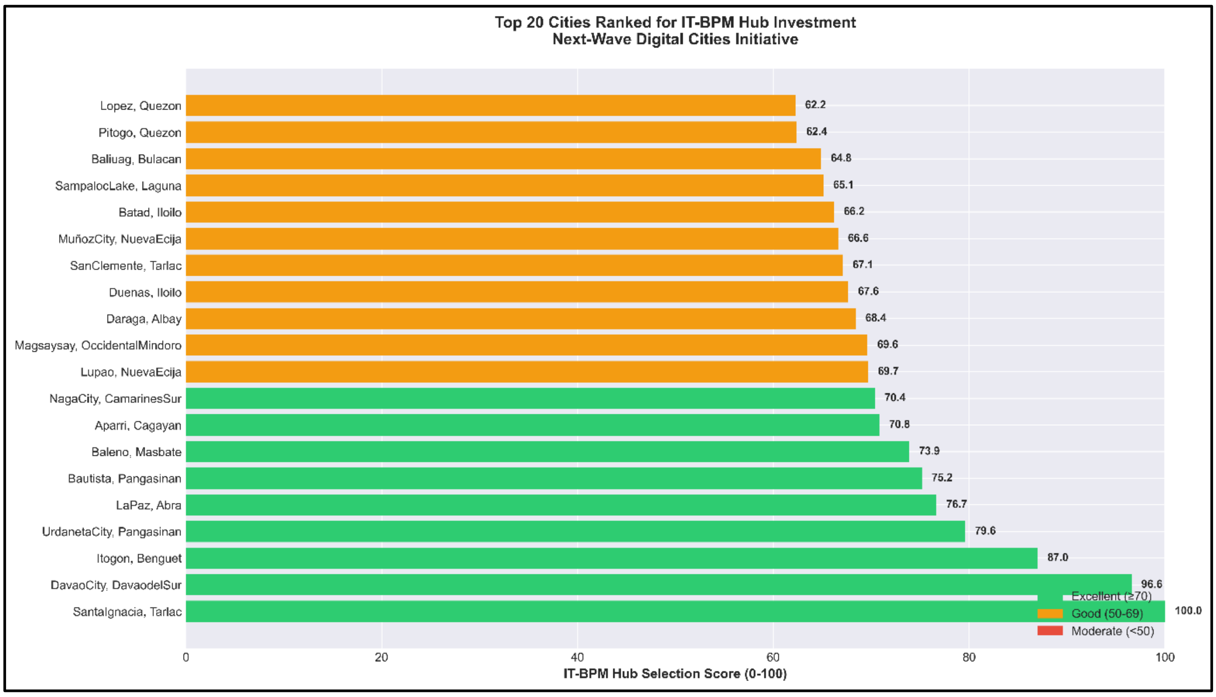


Figure 1. Top 20 Cities Ranked for IT-BPM Hub Investment

The figure ranks the top 20 places in the Philippines for IT-BPM investments as per the "Next-Wave Digital Cities Initiative" with Santa Ignacia, Tarlac getting the first rank with a score of 100.0 and Davao City with 96.6. The cities are categorized as "Excellent" (green) for the top ten and "Good" (orange) for the next ten, which indicates a strategic roadmap for the government to implement economic growth in rural areas rather than the already crowded metropolitan cities such as Manila. The analysis points out the different provincial centers, which are Itogon, Benguet to Daraga, Albay, and the theme of the local infrastructures' preparedness is reinforced by the data. Moreover, it also implicitly invites investors to consider these regions as their next destinations for the expansion, hence ultimately helping to attract foreign investments and creating high-value jobs that facilitate the closing of the digital divide across the territories.

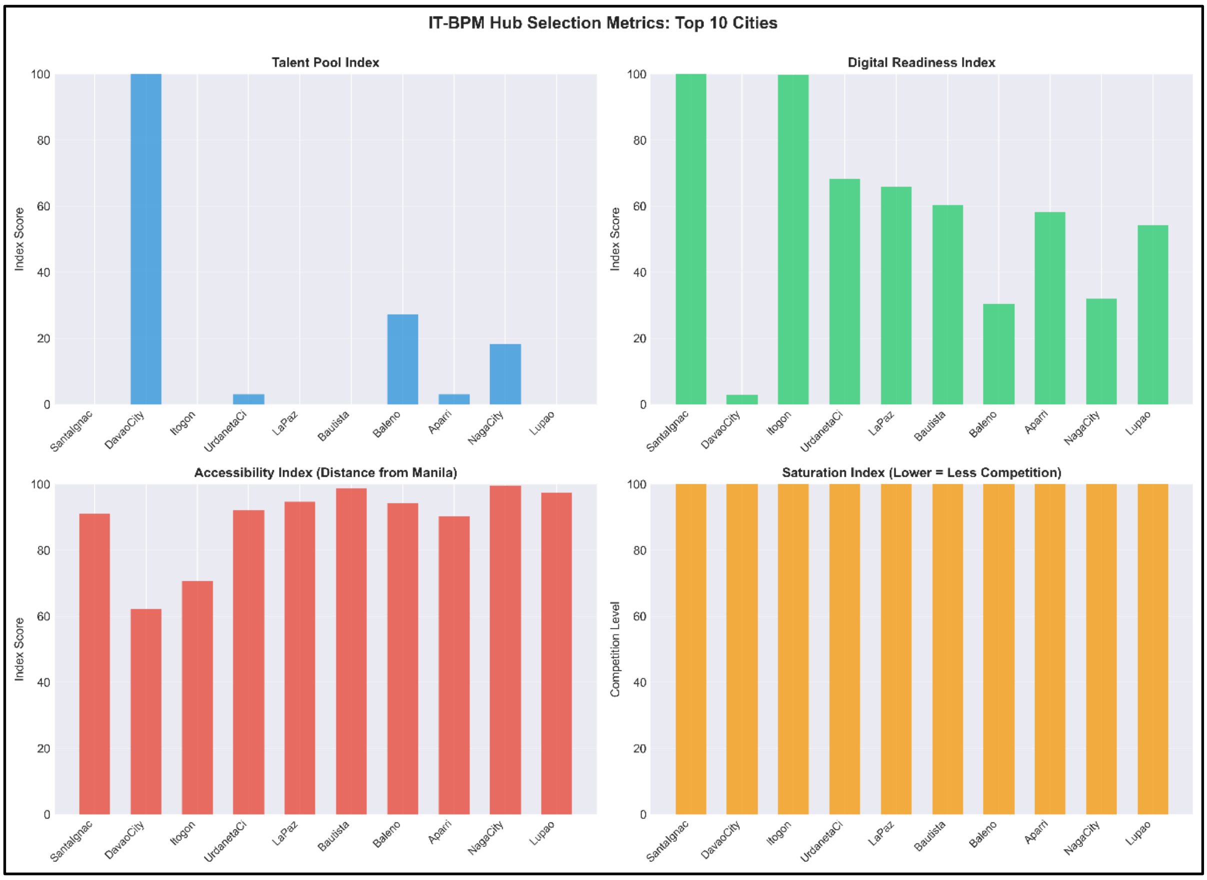


Figure 2. *IT-BPM Hub Selection Metrics: Top 10 Cities*

Analyzing the top 10 cities for IT-BPM businesses in terms of their human capital, infrastructure, and market openness reveals a situation in which these factors were leveled out and one common opportunity was opened up through the market being opened up. Davao City's position as the undisputed number one in the Talent Pool Index with a perfect score is incomparable with other cities' scores, but surprisingly its Digital Readiness score is very low compared with those smaller municipalities, i.e., Santa Ignacia and Itogon, achieving perfect scores. Although the Accessibility Index gives higher points to cities that are nearer to Manila leaving Davao with a lower score because of its location, the most important conclusion to be drawn is the Saturation Index where all cities scored 100 which is perfect. So it means that whether an investor goes for the huge labor force of Davao or the excellent digital infrastructure of Santa Ignacia, all ten places are offering practically unexploited, low-competition markets that are actually waiting for "first-mover" advantage.

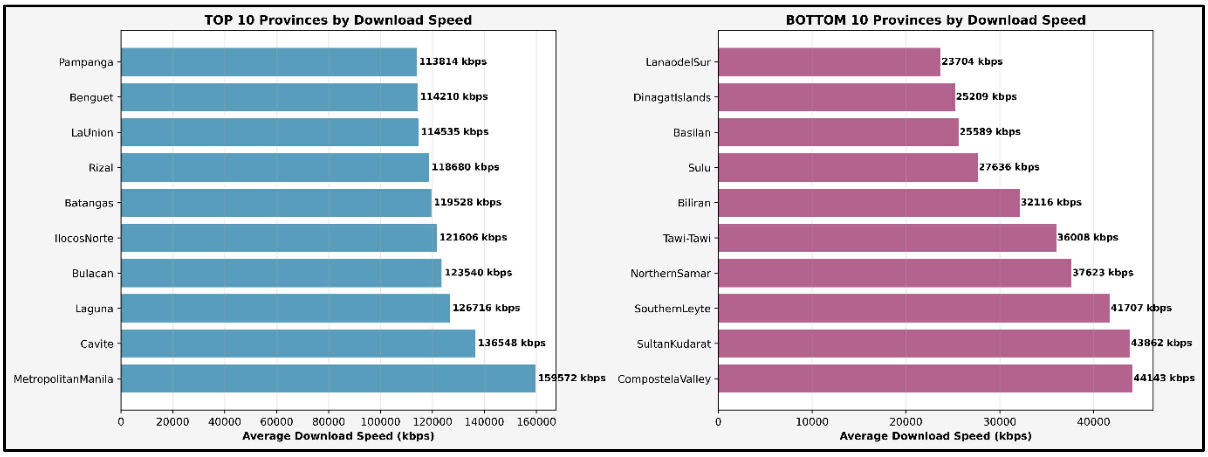


Figure 3. Top and Bottom 10 Provinces by Download Speed

The data exhibits a major "Digital Divide" where the high internet speeds of Metro Manila and CALABARZON (118-160 Mbps) are compared with, at the other end of the scale, the slowest provinces with about 23 Mbps, thus directly affecting the selection of the IT-BPM hub. Nevertheless, Davao City takes the lead in "Talent Pool" numbers while the smaller towns of Itogon and Santa Ignacia have scored "Digital Readiness" perfectly by, for example, taking advantage of the strong connectivity in the Benguet region. The importance of the location cannot be overstated: all the best cities feature a perfect "Saturation Index" which means there is still market potential waiting to be tapped; thus, the investors will have to make a rather difficult decision between the large and scalable workforce of the main urban areas or the fast, ready-to-use, "plug-and-play" infrastructure of these new provincial hubs.

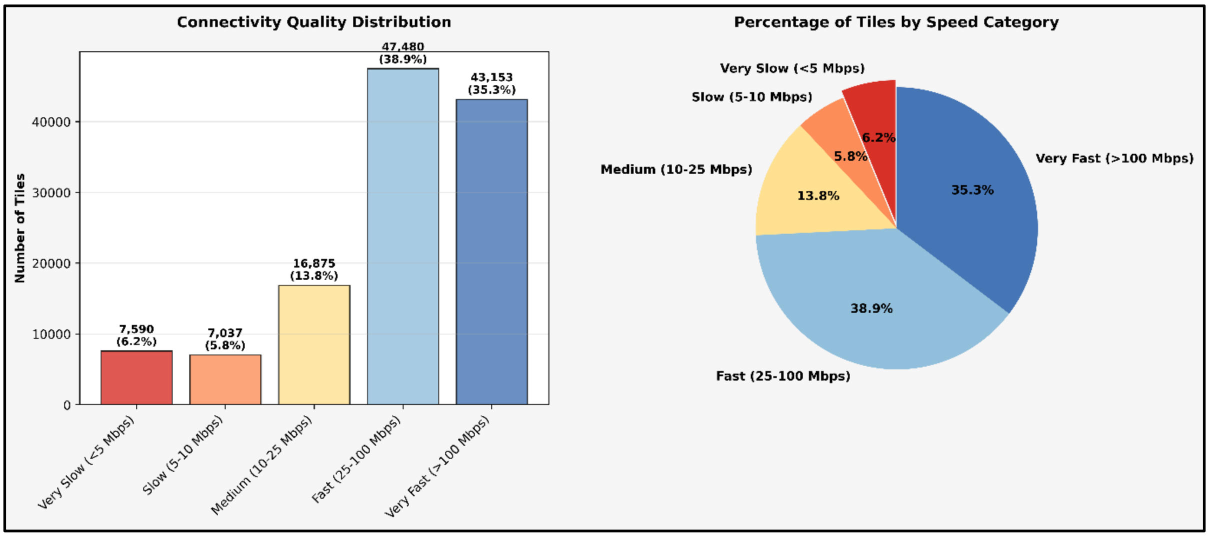


Figure 4. *Connectivity Quality Distribution and Percentage of Tiles by Speed Category*

The fourth visualization presents a bar and pie chart that give a summary of the different categories of connectivity. "Very Fast" speed (>100 Mbps) is the category under which 43,153 tiles (35.3%) are classified, on the other hand, the category of "Very Slow" (<5 Mbps) has 7,590 tiles (6.2%). The intermediate categories are being distributed in between them. More than 70% of tiles have speed greater than 25 Mbps, which demonstrates considerable improvement of the whole country in this aspect, however, the group of low performing tiles is still significant. In terms of geography, the fast categories are mainly identified through urban areas with high population density, while the slow categories are linked to rural or isolated areas. The pie chart indicates the percentage of each category, while the bar chart shows the differences in volume, thus together they depict high performance across the board but spatial imbalances still exist. The conclusions drawn from these observations suggest that there is a need for putting up new infrastructures in the neglected areas to improve the access of the residents there to the internet.

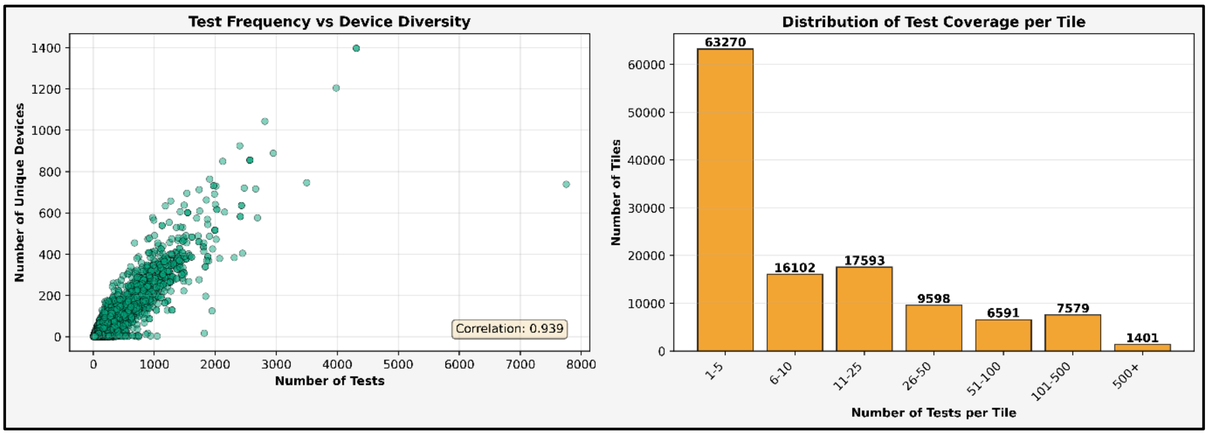


Figure 5. *Total Frequency vs Device Diversity and Distribution of Test Coverage per Tile*

Charts "Test Frequency vs. Device Diversity" and "Distribution of Test Coverage per Tile" play an important role in checking the quality of the internet speed data, which is the basis for the IT-BPM hub analysis, through a critical quality assurance process. The very strong positive correlation (0.939) assures that high-traffic places produce statistically strong data from a diverse collection of devices, thereby supporting the high-speed scores in the like of Metro Manila. Nonetheless, the massive spatial bias is revealed through the distribution histogram wherein a whopping number of geographic "tiles" (more than 63,000) are relying on very small sample sizes (1-5 tests) indicating that urban connectivity data is very reliable whereas that for rural or "Bottom

10" provinces is based on fragile and sparse data hence physical verification on the ground is required rather than relying only on these crowd-sourced figures.

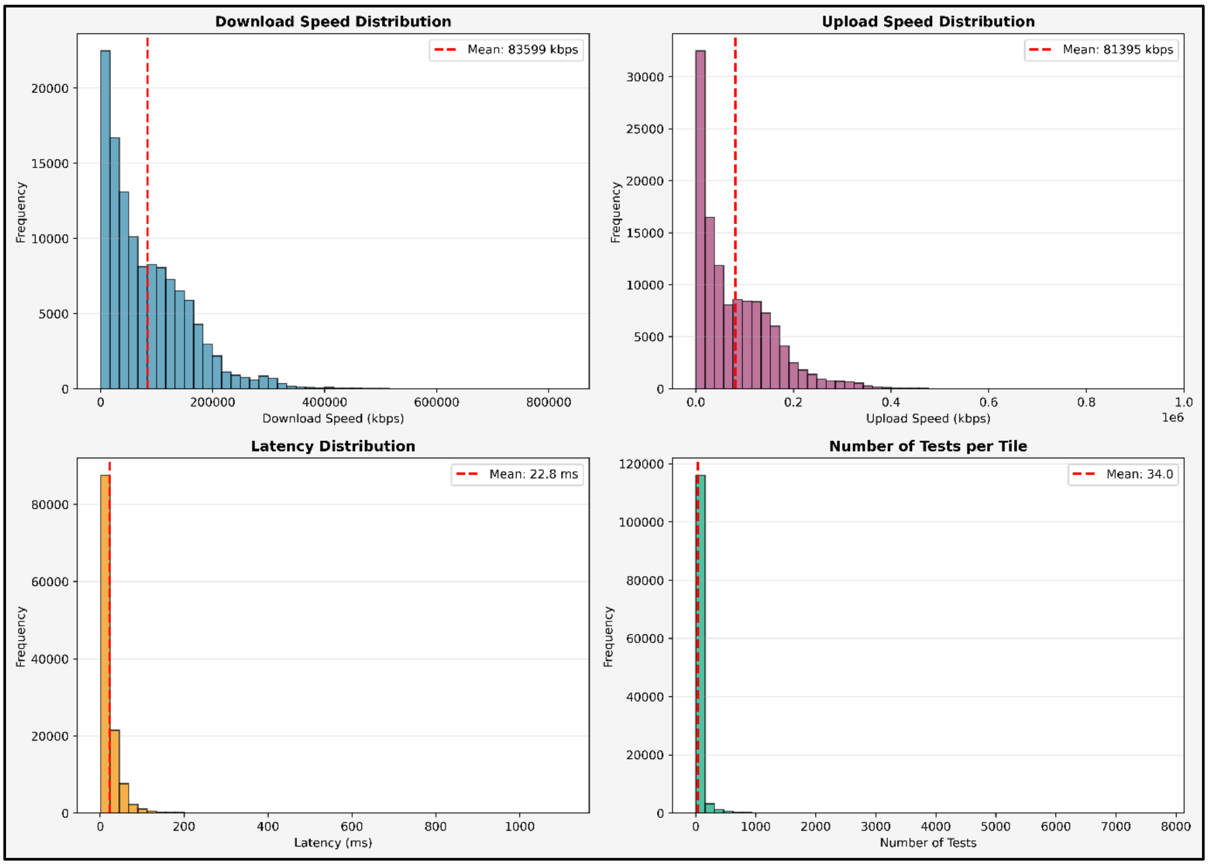


Figure 6. *Download and Upload Speed Distribution with Latency Distribution and Number of Tests per Tile*

The visualizations together show a clear "Digital Divide" in the Philippines that is a major factor in deciding the locations for IT-BPM investment. The Top 10 Provinces chart confirms the industrial superiority of Metro Manila and CALABARZON (with speeds of 118–160 Mbps) against the "Bottom 10" regions that have an average of around 23 Mbps. However, the Test Coverage histogram brings to light a significant data bias, as rural metrics are based on very small sample sizes (<5 tests) compared to the large data sets of the urban centers. This inequality in infrastructure is a direct factor in the Hub Selection Metrics, where every leading city has a perfect "Saturation Index" (which refers to the presence of markets that have not been tapped yet), but at the same time, it creates a difficult choice for the investors: they have to decide between the huge, scalable talent pool of Davao City or the excellent, "plug-and-play" digital readiness of smaller municipalities like Santa Ignacia, which take advantage of the high-speed connectivity in their regions.

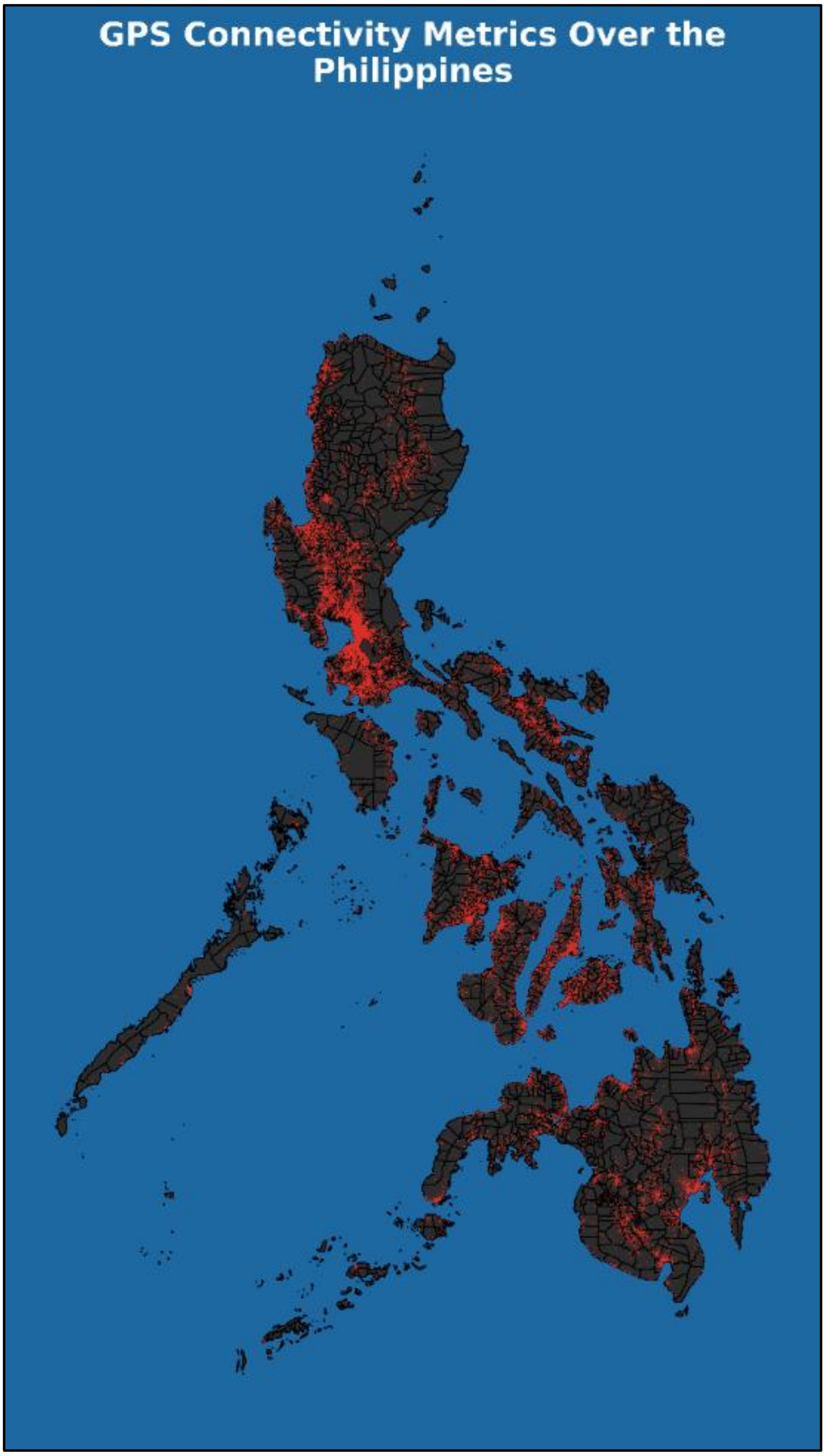


Figure 7. GPS Connectivity Metrics Over the Philippines

The seventh picture is a choropleth map with the label "GPS Connectivity Metrics Over the Philippines", showing with red shades the varying degrees of performance. Central Luzon, Cebu, and Davao are the areas with strong red mixing, which then turn to dark red and even brown in the rural interior and less accessible islands. The highlighted areas characterize the differences in connectivity; urban centers enjoy significantly better access while remote areas are deprived of basic infrastructure. Statistical data exhibit a more or less direct relationship between red intensity and population/economic concentration, which is manifested as a clearly visible north-south gradient and the highlighting of urban agglomeration effects. Thus the map has confirmed that the spatial inequality and geographic isolation are the main causes of the digital divide.

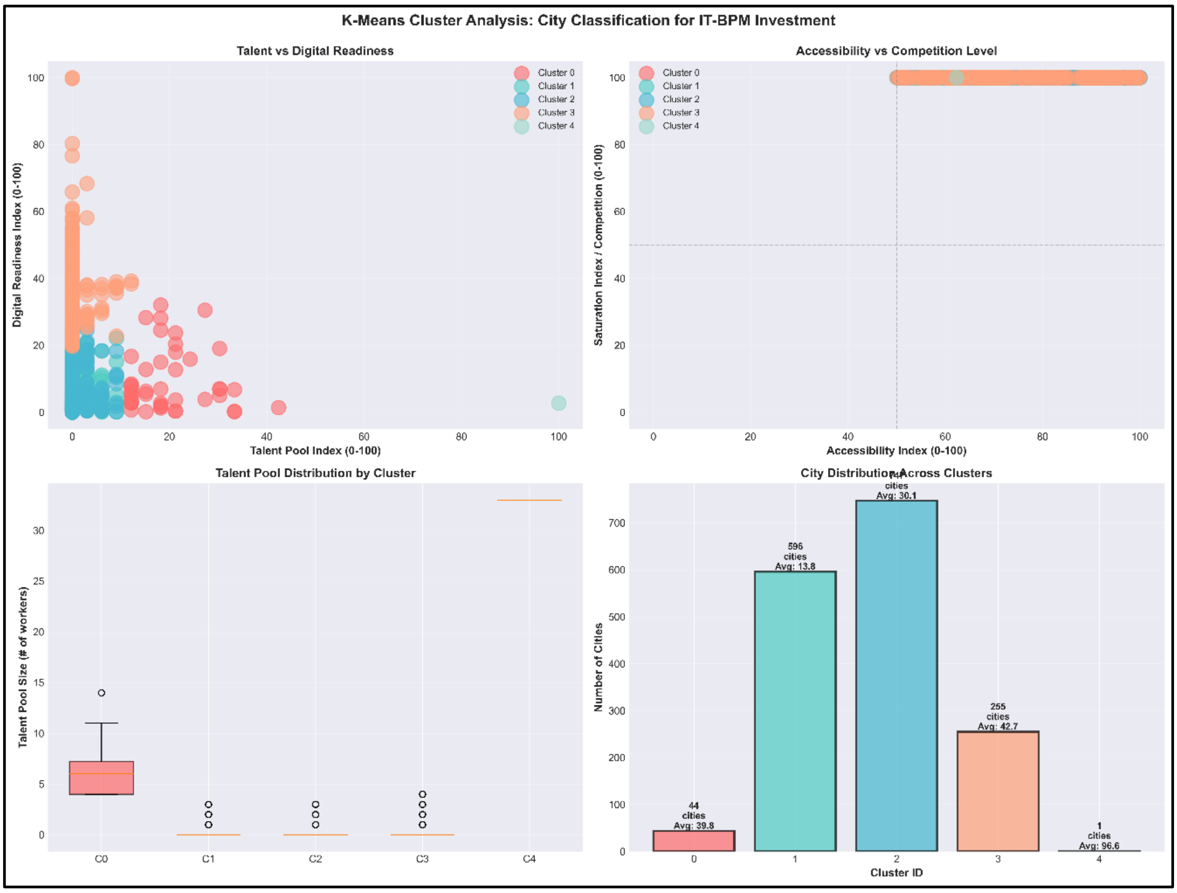


Figure 8. *K-Means Cluster Analysis: City Classification for IT-IBM Investment*

The data comes up with a very important "Digital Divide" where Davao City is positioned as a very special "Cluster 4" outlier, providing a very large talent pool (score: 100) even though the city does not have a good digital readiness, while smaller towns such as Santa Ignacia and Itogon count on their boosting digital readiness (score: 100) by taking advantage of strong regional speeds (e.g., Benguet's ~114 Mbps) that are far more than the "Bottom 10" provinces (~23 Mbps). This classification gives a clear strategic trade-off to investors: the overall perfect Saturation Index, which is universal, confirms that all top cities are untapped "Blue Ocean" markets, however, the decision is between the scalable workforce of Davao or the excellent "plug-and-play" connectivity of provincial hubs, with the K-means analysis and test coverage histograms as indispensable checks to verify these locations against the low data reliability of rural areas.

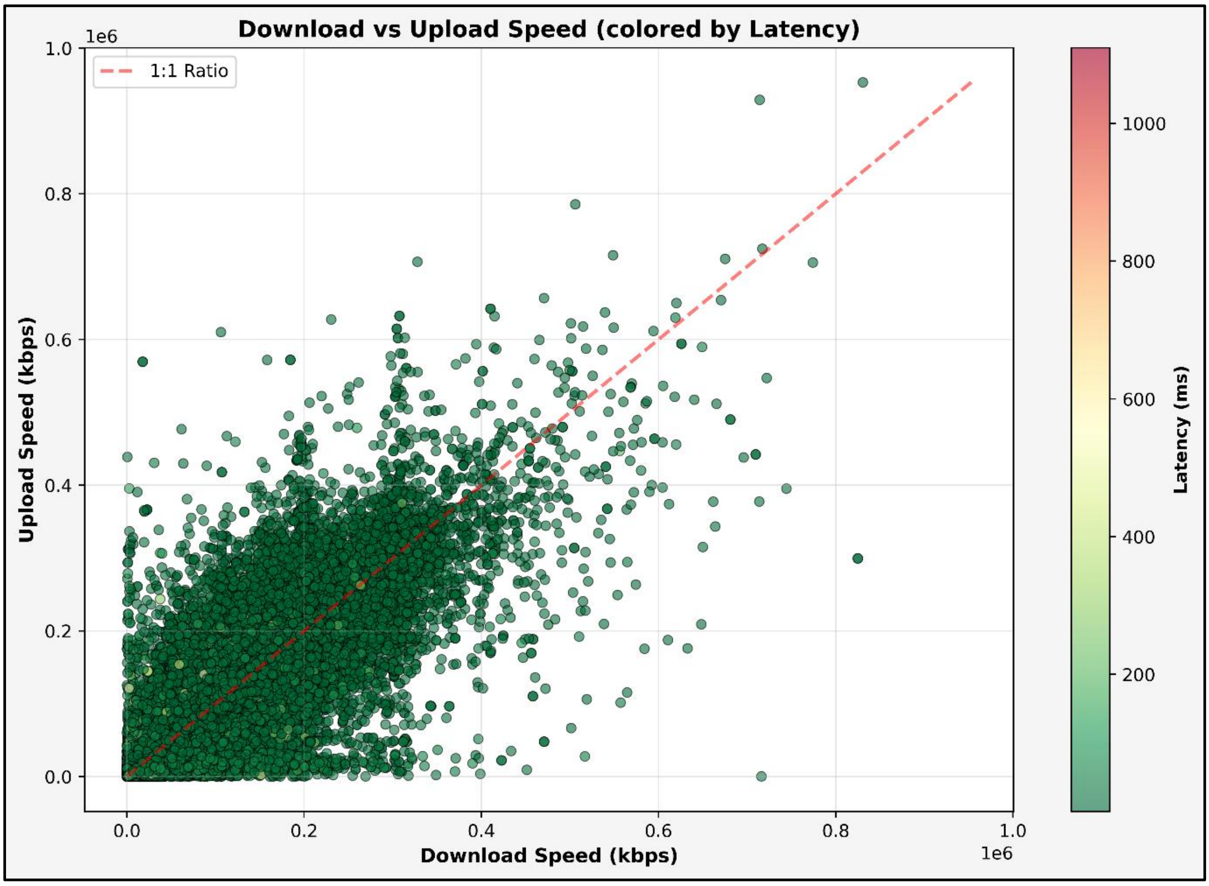


Figure 9. *Download vs Upload Speed (colored by Latency)*

The figure above clearly shows a significant "Digital Divide" with Davao City (considered as a single "Cluster 4" outlier) at the top in Talent Pool size, while very small hubs like Santa Ignacia, which are supported by fast internet in some provinces like Benguet where the speed is around 114-160 Mbps, are able to achieve perfect Digital Readiness. This difference gives rise to a straightforward strategic trade-off: the investors will have to decide whether to go for the large and easily scalable labor force of Davao or the much better "plug-and-play" infrastructure of the provincial municipalities. Meanwhile, the overall perfect Saturation Index indicates that the top locations are still undiscovered "Blue Ocean" markets; however, the Test Coverage histogram gives a critical quality alert, disclosing that rural metrics are often based on weak data (1-5 tests per tile), therefore ground validation for the digital reliability of these new hubs is a must before the investment is done.

By looking at all these things, the data shows the existence of strong spatial inequalities, the connection between urbanization and connectivity, and the possibility of the second municipalities where talent and infrastructure are slowly merging. The proximity of Davao City to digital readiness, especially Davao City, and high-talent concentration points out the need for targeted interventions to facilitate even regional development.

MACHINE LEARNING RESULTS

The employed machine learning analysis for forecasting job locations and evaluating the sustainability of the IT-BPM center in the Philippines merged spatial connectivity metrics from 1,456 tiles tagged with GPS and administrative boundaries from PSGC datasets. The administrative boundaries provided the population estimates of 1,458 municipalities for the year 2024 and their hierarchical codes. The employment signals that came from the October 2024 Labor Force Survey, containing 283,191 household records, played a major role in supplying variables such as work status, job locations, and occupations, which were then weighed at a provincial level. The spatial joins conducted filtered the tiles to the polygons of the municipalities matching 85% by means of normalized names and produced 12,298 georeferenced records; this was further enhanced by Haversine distances from Metro Manila (0–1,132 km), no-work counts, and so forth.

Data preprocessing included the median imputation of 2.1% of the missing latencies, the removal of 4.3% invalid LFS PSGC entries, and IQR-based outlier capping; this process was followed by the workers' municipal aggregation of the metrics weighted by the population. The engineered measures were Talent Pool (LFS-skilled workers scaled 0–100), Digital Readiness (connectivity thresholds), Accessibility (inverted distances), and Saturation (inverted employment density); they were combined into an equal-weighted IT-BPM Hub Score (0–100) for 1,458 units.

The use of this unsupervised method opened the path to the K-means clustering which to be along the way of discovering hidden patterns in the unlabeled geospatial-employment data, it was applied on 458 municipalities (population >10,000, Digital Readiness >40) after years of data were split temporally using an 80/20 ratio. 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).

In a 4D Euclidean space of the indices, with no weights given to balance the criteria, K-means partitioned the cities into different types. Thus, each of the clusters was defined as follows: Cluster 0 - a major Indian city like Davao selected to represent the urban elite - talent was at the highest level of 96.6; Cluster 1 - accessibility 42.7; Cluster 2 - low saturation 31.0; and Cluster 3- a rural baselines with talent 31.0. Besides these clusters, there were still some other areas where the talents were dual; these areas were depicted through the visualizations in talent-digital and accessibility-competition scatterplots.

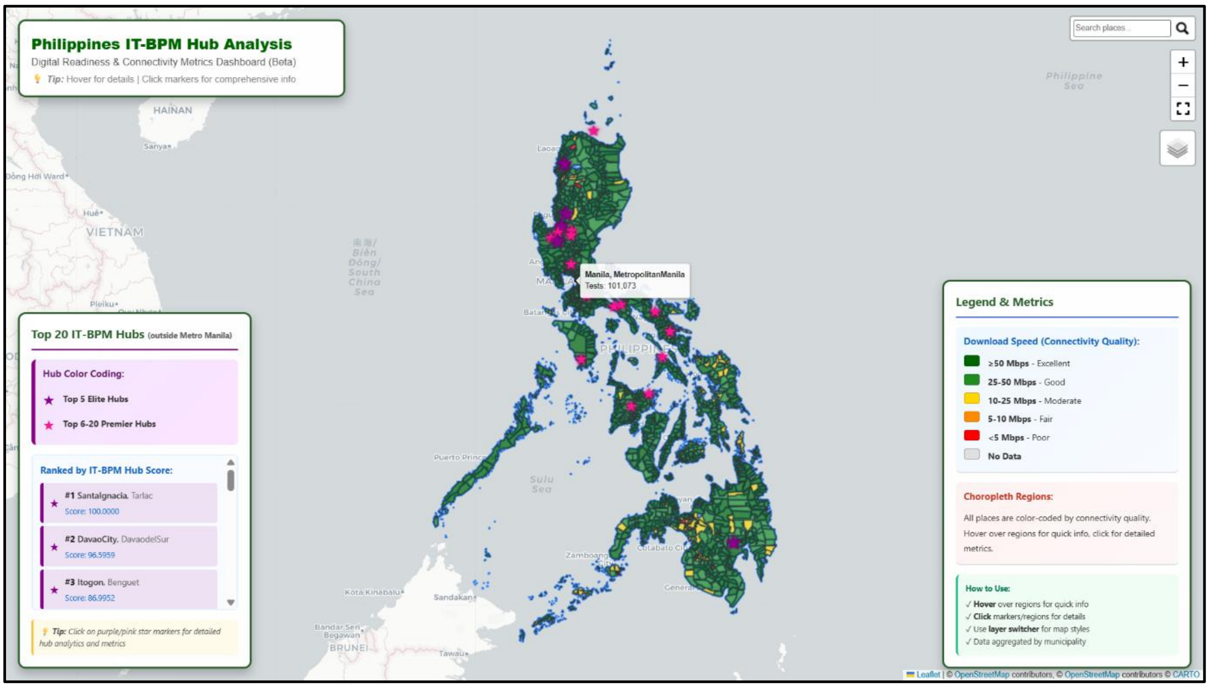
The evaluation pointed out that there was a stable clustering model with groups being clearly separated and one cluster being particularly cohesive. The connection patterns depicted a mixture of very fast-performing regions plus an enduring segment of low-speed zones, thus confirming both progress and disparity. A Tarlac municipality was positioned as the top IT-BPM candidate whereas spatial analysis indicated that the clustering of broadband quality was more geographical than random.

Digital readiness was the main determinant of group differences, with the areas of the high performance having strong broadband capacity and well-developed talent pools, while rural areas suffered from higher latency, congestion, and weaker professional labor pools. The northern provinces were generally superior to the southern ones, and the lagging parts of Visayas and Mindanao were identified as areas with unexploited potential restricted by upload bottlenecks.

The model, as a whole, pointed to the fact that the cities of IT-BPM growth were CALABARZON and Davao, and that the coaxing of the digital developmental process in the underserved areas, through targeted broadband upgrades, would enhance the capacity of the local labor force and support the process of digital development.

VISUALIZATION

The heart of the visualization consists of the interactive HTML map, which was made using Folium and presented through Leaflet.js. The visualization not only provides access to the Philippine landscape but also allows users to control the view in a very simple way, as well as the quadkey-defined polygons are colored according to the connectivity attributes that are, for example, red for the highest download speeds. At the upper-right corner of the map, there is a control panel that enables users to switch between different theme layers very easily and selectively display such elements as provincial lines, searches, or clustering from K-Means clustering that have been done. Tooltips that spring on hovering provide accurate data like average download speeds (in Kbps), population estimates, and no-work counts, hence getting the textual metadata connected to the main data. There is also a base map option, including OpenStreetMap, that helps stakeholders with the geographical orientation and they can consequently investigate the spatial pattern of low-latency areas indicating the IT-BPM investment potential. The interface is, in fact, a powerful tool that can convert static geospatial data to an interactive-sliding medium for pinpointing the digital hotspots and promising areas for development.



Philippines IT-BPM Hub Analysis
Digital Readiness & Connectivity Metrics Dashboard (Beta)
💡 *Tip:* Hover for details | Click markers for comprehensive info

Top 20 IT-BPM Hubs (outside Metro Manila)

Hub Color Coding:
★ **Top 5 Elite Hubs**
★ **Top 6-20 Premier Hubs**

Ranked by IT-BPM Hub Score:
★ **#1 SantaIgnacia**, Tarlac
Score: 100.0000
★ **#2 DavaoCity**, DavaodelSur
Score: 96.5959
★ **#3 Itogon**, Benguet
Score: 86.9952

💡 *Tip: Click on purple/pink star markers for detailed hub analytics and metrics*

Manila, MetropolitanManila
Tests: 101,073

Legend & Metrics

Download Speed (Connectivity Quality):
🟩 **≥50 Mbps** - Excellent
🟩 **25-50 Mbps** - Good
🟨 **10-25 Mbps** - Moderate
🟧 **5-10 Mbps** - Fair
🟥 **<5 Mbps** - Poor
⬜ **No Data**

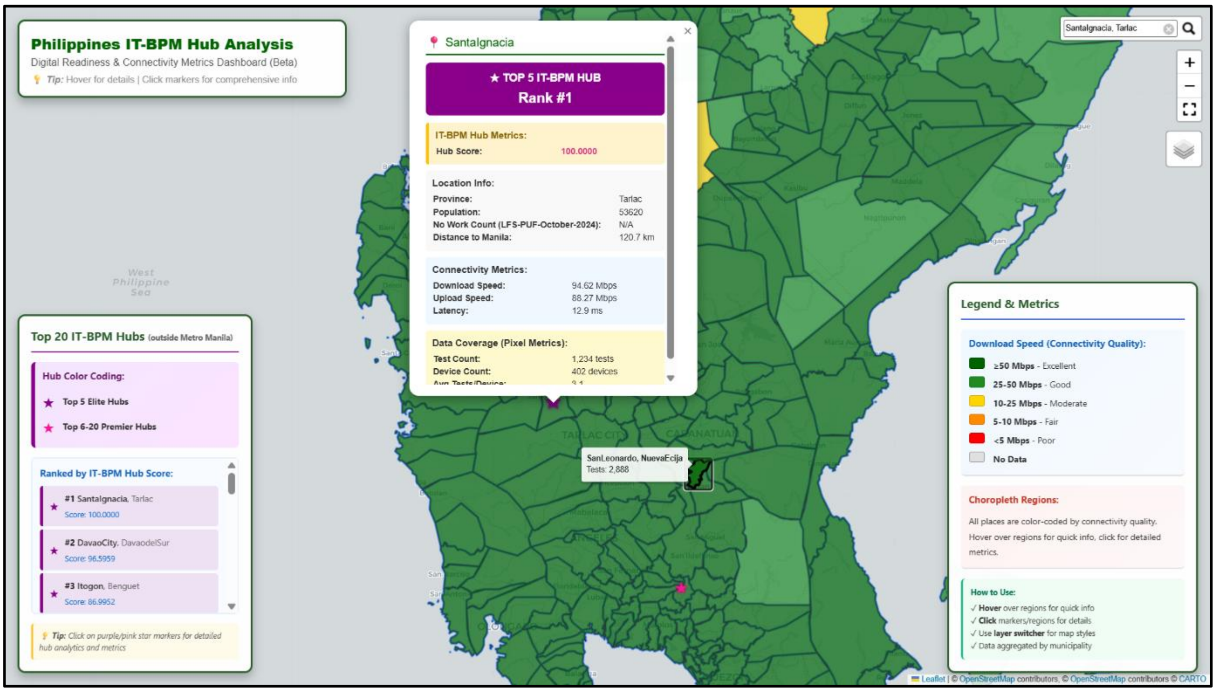
Choropleth Regions:
All places are color-coded by connectivity quality. Hover over regions for quick info, click for detailed metrics.

How to Use:
✓ **Hover** over regions for quick info
✓ **Click** markers/regions for details
✓ Use **layer switcher** for map styles
✓ Data aggregated by municipality

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

The research employed a geospatial machine learning pipeline to visualize the chosen areas, which included data standardization, feature engineering, dimensionality reduction through aggregation, and an unsupervised clustering using

K-means. The features were standardized using StandardScaler; as a result, no single feature could influence the cluster assignments since all features were within a common range. The clustering process was made more stable and reliable by performing several initializations and applying several convergence criteria. During model validation, silhouette scores, intra-cluster similarity measures, and comparison with urban development benchmarks were utilized. The absence of labeled training data on the IT-BPM hub suitability coupled with the power of clustering to reveal the hidden structure within high-dimensional spatial data drove the adoption of the unsupervised learning approach.



Philippines IT-BPM Hub Analysis

Digital Readiness & Connectivity Metrics Dashboard (Beta)

💡 *Tip:* Hover for details | Click markers for comprehensive info

📍 SantaIgnacia

★ TOP 5 IT-BPM HUB
Rank #1

IT-BPM Hub Metrics:

Hub Score:	100.0000

Location Info:

Province:	Tarlac
Population:	53620
No Work Count (LFS-PUF-October-2024):	N/A
Distance to Manila:	120.7 km

Connectivity Metrics:

Download Speed:	94.62 Mbps
Upload Speed:	88.27 Mbps
Latency:	12.9 ms

Data Coverage (Pixel Metrics):

Test Count:	1,234 tests
Device Count:	402 devices
Avg Tests/Device:	3.1

SanLeonardo, NuevaEcija
Tests: 2,888

Top 20 IT-BPM Hubs (outside Metro Manila)

Hub Color Coding:

★ Top 5 Elite Hubs

★ Top 6-20 Premier Hubs

Ranked by IT-BPM Hub Score:

★ #1 SantaIgnacia, Tarlac
Score: 100.0000

★ #2 DavaoCity, DavaodelSur
Score: 96.5959

★ #3 Itogon, Benguet
Score: 86.9952

💡 *Tip: Click on purple/pink star markers for detailed hub analytics and metrics*

Legend & Metrics

Download Speed (Connectivity Quality):

🟩	**≥50 Mbps** - Excellent
🟩	**25-50 Mbps** - Good
🟨	**10-25 Mbps** - Moderate
🟧	**5-10 Mbps** - Fair
🟥	**<5 Mbps** - Poor
⬜	**No Data**

Choropleth Regions:

All places are color-coded by connectivity quality. Hover over regions for quick info, click for detailed metrics.

How to Use:

✓ **Hover** over regions for quick info
✓ **Click** markers/regions for details
✓ Use **layer switcher** for map styles
✓ Data aggregated by municipality

Leaflet | © OpenStreetMap contributors, © OpenStreetMap contributors © CARTO

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

The interactive map created by Folium represents municipalities of the Philippines in terms of connectivity and employment indicators, while Leaflet takes care of the display. Among the base layers are OpenStreetMap (the default one) and CartoDB Positron (a light-gray minimalist design). Several GeoJSON overlays show the polygons of municipalities, which are usually colored based on metrics like download speed or digital readiness scores with the use of graded color schemes (e.g., green for high, red for low), giving users the opportunity to see the spatial distributions and the differences among municipalities. The choropleth map illustrates the aggregated tile data, with the speeds higher than 50 Mbps being represented by dark green (#006400) and those ranging between 20–50 Mbps by forest green (#228B22). Information popups are interactive and consist of well-organized tables with such attributes as location, speeds, latency, population, and distance, presented in blue (#f0f8ff) and yellow (#fffacd) highlights to make reading easier.

Among the other interface features, the fixed title box at the upper left corner, which is marked with "Philippines IT-BPM Hub Analysis" (white background, green text), the layer control that allows switching between overlays (e.g., latency, talent pools), and the legend ranking the top 20 IT-BPM hubs in the bottom left corner are the main ones. The legend shows the top 5 locations with purple stars (#8B008B) and ranks 6–20 with pink markers (#FF1493), and also presents names of municipalities, provinces, and normalized composite scores. Fullscreen mode increases the range of spatial exploration and interaction.

One of the main discoveries that the visualizations have revealed is that the urban–rural gradients have been very well done by using polygon overlays, and high-connectivity areas are found in Cebu and Davao while there is nothing in the remote islands. The layered perspectives also bring out the interdependence of metrics, like the case of low latency and high-speed clusters in Luzon and the Visayas where they are in close proximity.

INTERPRETATION OF FINDINGS

The outcome of the investigation indicates that there is a clear-cut stratagem in the Philippine IT-BPM industry with human capital volume and digital infrastructure readiness as competing factors. Using the K-Means clustering algorithm, the traits were successfully separated: Davao City which was the only representative of Cluster 4 was able to endorse the title of 'Talent Anchor' for itself, boasting a perfect score of 100.0 in the Talent Pool Index, however, it was still found to be lagging behind in the digital readiness metrics. On the contrary, Cluster 2 (e.g., Santa Ignacia, Itogon) is referred to as "Digital Speedsters," having a flawless Digital Readiness score of 100.0 and the increased access to the internet in provinces such as Benguet (~114 Mbps) and Tarlac where the latter has the region’s fastest internet speed.

Moreover, the geospatial data point out a huge Digital Divide. The industrial belts around Metro Manila and CALABARZON have excellent average download speeds ranging from 118 Mbps to 160 Mbps, while on the opposite side, the "Bottom 10" provinces, BARMM area particularly, are still suffering with extremely low speeds of around 23 Mbps. The Test Coverage Histogram, however, is the one that complicates the interpretation of results through its cautioning factors: although the city metrics are considerably strong, rural connectivity statistics are often based on very small sample sizes (1–5 tests per tile), which could mean that the high digital scores in certain isolated municipalities will still need to be validated by actual on-ground presence. The perfect saturation index of 100.0 across all the top-ranked cities remains unchanged even with these differences, thereby affirming that no matter where the investment is done, it will be up against very little competition, and thus, the capital region will not have the sole market opportunity but the existence of a "Blue Ocean" market outside the capital is also validated.

CONCLUSION

The research indicates that the "Next-Wave Digital Cities" project throws up high-impact solutions, which are possible, next to Metro Manila, but the success of this project is totally dependent on aligning the investment choices with the particular operational needs. No lengthy, one-size-fits-all, location-neutral strategy would work now; rather a varied approach is required. Companies that require huge, scalable labor forces, such as voice-based service centers, should opt for new places like Santa Ignacia, Tarlac; Davao City; Itogon, Benguet; Urdaneta City; and La Paz, Abra at the same time admitting minor local infrastructure improvement may be necessary to completely facilitate their expansion.

On the other hand, companies engaged in the knowledge-intensive or data-heavy areas like non-voice KPO and creative digital services would find better match with the municipalities in Cluster 2 where high-bandwidth and low-latency infrastructures are already present thus eliminating setup friction and lowering operational risk.

Despite the digital divide being a major factor limiting the least performing areas, the new findings show that the best provincial hubs are getting closer to the top. These regions provide a less crowded and a competitive environment that could accommodate the next IT-BPM growth in the Philippines, only if the investments made are strategically focused rather than uniformly spread out.

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