CGPM: Poverty Mapping Framework based on Multi-Modal Geographic Knowledge Integration and Macroscopic Social Network Mining

Zhao Geng[1], Gao Ziqing[2], Tsai Chihsu[1], and Lu Jiamin[1](✉)

[1] Department of Mathematics, Department of International Economics and Trade, College of Artifical Intelligence, Jinan University, Guangzhou, China
zg1063316621@outlook.com, tttjlu@jnu.edu.cn

[2] Department of Chinese Language and Literature, Xi'an Jiaotong University, Xi'an, China

Abstract. Having high-precision and high-resolution poverty map is a prerequisite for monitoring the United Nations Sustainable Development Goals(SDGs) and for designing development strategies with effective poverty reduction policies. Recent deep-learning-related studies have demonstrated the effectiveness of the geographically-fine-grained data composed with satellite images, geolocated article texts and Open-Street-Map in poverty mapping. Unfortunately, there is no presented method which considers the multimodality of data composition or the underlying macroscopic social network among the investigated clusters in socio-geographic space. To alleviate these problems, we propose CGPM, a novelty end-to-end socioeconomic indicator mapping framework featured with the cross-modality knowledge integration of multi-modal features, and the generation of macroscopic social network. Furthermore, considering the deficiency of labeled clusters for model training, we proposed a weak-supervised specialized framework CGPM-WS to overcome this challenge. Extensive experiments on the public multimodality socio-geographic data demonstrate that CGPM and CGPM-WS significantly outperforms the baselines in semi-supervised and weak-supervised tasks respectively of poverty mapping.

Keywords: Sustainability · Poverty Mapping· Multi-Modality· Social Networks Mining.

1 Introduction

Recently, the application of data mining in the field of sustainable development and global human rights protection has attracted a lot of attention [12, 22, 18]. One of the important applications is intelligent poverty mapping [19, 1]. The main content of poverty mapping is to obtain high-precision key socioeconomic indicators [4] that measure the wealth level or poverty level of geographically distributed clusters. Having high-precision poverty maps is a crucial prerequisite for monitoring the UN Sustainable Development Goals(SDGs) and designing development strategies [10, 22] for effective poverty reduction policies.

Driven by this, with the continuous progress of cutting-edge research in data mining and the increasing availability of geospatial data, many recent studies have proposed frameworks that combine deep learning algorithms with geospatial information as a highly-accurate, low-cost, and scalable technical system [2] to conduct poverty mapping.

Existing work on mapping poverty generally collects and uses multimodal geospatial data. Among the many data sources, the image data from Google's satellite map [9, 26], Geolocated Article Texts data [15, 3] and Open Street Map data [10, 14] are currently used by the mainstream. Open-sourced Demographic and Health Survey(DHS) data is often used as professional socioeconomic indicators.

However, we must emphasize that most of the existing A.I. poverty mapping algorithms simply aggregate the training results of various specialized models. Such algorithms cannot overcome the two challenges faced by the current poverty mapping: 1) The ideal poverty mapping algorithm should realize the integration and complementary enhancement of multimodal data, to obtain the optimal cluster-level features representations. However, the proposed integrated or end-to-end splicing frameworks do not have this function. 2) The various clusters on the poverty map do not exist in isolation, and they constitute a potential social network with multiple semantics. We hope to be able to intelligently mine and generate the potential macro-social network structure at the cluster level, and synergistically apply it to the optimization of node (cluster) level feature representation, which will help to improve the accuracy of the poverty mapping framework.

Therefore, to alleviate the above problems, we propose CGPM, a novelty Poverty Mapping Framework simultaneously considers the Integration of Multi-Modal Geographic data and the mining/generating of the underlying Macroscopic Social Network. The framework mainly includes two core components, Cross-modal Feature Integration Module and Feature-based Macroscopic Social Network Generating Module. In the Cross-modal Feature Integration Module, we construct a cross-modality feature transformer based on the cross-modality attention mechanism, and use it to conduct cross-modal feature transformation and integration. In the Macroscopic Social Network Generating module, we generate multiple feature-based candidate macroscopic social network structure graphs. Furthermore, we jointly train the integrated representations, the generated social network structure (parameterized), and the networks' semantic embedding (parameterized)under the task, and eventually conduct the high-precision poverty mapping. Meanwhile, to alleviate limitedness of the DHS labels that we often face in actual surveying and mapping, we improved the CGPM and proposed a specialized architecture for weakly supervised scenarios (CGPM-WS). CGPM-WS performs refinement operations based on the pseudo-labeling technique to obtain a better feature representation, and effectively overcome the above challenges. In conclusion, our contributions are as follows:

(a) CGPM is the first method which realizes cross-modal fusion and the underlying macroscopic social networks generating simultaneously, thus to augment the representations of the multimodal data in this field. Whether it is in the field of poverty mapping or the related frontiers of data mining, CGPM is significantly ahead of the baselines in terms of novelty and design, with relatively strong academic significance.

(b) We designed a specialized architecture for weakly supervised scenarios-CGPM-WS, as a migration variant of CGPM, to alleviate the problem of the low coverage of the cluster areas marked by DHS.

(c) Extensive experiments in six typically developing countries demonstrate that CGPM has a significant accuracy advantage over current baselines. CGPM-WS has a significant accuracy advantage while maintaining granularity. It is validated that CGPM has considerable application prospects in intelligent poverty mapping research.

2 Related Work

2.1 Existing Poverty Mapping Framework

In recent years, some progress has been made in the research on integrated/migratory poverty mapping based on artificial intelligence & data mining algorithms. Dohyung Kim [19] et al. proposed a migration algorithm based on satellite image data to achieve high-precision poverty prediction. Evan Sheehan [15] et al. established an integrated poverty-mapping-oriented deep neural network architecture based on Wiki geographic text comment data and high-definition satellite data. Chiara Ledesma [9] et al. introduced textual data and statistics from social media to optimize the accuracy of poverty prediction from the perspective of interpretable learning. Masoomali Fatehkia et al. [3] transferred the interpretability scheme to a wider range of socioeconomic indicator predictions and assessed the generalization of existing methods compared to baselines. Kumar Ayush [1] designed a dynamic mapping framework for poverty maps based on Reinforcement Learning, which achieved high-precision poverty detection with extremely high computational overhead. Lee.K [10] et al. proposed a technical approach to anchor clusters to sample alignment under multimodal data, while designing a simple architecture suitable for weakly supervised environments. However, existing research has significant shortcomings in multimodal data fusion and representation optimization, as well as the isolation assumption of individual clusters. Therefore, to alleviate the above two problems has become the motivation of our research.

2.2 Discussion: Underlying Macroscopic Social Network Mining & Generating—Why and How

In our research, we note that the existing A.I. poverty mapping frameworks treat each cluster as an isolated sample point. However, we know that in a spatial geographic area, due to the existence of social and economic ties, the clusters in the area are not completely independent from each other, and their subsets will form multiple macro-social networks based on these linkages and ties. We interpret the implications of the underlying macro-social network structure in poverty mapping as mobility and homogeneity. The mobility structure refers to the existence of important linkages or ties between two clusters in terms of economic and social activities. For example, the satellite city structures and core industrial chains in the metropolitan areas. Due to the existence of the social network formed economic circles, economic complexes and other entities among the clusters [24], the social functions of each cluster often show different characteristics. There exists uneven developments of the transformation, such as satellite cities tend to undertake more housing and basic medical care, but lack of other positive socio-geographical characteristics. This economic phenomenon is not clearly revealed in either the OSM features, the satellite images, or the lighting data. It is difficult for the existing frameworks to compare the low-poverty clusters with high-poverty clusters while recognizing their with uneven values in such indicators with precise distinctions. The realization of feature sharing and dissemination among nodes in such a social network structure can effectively alleviate this problem. Homogeneity structure means that, for a aggregated feature representation measured by weight, if the values of two clusters show a sufficiently high similarity, we can consider the two as homogeneous clusters. Therefore, we can smooth the node feature vector of each cluster in such a homogeneous social network to a certain extent [8], so as to alleviate the observation error. caused by the operation in data collection and cluster anchoring.

Meanwhile, in sociological investigations, researchers often determine the semantics of social network structures by means of subjective definitions[27]. However, in the A.I. application scenarios, this would not be an optimal graph-structured representation. Therefore, we hope that the framework adaptively learns a better latent social network structure representation from the node (cluster) features, and cooperates them with the spatial node-level message passing layer (with graph structure and node features as trainable inputs) training to optimize the feature representations of each cluster. Such practices can be realized with the Graph Structure Learning [28, 21, 13, 17] and Spatial Encoding [25] techniques.

3 Data Acquisition and Preprocessing

Open Street Map Data & Preprocessing

OpenStreetMap (OSM) contains open-source geospatial and infrastructure data open to the world[18]. A recent study shows that user-generated road maps in OSM are about 86% complete by 2020, and more than 40% of countries have a complete OSM street network[7]. We can obtain OpenStreetMap (OSM) data for the target area from Geofabrik, an online repository for OSM data[6]. From this, we extract extensive information about the number of roads, buildings and points of interest in a specific area, which will be presented as tabular features. We further discretize it to obtain Categorical Features.

In our feature engineering, OSM feature extraction ranges from rural areas with a 5 km radius and urban areas with a 2 km radius, each centered on the cluster location. We identified five road types in the dataset: arterial, arterial, paved, unpaved, and intersection. In terms of engineering road features, the preprocessing techniques we employ are as follows: for each type of road, we calculate the distance from the current cluster to the nearest road, the total number of roads, and the total road length for each cluster.

Satellite Imagery Data & Preprocessing

High-definition satellite image data is the most used geospatial data with the most stable acquisition channels in Poverty Mapping researches. We use the Google Static Maps API to obtain satellite images of clusters in rural areas within a radius of 0.5-5 km, and satellite images of clusters in urban areas within a radius of 0.5-3 km. To support this study, we prepared a total of 77960 images for download with a zoom level of 17, a scale of 1, and a pixel resolution of about 1.25 meters. And after matching it with the area covered by a single data point of nighttime light data, typically each image can cover a land area of 0.25 km.

The Night Light (NTL) data we use is from the 2019 Visible Infrared Imaging Radiometer Suite Day/Night Band Dataset(VIIRS DNB). The VIIRS DNB dataset has a nighttime luminance resolution of 15 arcseconds, and contains geophotometric data on the ground at continuous photometric levels from 0 to 122. It can be used with Satellite or by calculating statistics (for example, nighttime light intensity within 1x1 square kilometers around the area) as tabular features. In this study, we choose the latter.

To more fairly demonstrate the efficiency of our proposed framework, align with most Poverty Mapping methods, we deploy the trained open-source VGG16 model accepting 400 × 400 pixel images. Further, we augment the data with random horizontal mirroring and use 30% dropout on the convolutional layers instead of the fully connected layers. Finally, we start to fine-tune the entire network using the Adam optimizer, and get a preliminary 3200-dimensional imagery embeddings.

Wiki Text Data & Preprocessing

In terms of text encoding, we completely follow the preprocessing scheme of MMPM [15]. We use the pre-trained Doc2vec model open sourced by Genism to encode text data from about 1.2K articles. In terms of parameters, we set the Windows Size as 8, and obtain a 400-dimensional text embeddings.

Labeling: Demographic and Health Survey Indicator

In this poverty mapping study, we use the Resident Wealth Index (DHS-WI) published by the International Population and Health Organization as a dependent variable indicator to measure the poverty level of the clusters. Through past researches [16], scholars from various countries have widely recognized that DHS-Program data can be used as the basic fact for constructing indicators to measure social economic activities.

DHS-WI is a comprehensive wealth measurement index constructed by DHS Program officials based on its surveys. In the existing work [23], the researchers proved that the DHS-WI indicator has a strong correlation with the international wealth index [10](IWI, a common set of asset weighting calculations which is widely accepted as a measure of wealth index or poverty level, and is difficult to be calculated).

Eventually, the representations of our pre-trained multi-modal features (tabular-categorical features, image embeddings, text embeddings) can be written as $F \in \mathbb{R}^{n \times d_f}$, $I \in \mathbb{R}^{n \times d_I}$, $T \in \mathbb{R}^{n \times d_t}$, which is the inputs of CGPM.

4 Methodology

Framework: Figure 1 demonstrate the architecture of CGPM, which is the core algorithm of our proposed A.I. poverty mapping module based on cross-modality integration and graph structure learning (social network generating). We firstly construct a cross-modality feature transformer based on the cross-modality attention mechanism in order to implement the cross-modal feature transformation and integration. In the social network generating module, we generate multiple feature-based candidate macroscopic social network structure graphs. Furthermore, we jointly train the integrated representations, the generated social network structure (parameterized), and the semantic embedding (relation embedding, parameterized) of each graph structure under the task, eventually achieve/conduct? the high-precision poverty mapping.

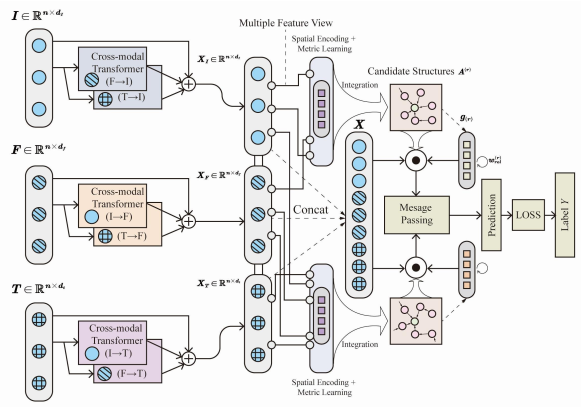
$I \in \mathbb{R}^{n \times d_I}$
Cross-modal Transformer (F→I)
(T→I)
$F \in \mathbb{R}^{n \times d_f}$
Cross-modal Transformer (I→F)
(T→F)
$T \in \mathbb{R}^{n \times d_t}$
Cross-modal Transformer (I→T)
(F→T)
$X_I \in \mathbb{R}^{n \times d_I}$
$X_F \in \mathbb{R}^{n \times d_f}$
$X_T \in \mathbb{R}^{n \times d_t}$
Multiple Feature View
Spatial Encoding + Metric Learning
Integration
Candidate Structures $A^{(r)}$
$g_{(r)}$
$w_{rel}^{(r)}$
X
Concat
Mesage Passing
Prediction
LOSS
Label Y
Integration
Spatial Encoding + Metric Learning

Fig. 1. Framework of CGPM

4.1 Cross-modality Feature Integration

The motivation of the Cross-modality Feature Integration module is to utilize the features from other source modalities to achieve the augmentation and supplementation of the target modality [11]. Therefore, we deploy a multi-head

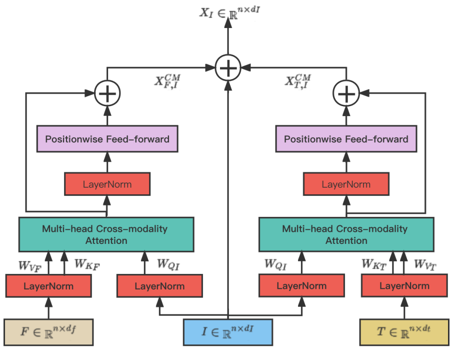
$X_I \in \mathbb{R}^{n \times dI}$
$X_{F,I}^{CM}$
$X_{T,I}^{CM}$
Positionwise Feed–forward
LayerNorm
Multi–head Cross–modality Attention
Positionwise Feed–forward
LayerNorm
Multi–head Cross–modality Attention
W_{VF}
W_{KF}
W_{QI}
W_{QI}
W_{KT}
W_{VT}
LayerNorm
LayerNorm
LayerNorm
LayerNorm
$F \in \mathbb{R}^{n \times df}$
$I \in \mathbb{R}^{n \times dI}$
$T \in \mathbb{R}^{n \times dt}$

Fig. 2. Structure of Cross-Modal Transformer

cross-modality transformer to achieve the integration of multi-modal information of clusters to be mapped.

Note that we use Image as the target modality in our discussion. While in the model architecture, all source modalities will be treated as target modality separately. In the computation of each channel of multi-head cross-modality attention, we construct an attention matrix, then obtain the transferred representations of features from source modalities, that can be written as:

$$X_{F,I}^{AttT} = MulH\left(\sigma\left(\frac{\left(IW_{Q_I}\right)^T\left(FW_{K_F}\right)}{\sqrt{d_k}}\right)\left(FW_{V_F}\right)^T\right) \tag{1}$$

Where $W_{Q_I} \in \mathbb{R}^{d_I \times d_k}$, $W_{K_F} \in \mathbb{R}^{d_f \times d_k}$ and $W_{V_F} \in \mathbb{R}^{d_f \times d_k}$ denotes the weighted parameter matrix of cross-modality attention. $\sigma(\cdot)$ is defaulted to a softmax activation function. $MulH(\cdot)$ is a multi-head function. $X_{F,I}^{Att} \in \mathbb{R}^{n \times d_k}$ is the output of multi-head cross modality attention. We added residual connections into the calculation of the mapping and deployed position-wise feed-forward to form a complete cross-modality transformer, which is written as:

$$X_{F,I}^{CM} = \text{conv } 1d\left(X_{F,I}^{Att}\right) + \text{relu}\left(X_{F,I}^{Att}W_1^{F,I} + b_1^{F,I}\right)W_2^{F,I} + b_2^{F,I} \tag{2}$$

Where $\text{conv } 1d(\cdot)$ is a 1-dimension convolutional layer deployed to adjust the residual dimension, and others are the formulation description of the positionwise feed-forward networks. $X_{F,I}^{CM}$ denotes the output of the cross-modal transformer about the transition from the source modality of discretized tabular features to the target modality of the image embedding. Eventually we set an attenuation coefficient to fuse the cross-modality information with the original information of the target modality, which is:

$$X_I = \alpha I + (1-\alpha)\left(\alpha^{ad}X_{F,I}^{CM} + \left(1-\alpha^{ad}\right)X_{T,I}^{CM}\right) \tag{3}$$

α_I and α_I^{ad} are settable attenuation coefficients. $X_{T,I}^{CM}$ denotes the output of the cross-modality module on the transition from text-modality features to the image modality. In the calculation process of the entire cross-modality transformer, the BatchNorm operation is deployed in each specific layer. Since it is a conventional optimization method, it is omitted from the formula description. Considering that our data source has three different modalities: Imagery, Text, and (tabular/categorical) Feature, we need to deploy 6 parallel channels of multi-head cross-modality transformers.

Finally, we concatenate integrated embeddings of the three target modalities, which is written as:

$$X = \mathrm{concat}\left(X_I, X_F, X_t\right) \tag{4}$$

Where concat$(\cdot)$ denotes a horizontal concatenation operation, X denotes the integrated cross-modality feature representation of clusters that we expect for the multimodal knowledge fusion process.

4.2 Feature-based Macroscopic Social Network Mining & Generating Module

The core function of this module is to generate the underlying macroscopic social network structure and optimize the feature representation of clusters(referred as nodes) using the learned graph structure. For nodes with high-dimensional features, inspired by Graphformer[25], we deploy the spatial encoding of the transformer layer to extract the global information of latent interactions, then generate candidate graphs to embed each cluster into high-dimensional space of underlying Macroscopic social networks, which can be written as:

$$G_{i,j}^{r} = \sigma\left(\frac{\left(x_i W_r^Q\right)\left(x_j W_r^K\right)^T}{\sqrt{d}}\right) \tag{5}$$

Where $W_r^Q, W_r^K \in \mathbb{R}^{d_k \times d_g}$ denotes weighted parameter matrices of the spatial coding, $G^r \in \{G^r\}_{r=1}^{R}$ denotes the r-th generated graph of the underlying social networks. (In order to modeling multiple semantics of interactions in social networks, we project to obtain $|R|$ candidate graph in total) Meanwhile, considering the assumptions of social networks that linkages tend to exist between pairwise nodes with significant homogeneity, we implement a H-head multi-channel metric-based approach to compute the similarity of features/embeddings of pairwise nodes using a cosine kernel function as a weight for candidate edges, which is:

$$S_{i,j}^{r} = \frac{1}{|H|}\sum_{h}^{H^r} \cos\left(w_s^{r,h} \odot x_i^T, w_s^{r,h} \odot X_j^T\right) \tag{6}$$

While elements of $w_s^{r,h} \in \mathbb{R}^{d_k}$ represent the importance of features in the measurement of similarity.

Eventually, we integrate the spatial encoding graph and node similarity graph by a structure propagation operation, and to augment the representation of underlying macroscopic social networks among clusters, which is

$$\tilde{A}^r = \mathrm{spar}\left(\sigma\left(\left(\eta_g G^r + \left(1-\eta_g\right) I\right)\left(\eta_s S^r + \left(1-\eta_s\right) I\right)\right), \varepsilon\right) \tag{7}$$

While η_g, η_s denote restriction coefficients of the propagations. $\operatorname{spar}(\cdot)$ is a sparsification function that enhances the sparsity of learned graph structure through dropping elements smaller than ε. $\left\{\tilde{A}^1, \ldots, \tilde{A}^R\right\}$ describes the generated underlying macroscopic social network obtained by CGPM. Furthermore, we initially distribute a semantic embedding g_r for each candidate graph/social network (defined as basis vectors or obtained by random walk). We compute node-level message passing[5, 20] under the precondition of considering the heterogeneous semantics of the underlying social networks, which is:

$$x_i^{(l)} = \sigma\left(\sum_{(j,r)\in\mathcal{N}(i)} A_{i,j}^r W_{mp}^{r,(l)T}\left(x_j^{(l-1)} \odot g_r^{(l-1)}\right) + h\left(x_i^{(l-1)}\right)\right) \tag{8}$$

Where l represents the depth of the message passing layer (deployed as a spatial graph convolutional layer), $\odot$ denotes a composition multiply operation, $\mathcal{N}(i)$ represents the set of clusters that have any social network structural connection with the cluster i under various semantics. $h(\cdot)$ and $g_r^{(l)} = W_{rel}^{(l)} g_r^{(l-1)}$ both denote a dimension alignment operation.

Eventually, the loss function of CGPM can be written as:

$$\mathcal{L}_{task} = \mathcal{L}_{rmse}\left(X^{\text{logit}}, Y\right) + \lambda\mathcal{L}_{reg}\left(X^{(l)}, \{A_r\}_{r=1}^R\right) \tag{9}$$

While $\mathcal{L}_{rmse}(\cdot)$ denotes a standard RMSE loss function and $\mathcal{L}_{reg}(\cdot)$ denotes a constraint function that prevents over-smoothing the vector representation of nodes and the learned macroscopic social network from being excessively dense.

4.3 CGPM-WS: for Weak Supervised Learning

Pseudo Labeling technology is the current solution to alleviate the challenge of excessive weak supervision in poverty mapping. Unlike existing general solutions which pseudo labels generated by other sub-models provide pseudo labeling refinement for CNNs(not mentioned previously), our proposed framework CGPM-WS is discussed as follows:

Cold-start Stage: We train 50 iterations of CGPM to obtain cross-modality integrated representations, as well as the parameters of each layer. Subsequently, CGPM-WS (Weak Supervision System) starts. The Cross-modality Transformer module (Paragraph 4.1) will be supplemented with the MLP(not mentioned previously layer and downstream tasks and moved to the **C area (Cross-modality Embedding Area)**, and the Macroscopic Social Network Mining module (Paragraph 4.2) will be moved to the **G area (Social Networks Generating Area)**. Meanwhile, CGPM-WS includes an **F area (Feature-based Model Area)**.

F Area: Take OSM data, and the features obtained from statistical description and artificial feature engineering of nightlight image and text data as inputs, construct LightGBM model (Adaboost's ensemble mode) for only DHS-WI labeled clusters. Furthermore, only when the first generation of the framework is executed, LightGBM will make predictions for all clusters, and select clusters

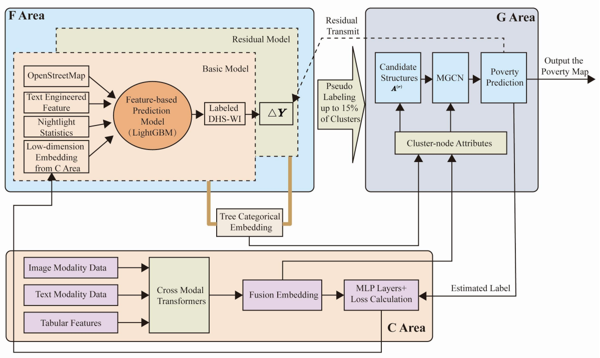

Fig. 3. Framework of CGPM-WS

whose predictions are the closest to the output of the pre-training, then utilize the predictions as pseudo labels, so that the proportion of labeled clusters can be supplemented to 15%. In all iterations of the system, 15% (ground-truth+ pseudo) will be input into the Macroscopic Social Network Generating module as estimated labels. In addition to the cross-modality representation output by the Cross-modality Integration Module, we concatenate the Tree Categorical Embedding (Pred_leaf parameter) output by LightGBM with it to complement the discretized tabular features.

G Area: We train the Social Networks Generating Module with received labels and features. The output will be submitted to the C area as the Refinement Label. After completing all training epochs, the output of the G Area is applied to draw the poverty map.

C Area: We directly connect MLP layers, Refinement Labels and standard loss functions to the downstream of the Cross-modality Transformer Module to obtain the updated features embedding of the sampled clusters. We weight and fuse the embedding vector on the Cross-modality Transformer side with the corresponding part of the node representation in the G area at a decay rate of 0.1, and submitted this embedding vector output from one layer of the MLP layers? (the second layer was selected in the experiment) to the F area. , concatenate with the features of the F area to provide a more informative representation for the LightGBM model.

Residual Transmittal: In early iterations of training, we make a residual between the output of the current round and the output of the previous round of the ground-truth labeled clusters and input the residual into the F area. The F area utilize the residual as labels, and take all the feature vectors received in this area as the input to append an additional LightGBM sub-model. When the F area provides estimated labels to the G area, the output is the summation of the prediction of all LightGBM sub-models.

Meanwhile, we emphasize/conclude? that CGPM-WS is a specialized model suitable for typical weak supervised environment.

5 Experiment

5.1 Implementation Details:

We select six representative developing countries in southern Asia and Africa as the experimental subjects: the Philippines(PHL), India(IND), Bangladesh(BAN), Tanzania(TZA?), Uganda(UGA), and Nigeria(NGA). The statistics of the sample points we collected and used in the experiments are shown in Table 1. For baselines, we choose TMPM [19] (transferable model modeled with satellite images and OSM data), WI-MMPM [15] (the most recognized end-to-end multimodal model, short for MMPM or WIPM) and HEPM [10] (specialized poverty estimation framework for weakly supervised learning), the above three methods are the most representative open source baselines.

For parameters, we set α_I and α_I^{ad} as 0.8, 0.5. The depth of Cross-modal Transformers and Message Passing are both set to 2. $\varepsilon = 0.1$, $H = 4$, $|R| = 5$ to ensure fair evaluation, the linear layer depth of all models is set to 2.

Table 1. Clusters in our Experiment.

	PHL	BAN	IND	NGA	UGA	TZ
Total	9970	8705	33076	23649	10700	8935
Labelled	1213	600	2058	1681	1011	1044
Precisely Geolocated	6488	5729	8914	11057	5163	5597

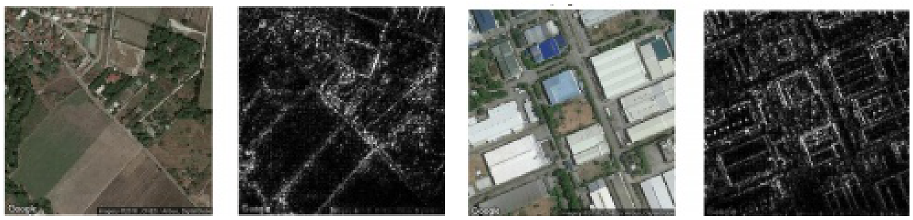

Fig. 4. Satellite data

5.2 Evaluation

Following previous research, we measured the coefficient of determination (R-squared) between the estimated wealth index of the model and the observed wealth index in the recent 5-year DHS surveys. The R-squared can be interpreted as the proportion of the variance for the observed wealth index that is explained by the estimated wealth index. Although the R-squared does not represent the accuracy of prediction precisely, it conveys a degree of performance in an intuitive way. Compared to RMSE, R-squared can be derived from econometrics showing whether our estimated index can be used to replace the DHS-WI index for relevant empirical research.

In the experimental setting, we set up three groups:

1) In the Semi-Supervised Group, 50% of the clusters are labeled (containing all labeled data points we have). We use 80% labeled clusters as a training set, 10% labeled clusters as a validation set, and 10% labeled and other unlabeled clusters as de facto validation set.

2) In the Weakly-Supervised Group, 8%-18% of the clusters are labeled (contains all the labeled data points we have, since the number of unlabeled data points varies, this ratio fluctuates between 8%-18%). We use 80% labeled clusters as training set, 10% labeled clusters as validation set, and 10% labeled and other unlabeled clusters as de facto validation set.

3) Cross-National Group: the experimental group which is to verify the generalization performance of our proposed method. In this group, we use the Philippines as the training set, transfer the trained model to other countries, and verify the performance. In CGPM and CGPM-WS, since the Macroscopic Social Network cannot be migrated, during the training process, we merge the target country and the Philippine clusters as the input, while using the labeled Philippine clusters as the training set and the validation set, the target country as the test set, thus to fairly evaluate the generalization of the model.

Table 2. Comprehensive Evaluation.

Performance Evaluation: Metric: Mean Pearson's R-square + Values Deviation						
	PHL	BAN	NGA	UGA	TZ	IND
Semi-Supervised Group: Labeled Ratio: 50%						
TMPM	0.679±0.004	0.713±0.017	0.641±0.004	0.732±0.026	0.68±0.009	0.627±0.01
MMPM	0.715±0.011	0.74±0.006	0.675±0.002	0.717±0.009	0.765±0.005	0.695±0.006
HEPM	0.73±0.007	0.726±0.003	0.626±0.001	0.768±0.011	0.669±0.013	0.633±0.006
CGPM	**0.824±0.005**	**0.78±0.002**	**0.737±0.004**	**0.793±0.003**	**0.775±0.007**	**0.764±0.001**
Weakly-Supervised Group: Labeled Ratio: 10-20%						
TMPM	0.656±0.019	0.593±0.017	0.597±0.016	0.674±0.021	0.64±0.011	0.639±0.033
MMPM	0.732±0.024	0.575±0.022	0.625±0.012	0.665±0.037	0.669±0.009	0.652±0.025
HEPM	0.783±0.009	0.718±0.006	0.682±0.008	0.721±0.009	0.683±0.006	0.707±0.011
CGPM	0.819±0.014	0.705±0.015	0.696±0.008	0.726±0.014	0.705±0.005	0.716±0.02
CGPM-WS	**0.845±0.005**	**0.759±0.003**	**0.731±0.005**	**0.773±0.012**	**0.724±0.002**	**0.748±0.017**
Semi-Supervised Cross National Experiment on Cross-Modality Feature Integration Training (on PHL)						
TMPM	/	0.691±0.016	0.476±0.031	0.575±0.024	0.557±0.016	0.598±0.006
MMPM	/	0.713±0.028	0.569±0.029	0.668±0.035	0.541±0.02	0.622±0.014
HEPM	/	0.699±0.009	0.413±0.025	0.585±0.018	0.462±0.013	0.617±0.022
CGPM	/	**0.747±0.013**	**0.63±0.01**	**0.692±0.004**	**0.575±0.007**	**0.709±0.008**

The results of the evaluation are reported in Table 2, from which we have the following observations: (a): In the Semi-Supervised Group, the multimodal end-to-end approach significantly outperforms the multimodal model ensemble method, while the performance improvement ratio is as high as eight percentage points. At the same time, CGPM significantly outperforms all other baseline methods in all country experiments, with an average advantage of about 6 percentage points. And considering the setting in this experiment, each module of our CGPM uses a lightweight deployment scheme, implying that there is still considerable room for improvements in the performance of CGPM. (b): In the Weakly-Supervised Group, the weakly supervised learning methods are significantly more suitable for this experimental setting. Both TTMPM and MMPM suffer a large performance loss, while HEPM and CGPM-WS based on CGPM improvement expands the performance advantage by about 5 percentage points, compared to the previous set of experiments. Among them, CGPM-WS outperforms the baseline schemes in all countries and has higher stability. (c) Our proposed CGPM exhibits good generalization performance and stability beyond the baselines in all multinational experiments. According to the experimental results, the effect of transnational migration experiments is relatively good between countries in the same geographic region or countries(effect is good?) with similar economic patterns and social development levels. Considering the performance and the stability in the three sets of experiments, CGPM is more suitable for poverty mapping deployed in countries with highly complex socioeconomic environments (Philippines, Bangladesh, and India).

In order to visually demonstrate the accuracy and stability of CGPM, we calculate the average of the model's prediction results for the residential areas by province in the Philippines and displayed it in Figure 5. From this, we can conclude that the accuracy of TMPM is slightly insufficient, MMPM has an overall shift in the predicted value, and the prediction result of CGPM is the closest to the Ground Truth(not mentioned previously or you used another term in the previous text) and has the highest accuracy.

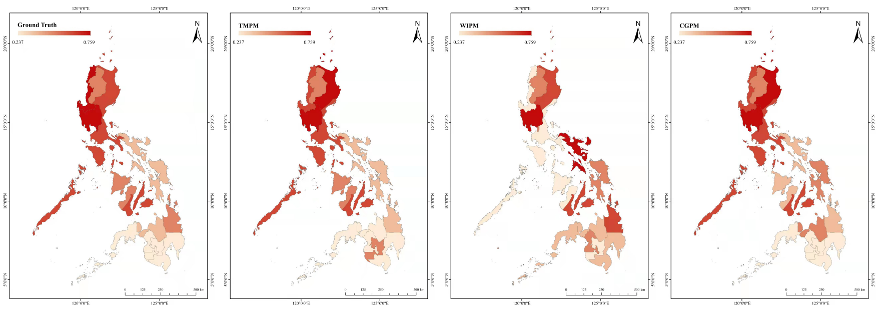

Fig. 5. Case Visualization, Philippines

Meanwhile, to demonstrate the advantages of CGPM-WS over CGPM in the weakly supervised learning domain, we run the CGPM and CGPM-WS models in the Weakly-Supervised Group respectively, and present them in the form of scatter plots in Figure 6. The visualization results show that CGPM-WS has higher accuracy in general, and the number of sample points with excessive prediction bias is significantly less than that of CGPM.

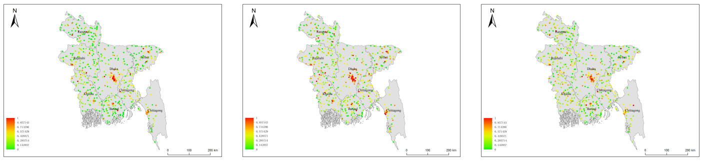

Fig. 6. Case Visualization for Weakly Supervised Learning, Bangladesh (Ground Truth, CGPM, CGPM-WS)

5.3 Ablation Study

In this section, we will verify the validity of each module of CGPM and CGPM-WS. Therefore we set reduction models CGPM-C, CGPM-N, CGPM-WS-C, CGPM-WS-N, which respectively remove the Cross-modality Feature Integration module, Feature-based Macroscopic Social Network Mining & Generating module from CGPM and CGPM-WS. Specifically, for CGPM-C, and CGPM-WS-C, we deploy a horizontal concatenation operation as the replacement of the margin. For CGPM-C, and CGPM-WS-C, we additionally deploy a 2-depth MLP layer for the output of Cross-modality Feature Integration module, as the connection with loss function.

Experimental Result in Figure 7 demonstrates the Social Network Generating module has significant effectiveness in countries with complex social structure, high degree of modernization, large population, frequent and prosperous economic activities. One possible reason is that there are numerous and important underlying macro-social network structures among the clusters in such countries, and they have interrelated socioeconomic effects that cannot be neglected. The Cross-modality Feature Integration module can stably improve the model performance in all countries, which means that modeling with knowledge-fused multimodal data will hopefully become a beacon for future poverty mapping research.

6 Conclusion

In this paper, we propose an end-to-end Poverty Mapping framework, CGPM, which is ahead of the academic frontier. CGPM innovatively realizes the cross-modal transformation and integration of multimodal data in this field, as well as the mining of underlying macroscopic social network structure, thus to optimize the representations of clusters. Meanwhile, we propose a variant of CGPM,

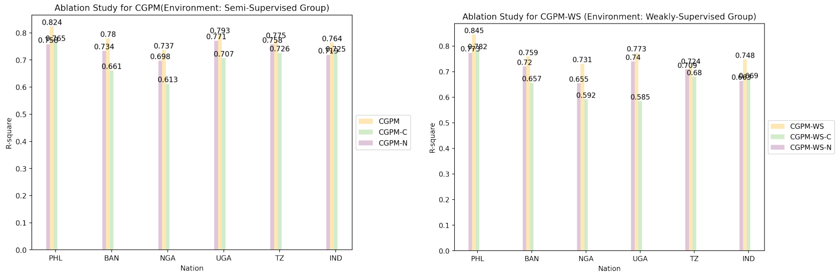

Fig. 7. Ablation Study

CGPM-WS, to specialize to overcome the weakly supervised learning challenges commonly found in Poverty Mapping. Extensive experiments demonstrate that CGPM and CGPM-WS significantly outperform the current baselines, and show more promising research prospects.

References

1. Ayush, K., Uzkent, B., Tanmay, K., Burke, M., Lobell, D., Ermon, S.: Efficient poverty mapping from high resolution remote sensing images. In: Proc. AAAI Conf. Artif. Intell. vol. 35, pp. 12–20 (2021)
2. Belhadj, B., Kaabi, F.: New membership function for poverty measure. Metroeconomica **71**(4), 676–688 (2020)
3. Fatehkia, M., Tingzon, I., Orden, A., Sy, S., Sekara, V., Garcia-Herranz, M., Weber, I.: Mapping socioeconomic indicators using social media advertising data. EPJ Data Science **9**(1), 22 (2020)
4. Flechtner, S.: Poverty research and its discontents: Review and discussion of issues raised in dimensions of poverty. measurement, epistemic injustices and social activism (beck, v., h. hahn, and r. lepenies eds., springer, cham, 2020). Review of Income and Wealth **67**(2), 530–544 (2021)
5. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Advances in neural information processing systems **30** (2017)
6. Htet, N.L., Kongprawechnon, W., Thajchayapong, S., Isshiki, T.: Machine learning approach with multiple open-source data for mapping and prediction of poverty in myanmar. In: 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). pp. 1041–1045. IEEE (2021)
7. Hu, S., Ge, Y., Liu, M., Ren, Z., Zhang, X.: Village-level poverty identification using machine learning, high-resolution images, and geospatial data. International Journal of Applied Earth Observation and Geoinformation **107**, 102694 (2022)
8. Jin, W., Ma, Y., Liu, X., Tang, X., Wang, S., Tang, J.: Graph structure learning for robust graph neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 66–74 (2020)
9. Ledesma, C., Garonita, O.L., Flores, L.J., Tingzon, I., Dalisay, D.: Interpretable poverty mapping using social media data, satellite images, and geospatial information. arXiv preprint arXiv:2011.13563 (2020)

10. Lee, K., Braithwaite, J.: High-resolution poverty maps in sub-saharan africa. arXiv preprint arXiv:2009.00544 (2020)
11. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Federated learning for vision-and-language grounding problems. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11572–11579 (2020)
12. Martínez, S., Rueda, M., Illescas, M.: The optimization problem of quantile and poverty measures estimation based on calibration. Journal of Computational and Applied Mathematics p. 113054 (2020)
13. Pilco, D.S., Rivera, A.R.: Graph learning network: A structure learning algorithm. arXiv preprint arXiv:1905.12665 (2019)
14. Roghani, H., Bouyer, A., Nourani, E.: Pldls: A novel parallel label diffusion and label selection-based community detection algorithm based on spark in social networks. Expert Systems with Applications **183**, 115377 (2021)
15. Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Burke, M., Lobell, D., Ermon, S.: Predicting economic development using geolocated wikipedia articles. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2698–2706 (2019)
16. Steele, J.E., Sundsøy, P.R., Pezzulo, C., Alegana, V.A., Bird, T.J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., De Montjoye, Y.A., Iqbal, A.M., et al.: Mapping poverty using mobile phone and satellite data. Journal of The Royal Society Interface **14**(127), 20160690 (2017)
17. Tang, J., Qian, T., Liu, S., Du, S., Hu, J., Li, T.: Spatio-temporal latent graph structure learning for traffic forecasting. arXiv preprint arXiv:2202.12586 (2022)
18. Thornton, P., et al.: Mapping poverty and livestock in the developing world, vol. 1. ILRI (aka ILCA and ILRAD) (2002)
19. Tingzon, I., Orden, A., Sy, S., Sekara, V., Weber, I., Fatehkia, M., Herranz, M.G., Kim, D.: Mapping poverty in the philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. In: AI for Social Good ICML 2019 Workshop (2019)
20. Vashishth, S., Sanyal, S., Nitin, V., Talukdar, P.: Composition-based multi-relational graph convolutional networks. arXiv preprint arXiv:1911.03082 (2019)
21. Wang, L., Chan, R., Zeng, T.: Probabilistic semi-supervised learning via sparse graph structure learning. IEEE transactions on neural networks and learning systems **32**(2), 853–867 (2020)
22. Watson, D., Whelan, C.T., Ma?tre, B., Williams, J.: Non-monetary indicators and multiple dimensions: The esri approach to poverty measurement. The Economic and Social Review **48**(4, Winter), 369–392 (2017)
23. Xie, M., Jean, N., Burke, M., Lobell, D., Ermon, S.: Transfer learning from deep features for remote sensing and poverty mapping. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
24. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 (2018)
25. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu, T.Y.: Do transformers really perform badly for graph representation? Advances in Neural Information Processing Systems **34** (2021)
26. Zhang, H., Xu, Z., Wu, K., Zhou, D., Wei, G.: Multi-dimensional poverty measurement for photovoltaic poverty alleviation areas: Evidence from pilot counties in china. Journal of Cleaner Production **241**, 118382 (2019)
27. Zhu, Y., Xu, W., Zhang, J., Du, Y., Zhang, J., Liu, Q., Yang, C., Wu, S.: A survey on graph structure learning: Progress and opportunities

28. Zhu, Y., Xu, W., Zhang, J., Liu, Q., Wu, S., Wang, L.: Deep graph structure learning for robust representations: A survey. arXiv preprint arXiv:2103.03036 (2021)