Targeted Policy Recommendations using Outcome-aware Clustering

ANONYMOUS AUTHOR(S)

Policy recommendations using observational data typically rely on estimating an econometric model on a sample of observations drawn from an entire population. However, different policy actions could potentially be optimal for different subgroups of a population. In this paper, we propose *outcome-aware clustering*, a new methodology to segment a population into different clusters and derive cluster-level policy recommendations. Outcome-aware clustering differs from conventional clustering algorithms across two basic dimensions. First, given a specific outcome of interest, outcome-aware clustering segments the population based on selecting a small set of features that closely relate with the outcome variable. Second, the clustering algorithm aims to generate near-homogeneous clusters based on a combination of cluster size-balancing constraints, inter and intra-cluster distances in the reduced feature space. We generate targeted policy recommendations for each outcome-aware cluster based on a standard multivariate regression of a condensed set of actionable policy features (which may partially overlap or differ from the features used for segmentation) from the observational data. We implement our outcome-aware clustering method on the Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) dataset to generate targeted policy recommendations for improving farmers outcomes in sub-Saharan Africa. Based on a detailed analysis of the LSMS-ISA, we derive outcome-aware clusters of farmer populations across three sub-Saharan African countries and show that the targeted policy recommendations at the cluster level significantly differ from policies that are generated at the population level.

ACM Reference Format:
Anonymous Author(s). 2022. Targeted Policy Recommendations using Outcome-aware Clustering. 1, 1 (January 2022), 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

INTRODUCTION

Policymakers and development practitioners aim at implementing policies designed to improve a population's outcomes. However, they often rely on little to no data on what impact the policy recommendations would have at the population level. In the scenarios when observational data is available, econometric models have allowed to determine which input variables have the strongest association with an outcome of interest and have provided guidance on policy recommendations aimed at changing the value of these inputs variables. A fundamental drawback of this approach is that the model would typically prescribe the same set of actions for each individual in a population. In reality, a policy which may appear as the optimal policy on average may not be the best fit at an individual or sub-population-level.

This paper specifically addresses the problem of determining targeted agricultural policy interventions for different sub-groups of the farmer population in Sub-Saharan Africa (SSA) to enhance agricultural outcomes with the ultimate goal of enhancing the livelihoods of the population in the region. The SSA region accounts for more than 950 million people, approximately 13% of the global population. By 2050, this share is projected to increase to almost 22% or 2.1 billion. Agriculture accounts for about 25% of Growth Domestic Product in SSA, and farming is the primary employment for about 60% of the population. Although that percentage is down from 80% a decade ago, it will remain a major component

Author's address: Anonymous Author(s).

Manuscript submitted to ACM

of economic activity in the SSA region in the coming decade. Given the key role of agriculture will continue to play, it is crucial to design policies aiming at promoting growth and sustainability in that sector.

In this paper, we propose *outcome-aware clustering*, a new methodology to segment a population into clusters that closely match the cluster feature variations with the outcome variations. Given a specific outcome of interest, the primary goal of outcome aware clustering is to segment the population into meaningful and related sub-groups. These clusters provide a framework to the development practitioners on the field, who can then personalize and choose the best outcome-specific predictive policy recommendation and customized support at a cluster-level granularity. This further bridges the gap between the econometric population level modeling, and the practical applicability on the field, where serving the development needs of individual clients is paramount.

Outcome-aware clustering fundamentally differs from the broad array of research on clustering and segmentation. Segmentation of a population, in general, focuses on grouping people into non-overlapping segments such that all the users in the same segment have similar needs and preferences. From a policy perspective, segmentation allows effective customization of policy recommendations to the particular preferences of each segment.

In outcome-aware clustering, the primary objective of clustering is centered on the outcome variable of interest. Conventionally, clustering algorithms have primarily centered around unsupervised learning. The popular k-means (and its variants k-medians, k-medoids, etc.), hierarchical clustering [29], and spectral clustering [26, 34] are notable examples. All these clustering approaches specify a distance/similarity measure between data points and determine the segments by optimizing a merit function that captures the quality of any given clustering. However, the distance function used in these clustering algorithms is independent of any outcome variable.

Outcome-aware clustering performs two key steps to directly tie the outcome variable with the clustering process. First, given a specific outcome of interest, outcome-aware clustering segments the population based on selecting a small set of features that closely relate with the outcome variable. Outcome-aware clustering measures distance between two users in the population in the reduced feature space. This step essentially makes the clustering process partially supervised. Second, the cluster generation algorithm aims to generate near-homogeneous clusters based on a combination of cluster size-balancing constraints, inter and intra-cluster distances in the reduced feature space.

While outcome-aware clustering normalizes each feature in the reduced space, it specifically does not tie the distance function used in the clustering algorithm to variations in the outcome variable. This is specifically to avoid any specific distance biases that the outcome variable may introduce with respect to specific features in the reduced space. Outcome-aware clustering is also designed for highly noisy contexts where the reduced features may only be weakly correlated with the outcome variable and may only provide limited information about the user with regards to the outcome of interest. Across many survey-based observational studies, especially with missing and noisy entries, we often encounter very few features (sometimes even zero) variables that may exhibit strong correlation with a given outcome variable. Outcome-based clustering is specifically designed to be robust in the face of the observational data having missing values or noisy features or the absence of any features that strongly correlate with an outcome variable.

Outcome-aware clusters can enable field staff to provide customized support based on cluster-level policy recommendations. The basic approach we use to generate targeted policy recommendations for each outcome-aware cluster is a standard multivariate regression based on a condensed set of actionable policy features that are regressed with the outcome variable. These condensed set of variables need to satisfy three properties: (a) Every variable from a policy perspective, needs to be *actionable*, where the policy recommendation is possible on the variable; (b) Every variable should have at least weak correlation with the outcome variable at the cluster level; (c) If a group of two or more variables, exhibit strong co-linearity among themselves, we reduce these set of variables to the most appropriate variable for the regression analysis.

We demonstrate how the outcome-aware clustering method can be used to the address the problem of improving farmers outcomes in several countries in sub-Saharan Africa (SSA), using data from the World Bank's Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA). Based on a detailed analysis of the LSMS-ISA, we derive outcome-aware clusters of farmer populations across three sub-Saharan African countries and show that the targeted policy recommendations at the cluster level significantly differ from the policies that are generated at the population level. Based on multiple years of LSMS-ISA surveys, we then demonstrate early evidence of movement of populations across clusters for the dominant cluster-specific policy recommendations.

RELATED WORK

The terms clustering and segmentation have typically been used interchabeably across a broad array of literature spanning multiple disciplines including statistics, machine learning and econometrics. We outline some of the key works that closely relate in spirit to our work. We refer the reader to [47] and [15] for a detailed review of the literature.

The most popular class of clustering algorithms is similarity based clustering, where each algorithm uses a specific distance/similarity measure between data points and determine the segments by optimizing a merit function that captures the "quality" of any given clustering. The popular k-means (and its variants k-medians, k-medoids, etc.), hierarchical clustering [29], and spectral clustering [26, 34] are notable examples. Another class of clustering algorithms is model-based clustering techniques [21, 49] which assume that each cluster is associated with an underlying probabilistic model and different clusters differ on the parameters describing the model. They estimate a finite mixture model [25] to the data and classify customers based on the posterior membership probabilities. However, as mentioned earlier, outcome-aware clustering fundamentally differs from these algorithms in that all these algorithms are completely unsupervised and are not tied to any specific outcome variable or objective.

Outcome-aware clustering also closely relates to customer segmentation literature in operations and statistics. One traditional method for predictive clustering is automatic interaction detection (AID), which splits the population into non-overlapping groups that differ maximally according to a dependent variable, such as purchase behavior, on the basis of a set of independent variables, like socioeconomic and demographic characteristics [4, 23]. Kamakura [16] proposed hierarchical segmentation techniques tailored to conjoint analysis, which group users such that the accuracy with which preferences/choices are predicted from product attributes or profiles is maximized. Cluster-wise regression methods [43, 44] cluster users in a population such that the regression fit is optimized within each cluster.

Latent class (or mixture) methods offer a statistical approach to the segmentation problem. Mixture regression models [41] simultaneously group subjects into unobserved segments and estimate a regression model within each segment, and were pioneered by Kamakura and Russell [18] who propose a clusterwise logit model to segment households based on brand preferences and price sensitivities. This was extended by Gupta and Chintagunta [12] who incorporated demographic variables and Kamakura et al. [17] who incorporated differences in customer choice-making processes, resulting in models that produce identifiable and actionable segments. Existing deep learning based clustering approaches use the dimensionality reduction capabilities of neural networks [50, 51] and learn clustering assignments from the resulting representation [52], but they lack interpretability with respect to the desired outcome. While outcome-aware clustering makes no specific assumptions about the features or the characteristics of the population, many of these latent approaches implicitly assume a mixture distribution characterization that describes the population.

ACHIEVING AGRICULTURAL TRANSFORMATION IN SUB-SAHARAN AFRICA

Dataset

To understand the factors improving farmers' standards of living, we use data from the LSMS-ISA survey. This survey consists in a nationally representative household panel data with a strong focus on agriculture and rural development. It was designed to improve the understanding of development in the SSA region, in particular of the linkages between farm and non-farm activities.

This survey has been implemented in eight countries in multiple waves. Most of our analysis will focus on the 2015 survey for Ethiopia. In section , we also show how our results can be extended to Tanzania and Uganda, comparing our main policy results across countries.

Before delving into the analysis, it is important to understand some of the limitations associated with using the LSMS-ISA dataset to conduct this analysis. First, a significant number of zeros and missing values limits the ability to draw inferences at a subpopulation level. We choose to discard survey answers with more than 30% of missing values. Second, we also drop variables which are not observed across multiple waves.

Relevant Outcomes and Inputs

A policy maker aiming at improving the living conditions of farmers in sub-Saharan Africa could choose to focus on a variety of outcomes: their revenue, level of expenditure, food expenditure diversification, whether they receive medical assistance when they are ill, whether they face food deficiency, etc. We find that among these outcomes of interest, the correlation is only 9% on average (Fig. 1a). This suggests that each outcome follows its own path, hence policy recommendations should be independently evaluated for each outcome.

In addition, while a large number of inputs could in principle play a role in farmers' living conditions, inputs with high correlation with outcomes are good candidates to consider when looking to improve farmers' outcomes. For the purpose of deriving policy recommendations, we distinguish between inputs that can be modified through short-term policy actions ("actionable") from those that cannot ("non-actionable").

We find that for inputs with high correlation with outcomes variables, while these correlations typically have the same sign across outcome variables, their magnitude tend to vary substantially (Fig. 1b and c). As correlation between outcomes are low, it is not surprising that the effect of a given input will vary across outcomes, reinforcing the conclusion that policy recommendations need to be outcome specific. We also find that even the most impactful input variables only have a 10% correlation with outcome variables on average, leading to a set of less than 10 actionable inputs likely to have an substantial impact on a given outcome.

METHODOLOGY

Generating policy recommendations can be thought of as a problem of extracting features which are predictive of an outcome intended by the policy. Given a set of n features F in an input variable matrix X, an outcome variable y, we intend to identify the best set of features P which would predict the outcome variable. We now describe our approach in the rest of the section. First, we cluster the features using a novel *outcome-aware clustering* algorithm. We then learn a regression model for each of these clusters separately to identify important actionable variables which significantly predict the outcome variable.

Fig. 1. **Relationship Between Farmers' Outcomes and Inputs:** (a) Spearman correlations between farmers' outcomes, showing a low average correlation equal to 0.09, and suggesting that policy recommendations should be derived for each outcome separately. We also show the Spearman correlations between farmers' outcomes and inputs, separating (b) non-actionable from (c) actionable inputs, and ranking inputs by their average correlation across outcomes. These subplots indicate that for inputs with the high correlations with outcome variables, correlations across outcomes are of similar sign but vary in strength, reinforcing that separate analyses should be conducted for each outcome of interest. For inputs with low average correlations with outcome variables, correlations across outcomes tend to vary both in sign and in strength.

Outcome aware clustering

We define outcome aware clustering as the problem of choosing a subset of features C such that the unsupervised clusters on these features effectively separate both the input features and the outcome variable across these clusters.

Prior to doing any clustering, it is essential to ensure that we don't incorporate features with a large fraction of missing values. Since most features in our study are categorical in nature, using any form of imputation or matrix completion techniques on these would not be sound. Hence, a simple threshold based filtering is used. Normalization of the features used for clustering is done by applying the z-score method.

In addition to finding the features to cluster on, we need to fix on the number of clusters to learn in a commonly used k-means clustering. During each step of making the choices of features to cluster on, we identified k using the elbow method and the average euclidean distance from the centroids across a range of k ∈ [1,10].

As explained in Algorithm 1, we initialize C as an empty set and iteratively add features to C in a greedy fashion. In each iteration, we choose a feature which maximizes a weighted silhouette coefficient for the k-means clustering obtained by including the feature in the clustering set C. This weighted silhouette coefficient (sc) combines the sc as measured in the clustering feature space as well as the single dimensional outcome space. The outcome awareness is controlled by a parameter $\alpha \in [0,1]$. We can see that $\alpha = 0$ is equivalent to traditional unsupervised clustering on the input feature space, whereas $\alpha = 1$ is equivalent to bucketization based only on the outcome variable. With α between 0 and 1, the clustering achieves two objectives. First, we identify a clustering which can separate the clusters based on the outcome variable, allowing to design policy recommendations at various outcome levels. Second, it separates the input features space which is critical to identifying these clusters when the outcome variable is not observed in an unsupervised manner.

Algorithm 1: Feature choice for clustering

$F := \{f_1, f_2, f_3, .., f_n\}$, input features
$y :=$ output feature
$\alpha \in [0,1]$, Output awareness parameter
$C := \emptyset$
$\epsilon :=$ Threshold of k-means silhouette coefficient (sc) improvement
while $\Delta sc > \epsilon$ **do**
  for f in $F \backslash C$ **do**
    $l_f = Kmeans(f \cup C)$
    $sc_{y,f\cup C} = \alpha * sc_y(l_f) + (1-\alpha) * sc_{f\cup C}(l_f)$
  end for
  $f_{opt} = \underset{f \in F\backslash C}{\text{argmax}}\, sc_{y,f\cup C}$
  $\Delta sc = sc_{y,f_{opt}\cup C} - sc_{y,C}$
  $C := f_{opt} \cup$ C
end while
return C

A benefit of choosing the features iteratively is that we don't end up with redundant features which explain the same feature space and outcome level. This ensures that the final set of features can distinguish between any pair of clusters using only a subset of these features. This can be thought of increasing the information criterion of the clusters iteratively. Hence, some of the features chosen during the iterative steps could have low outcome correlation values at the population level, but are instrumental in distinguishing certain specific outcome clusters. In each step, the k-means also enforces that each cluster is of a certain minimum size to avoid learning behavior of statistical outliers, and guarantee that we have enough observation to derive cluster-level policy recommendations.

The stopping condition of iterations is based on the improvement in the silhouette coefficient over the iterations, and the threshold (ϵ) can be chosen in a problem specific manner. Once the feature set C is chosen, we have also jointly learnt the corresponding k-means clusters. It can be noted that our algorithm is generic and can accommodate any unsupervised clustering method and operates as a layer above it.

Policy recommendations through regression

The fundamental contribution of our approach is that we learn different policy recommendations for different clusters of households. These variations in policy recommendations across clusters are not evident if done at a population level.

As shown in Algorithm 2, choosing features for regression is done in a principled two step approach. First, we used highly correlated features with the outcome, where a threshold (β) on the spearman correlation coefficient (ρ) was used for filtering. Second, in order to eliminate multi-collinearity in the correlated features, we iteratively eliminated the feature with the highest variance inflation factor (VIF) above a certain threshold (γ). These thresholds were identified using an appropriate grid search to ensure that a reasonable set of policy recommendations were identified. The filtered features are then used in a linear regression model to predict the outcome variable for each cluster. Statistically significant coefficients of this model are then used to derive policy recommendations for each cluster.

Algorithm 2: Regression based Policy Recommendations

$C := \{c_1, c_2, .., c_k\}$, the set of clusters
$F := \{f_1, f_2, f_3, .., f_n\}$, set of actionable features
$y := (obs, 1)$ output matrix
$X := (obs, n)$ input matrix
$\beta :=$ Output correlation threshold
$\gamma :=$ Input multi-collinearity threshold
$F_{corr} = \{f_i | \rho(y, X[f_i]) > \beta\}$
repeat
    $f_{max} = \underset{f \in F_{corr}}{\text{argmax}}\ VIF(X[F_{corr}], X[f])$
    $F_{corr}.remove(f_{max})$
until $(VIF(X[F_{corr}], X[f_{max}]) < \gamma)$
for c in C **do**
    $coeff_c = OLS(X_c[F_{corr}], y_c)$
    $P_c :=$ stat-significant $coeff_c$
end for
return $\bigcup_{c \in C} P_c$

RESULTS

Clustering Farm Households

Next, we experiment the clustering method that we have developed on the 2015 LSMS-ISA survey of Ethiopia. We focused on farmers' crop sales as our outcome of interest. Our algorithm suggested to cluster farm households based on the following inputs: their total land surface, household size, the number of oxen they own, the number of ploughs they own, whether or not they participate in an extension program, the quantity of chemical fertilizers they use, and their number of hired workers. These inputs are indeed among those having the highest correlation with crop sales. We then allocate households into four clusters as suggested by the Elbow method (Fig. 2c).

We find that our clustering method indeed allows to construct clusters in which households crop sales are similar within each cluster and different across clusters (Fig. 2a). On average, crop sales increases monotonically across clusters,

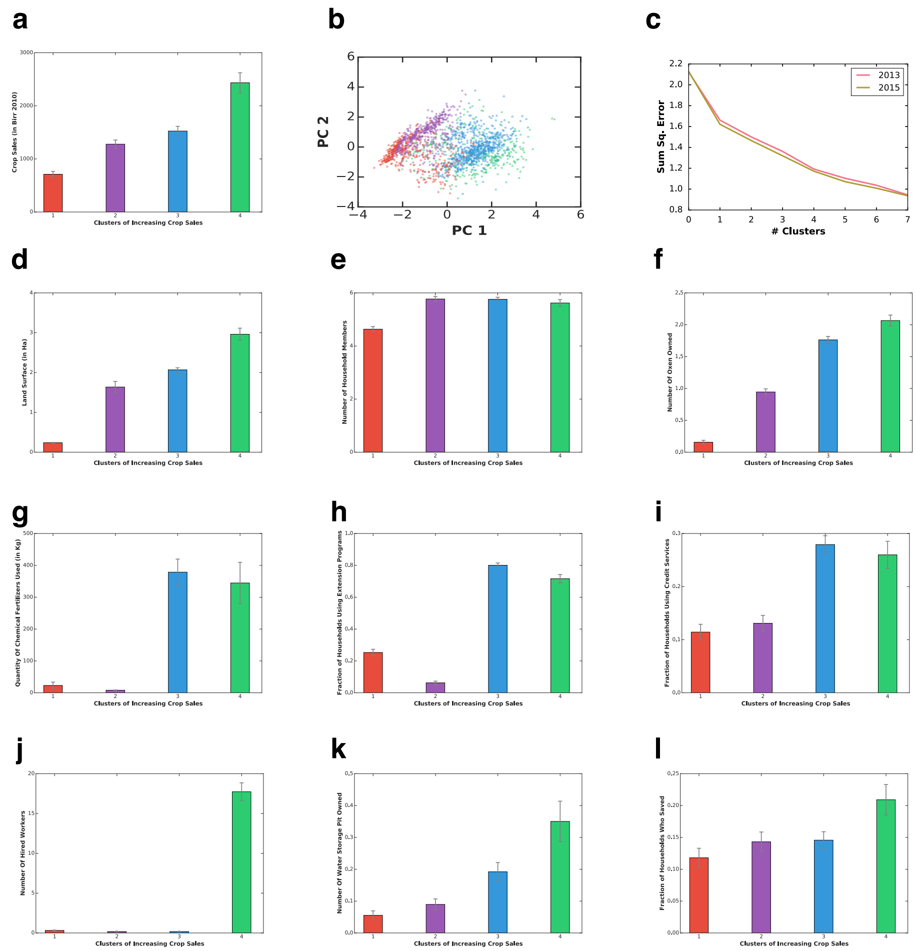

Fig. 2. **Clustering Results:** (a) Average crop sales across clusters, indicating that our method allows to construct clusters such that households outcomes are similar within each cluster and different across clusters. (b) The two principal components of our clustering features across households, indicating that our method allows to construct clusters such that households clustering inputs are similar within each cluster and different across clusters. (c) Sum of square errors of K-means clustering, showing that the error is stable across survey waves. The elbow method indicates that the optimal number of clusters is 4. To understand the composition of the resulting clusters, we then show the average value across clusters of the three features with the highest relative change occurring between cluster one and two (d-f), between cluster two and three (g-i), and between cluster three and four (j-k).

ranging from 711 Birr to 2,424 Birr. Projecting our clustering inputs on their first two principal components, we also find that our method allows to construct clusters in which clustering inputs are similar within each cluster and different across clusters (Fig. 2b).

Compared to all of the richer clusters, households in the first cluster only own 0.24 Ha of land on average, which is 6.8 times less than households in the second cluster (Fig. 2d). They are comprised of five members on average, compared to six for the other clusters (Fig. 2e). They are five times less likely to own an ox (Fig. 2f) and 2.4 times less likely to own a plough (Table 1) compared to households in the second clusters. 28% of them are female-headed households (Table 1), which is 1.8 times more than in the other clusters, and they are predominantly located in the SNNP region (Fig. 3). These are the poorest households in our sample; they do not have the means to own large properties nor the ability to purchase basic tools required to harvest efficiently.

Households in the second clusters generate 1.8 times more revenue and are better equipped than those of the first cluster. Yet, they still do not use significant amounts of fertilizers (Fig. 2g) or improved seeds (Table 1) to increase their productivity compared to those in the third or fourth cluster. Only about 13% of households in the first two clusters participate in an extension program, and only about 13% of them use damaged prevention techniques, compared to about respectively 76% and 22% of those in the last two clusters (Fig. 2h and Table 1). Only about 12% of households in the first two clusters use credit services, compared to about 27% of those in the third or the fourth cluster (Fig. 2i).

The richest households are located in the fourth cluster, with a average income 60% larger than those of the third cluster. They are mainly characterized by their ability to hire workers (Fig. 2j). 22% of them save money, compared to less than 15% of households in third clusters and below. They also tend to acquire more sophisticated or more expensive tools. They are 2.1 times more likely to own a pick ax (Table 1), and 1.5 times more likely to own an ax (Table 1) compared to those in the third cluster or below.

Taken together, these results show that the clusters derived from our *outcome-aware clustering* are robust and correspond to interpretable subpopulations of households.

Policy Recommendations

Having constructed robust and interpretable clusters, we now ask whether we can derive policy recommendations at the cluster level, and whether these recommendations differ from those obtained at the population level.

As our analysis is conducted on a relatively small dataset, we choose to estimate a multivariate regression model of crop sales using a restricted set of policy variables. We apply algorithm 2 choosing the two following parameters: (a) we remove any policy variable that has a correlation with crop sales of less than $\beta = 0.05$, and (b) we iteratively remove policy variables until the VIF scores of the remaining variables is less than $\gamma = 1.5$. This guarantees that the selected variables will have a substantial impact on the outcomes, and will remove collinear policy variables from the model. In a robustness check, we found that our results hold for a wide range of values for β and γ, other specifications typically leading to a larger set of insignificant variables being included in the model.

The number of hired workers has the strongest coefficient in the full sample regression (Fig. 4a). As the standard deviation of crop sales is equal to 1,169 Birr, hiring one additional worker is associated with an increase in income of $0.25 \times 1,169 = 292$ Birr. The effect of hiring workers on crop sales is U-shaped, with the largest effect concentrated in the first cluster where the coefficient is equal to 0.7. It indicates that policies should primarily focus on encouraging farmers to hire workers, especially in the first and the fourth cluster. Possible implementations could be to subsidize workers hiring costs, develop or improve systems providing information on labor market conditions, etc. It is important to note that our analysis does not account for the costs of implementing such policies. Hiring workers could be quite costly, especially for low income households.

The second most impactful factor corresponds to the use irrigation techniques (Fig. 4b). Households using irrigation have an average revenue that is 128 Birr higher than those who do not. Here, the effect is also U-shaped: it is positive and significant for households in the first and the fourth clusters, but it is insignificant for those in the second and third cluster.

An increase in the quantity of chemical fertilizers used by one standard deviation or in the number of axes owned by one unit are associated with a small increase in income of 105 Birr and 47 Birr respectively (Fig. 4c and f). This effect is concentrated on households in the first cluster, the effect being insignificant for the remaining clusters. This suggests

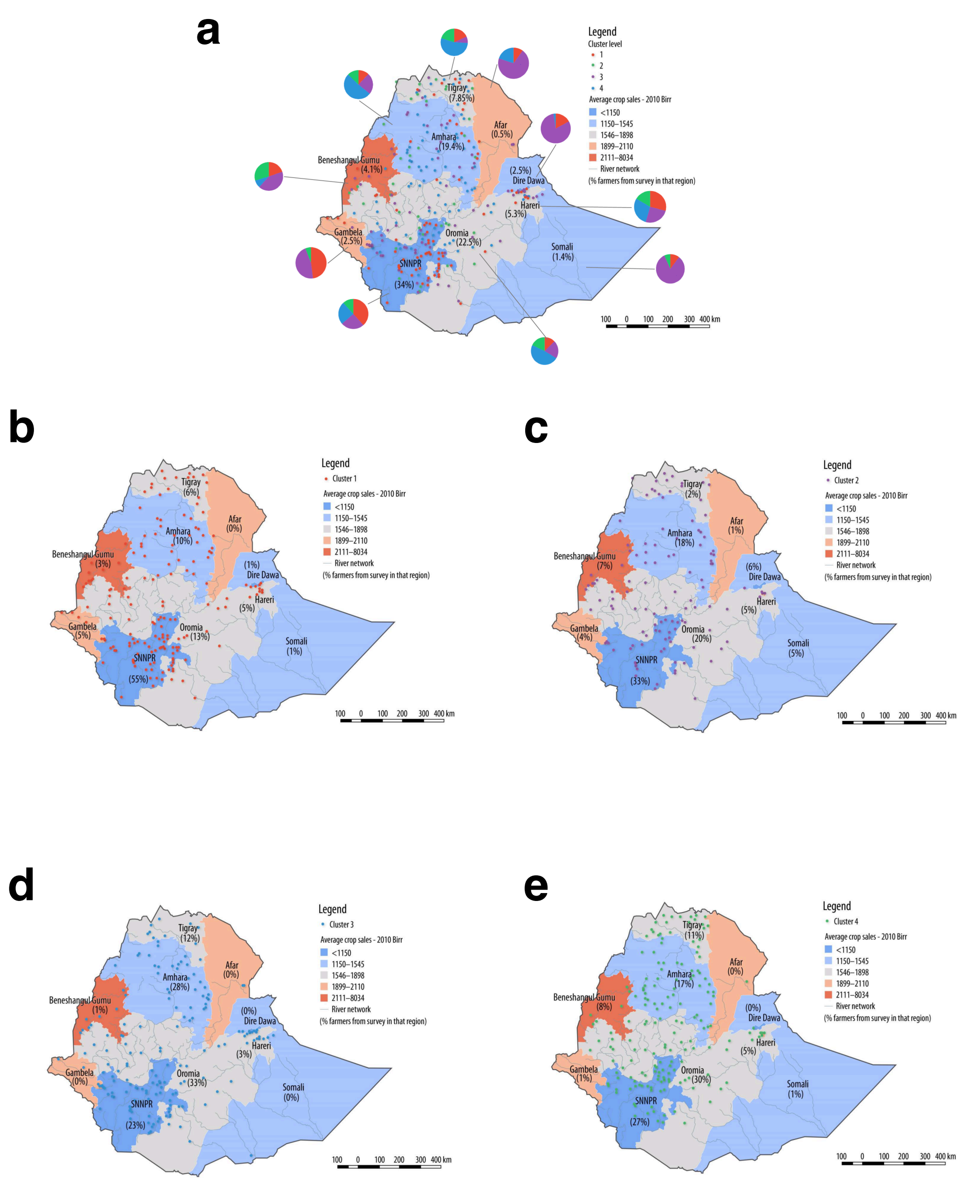

Fig. 3. **Geography of Clusters:** Each dot corresponds to a household colored by its cluster.

that policy aiming at improving the income prospects of households in the first and second clusters specifically could be targeted towards reducing the costs of acquiring additional tools or fertilizers through subsidies or conditional cash transfers.

Finally, households using damage prevention techniques or saving money generate on average 94 Birr and 82 Birr respectively more than those who do not (Fig. 4d and e). The effect is concentrated on households in the third and fourth cluster and is insignificant for households in the first and second cluster. This suggests that policies targeted towards the third or the fourth cluster could focus on raising awareness on the benefits of damage prevention techniques, or incentivize farmers to save money using their mobile phone.

Taken together, these results show that *outcome-aware clustering* allowed us to derive policy recommendations at the cluster level, showing that they often differ from those that would be optimal at the population level.

Cross-country Comparison

Next, we compare the results that we obtained in Ethiopia to other countries included in the LSMS-ISA survey. We apply outcome-aware clustering on the 2014 survey for Tanzania and the 2013 survey for Uganda, deriving policy recommendations at the cluster level. Although cross-country comparisons are limited by a lack of homogeneity in how key policy variables are measured across countries, it is nonetheless interesting to test whether some consistent patterns emerge.

The amount of pesticides used has the strongest association with crop sales, both for Tanzania and for Ethiopia. In both cases, the effect is slightly decreasing across clusters (Fig. 4g and l).

The next variable with the strongest association with crop sales both for Tanzania and for Uganda is the amount of fertilizers used (Fig. 4h and l). The strength of the effect is U-shaped across cluster for Tanzania, and has an inverted U-shaped for Uganda, which differ from the pattern observed for Ethiopia. These differences could be explained by variations in the variety of crops that are being grown, the relative returns to using fertilizers, or the types of fertilizers being used.

For Tanzania, owning a plough has an effect on crop sales that is mostly concentrated in the first cluster (Fig. 4i). This is consistent with the effect of owning an axe being concentrated in the first cluster in the case of Ethiopia.

In the case of Uganda, the effect of hiring workers is not as predominant as in the case of Ethiopia (Fig. 4m), yet we observe a similar U-shape behavior.

Finally, having a bank account in Tanzania is only associated with generating more revenue for households in the third and fourth cluster, which is similar to the effect of saving observed for Ethiopia. Similarly, borrowing is associated with a reduction in income only for households in the fourth cluster in the case of Uganda.

Validating Predictions Over Time

To validate our policy recommendations, we do a longitudinal evaluation tracking households across 3 waves of surveys done in Ethiopia, with a gap of 2 years between each wave.

For a majority of households, the value of key inputs remain constant between surveys, limiting the ability to test the validity of our predictions over time. We focused on households' “number of hired workers", as it is most impactful input coming out of the model predictions, the other inputs being associated with insignificant evidence of movement between waves.

We found evidence of a lift in the increase crop sales associated with hiring an additional worker being equal to 0.39, 0.23, 0.26 and 0.57 across clusters of increasing income (Fig 5). This indicates that households in the first cluster who hired an additional worker between two consecutive wave are 39% more likely to have had an increase in crop sales during the same period that those who did not, similar conclusion being drawn for the other clusters. Interestingly, we find a U-shape in the value of the lift factor associated with hiring an additional worker, which mimics the variations in coefficient strengths obtained in the multivariate regression. Although additional data would be needed to provide further evidence, this gives some initial validation for our approach.

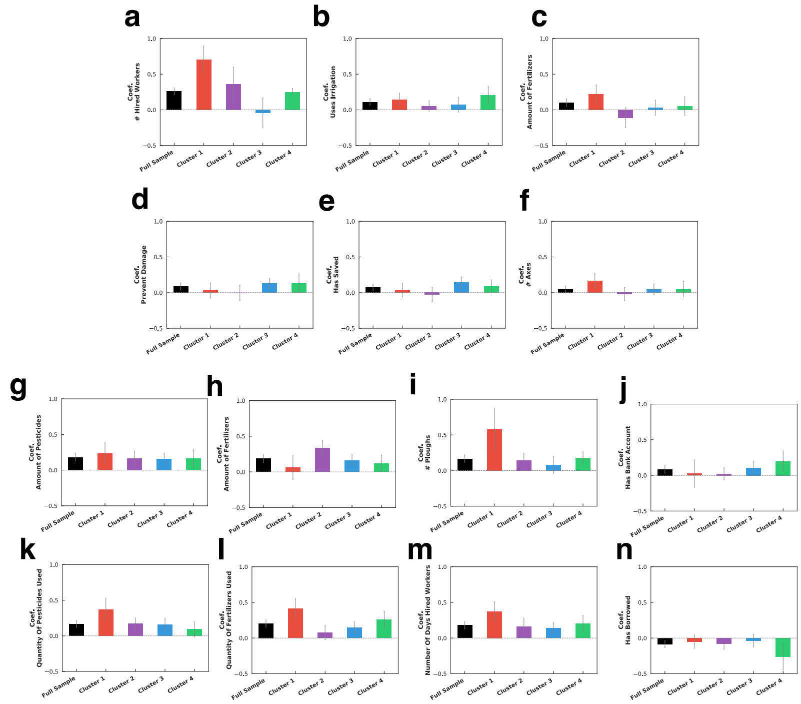

Fig. 4. **Policy Recommendations:** Regression coefficients of a multivariate regression of crop sales on a set of selected policy variables, for the entire sample (black), and per cluster of increasing crop sales. Coefficients are ranked by decreasing value on the entire sample. The first two rows corresponds to the 2015 survey for Ethiopia, the third row corresponds to the 2014 survey for Tanzania, and the fourth row corresponds to the 2013 survey for Uganda. This plot shows that the effect of the most impactful variables vary significantly across clusters, indicating that policy recommendations should indeed be cluster-specific.

CONCLUSIONS

This paper presents *outcome-aware clustering*, a new clustering methodology to segment a population into meaningful clusters corresponding to a specific outcome of interest. Unlike traditional unsupervised clustering and mixture modeling approaches for population segmentation, *outcome-aware clustering* relies on choosing a set of clustering features closely related to an outcome of interest, while minimizing intra-cluster and maximizing inter-cluster distances. We demonstrate the utility of this *outcome-aware clustering* methodology to enable field practitioners to provide personalized and customized cluster-level policy recommendations. Using data from the LSMS-ISA survey across three countries in Sub-Saharan Africa, we found that our method provides actionable and highly predictive cluster-level policy recommendations which significantly differ from those obtained at the population level.

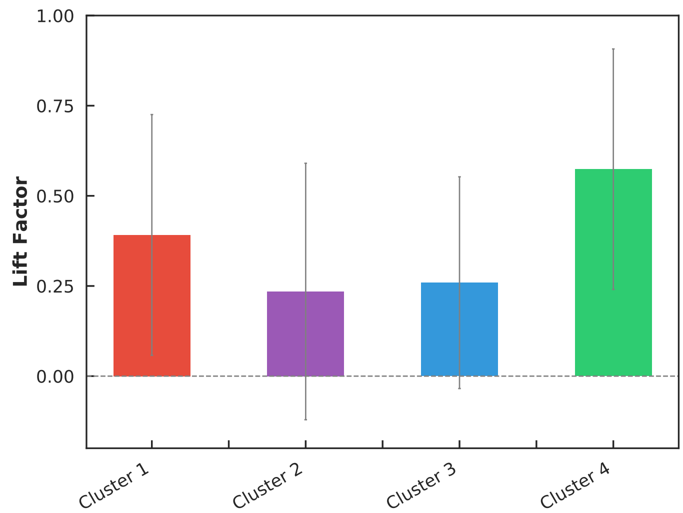

Fig. 5. **Evidence of Movement Between Clusters:** For each cluster, the lift factor associated with a given input measures the fraction of households whose income increases beyond a given threshold during two consecutive survey wave when the value of that input also increased, relative the fraction of households whose crop sales increased beyond the same threshold. We pick the threshold to correspond to the 25%ile of the distribution of changes in crop sales for each cluster and each wave. We only show the lift associated with hiring additional workers, the lift associated with less impactful policy inputs being insignificant.

	Cluster 1 Avg.	Stdev.	Cluster 2 Avg.	Stdev.	Cluster 3 Avg.	Stdev.	Cluster 4 Avg.	Stdev.
Amount Of Assistance Received	51.036	228.505	84.745	356.512	51.177	248.418	57.666	381.920
Attended School	0.320	0.389	0.295	0.384	0.290	0.379	0.363	0.427
Average Precipitation	1271.081	277.231	1210.551	354.563	1228.689	303.100	1260.963	326.223
Average Temperature	181.735	24.328	190.300	32.059	175.884	25.541	192.661	30.492
Children Education	0.696	0.331	0.700	0.343	0.740	0.310	0.719	0.326
Number of Crops Planted	3.638	2.642	2.948	3.184	2.831	3.332	2.778	3.754
Crop Sales (in Birr 2010)	711.440	1170.191	1277.643	1768.510	1524.358	2373.700	2427.224	3195.684
Distance To Market	63.565	42.324	72.449	48.111	60.292	42.312	67.126	46.472
Distance To Population Center	27.208	20.145	40.966	26.594	32.082	19.888	40.130	32.136
Distance To Road	11.713	12.511	17.654	20.798	12.230	11.732	11.940	13.820
Elevation	1998.337	411.404	1910.413	501.948	2138.493	412.222	1850.324	472.553
Non-food Expenditure (in Birr 2010)	1065.626	1460.336	1231.502	1287.209	1775.106	1358.489	2397.284	2195.733
Food Expenditure Diversification	0.840	0.169	0.846	0.147	0.871	0.120	0.875	0.106
Fraction of Households With A Bank Account	0.045	0.207	0.020	0.142	0.049	0.216	0.077	0.267
Has Borrowed	0.227	0.419	0.271	0.444	0.309	0.462	0.253	0.435
Fraction of Households Using Medical Assistance	0.198	0.291	0.210	0.249	0.231	0.268	0.272	0.273
Fraction of Households Who Saved	0.116	0.320	0.144	0.351	0.146	0.354	0.208	0.406
Heavy Rains Preventing Work	0.041	0.219	0.035	0.199	0.039	0.252	0.031	0.413
Household Head Age	47.068	16.779	48.417	15.360	47.426	13.861	46.572	13.659
Fraction of Divorced	0.073	0.261	0.030	0.168	0.016	0.123	0.005	0.072
Fraction of Female-headed Households	0.278	0.448	0.094	0.292	0.147	0.354	0.114	0.318
Fraction of Male-headed Households	0.722	0.448	0.906	0.292	0.853	0.354	0.886	0.318
Household Head Is Monogamous	0.718	0.450	0.854	0.351	0.845	0.361	0.825	0.379
Household Head Is Polygamous	0.024	0.152	0.038	0.191	0.023	0.149	0.067	0.250
Household Head Is Separated	0.003	0.057	0.002	0.042	0.002	0.047	0.008	0.091
Fraction of Widow	0.174	0.379	0.076	0.263	0.103	0.303	0.091	0.287
Household Head Never Married	0.007	0.085	0.000	0.020	0.012	0.109	0.003	0.053
Number of Household Members	4.637	2.087	5.773	2.196	5.754	2.085	5.621	2.158
Illness Of Household Member	0.300	1.032	0.389	1.223	0.294	0.817	0.345	0.864
Increase In Price Of Inputs	0.172	0.474	0.172	0.514	0.265	0.519	0.363	0.553
Land Surface (in Ha)	0.239	0.144	1.638	3.249	2.066	1.400	2.953	2.595
Latitude	7.879	2.021	9.057	2.320	9.361	1.880	9.076	2.058
Literacy Rate	0.325	0.381	0.318	0.369	0.335	0.372	0.405	0.392
Lives In Afar	0.000	0.015	0.001	0.036	0.000	0.015	0.000	0.000
Lives In Amhara	0.126	0.332	0.303	0.460	0.310	0.462	0.235	0.424
Lives In Benishangul Gumuz	0.014	0.119	0.030	0.171	0.004	0.064	0.035	0.183
Lives In Dire Dawa	0.001	0.038	0.007	0.086	0.000	0.009	0.000	0.000
Lives In Gambella	0.006	0.076	0.009	0.095	0.000	0.000	0.001	0.036
Lives In Harari	0.002	0.047	0.002	0.048	0.001	0.033	0.003	0.052
Fraction of Households Living in Oromiya	0.169	0.374	0.352	0.478	0.517	0.500	0.507	0.500
Lives In Snnp	0.650	0.477	0.266	0.442	0.121	0.327	0.161	0.367
Lives In Somalie	0.001	0.025	0.017	0.129	0.000	0.000	0.005	0.069
Lives In Tigray	0.031	0.173	0.011	0.106	0.046	0.210	0.054	0.225
Longitude	38.128	1.198	38.102	1.807	38.190	1.411	37.767	1.521
Fraction of Households Without Food Deficiencies	0.466	0.499	0.689	0.463	0.773	0.419	0.836	0.368
Number Of Axe Owned	0.651	0.695	0.682	0.851	0.545	0.848	0.888	1.065
Number Of Droughts	0.283	0.604	0.434	1.134	0.207	0.567	0.259	0.509
Number Of Hired Workers	0.317	1.075	0.170	0.589	0.177	0.574	17.680	19.054
Number Of Oxen Owned	0.157	0.648	0.950	1.124	1.759	1.449	2.058	1.461
Number Of Pick Axe Owned	0.581	0.720	0.776	0.861	0.831	1.121	1.715	4.761
Number Of Plough Owned	0.315	0.540	0.770	0.634	1.220	0.885	1.239	1.046
Number Of Sickle Owned	1.016	1.011	1.576	1.325	2.155	1.703	2.067	1.766
Number Of Water Storage Pit Owned	0.055	0.306	0.090	0.395	0.192	0.773	0.349	1.081
Fraction of Households Who Own A Land Certificate	0.429	0.486	0.541	0.478	0.665	0.443	0.622	0.447
Percentage Of Damaged Crop	12.551	16.531	21.273	23.839	17.784	20.482	17.693	19.463
Prevent Damage	0.133	0.310	0.124	0.241	0.236	0.288	0.205	0.264
Price Rise Of Food Item	0.304	1.204	0.372	1.365	0.155	0.446	0.158	0.610
Yield (in BIRR per Acre)	5626.935	29955.749	1264.107	1960.196	859.062	1066.618	1278.399	2122.692
Quantity Of Chemical Fertilizers Used (in Kg)	22.925	229.296	7.733	19.063	378.077	1093.361	343.620	1103.675
Quantity Of Improved Seeds Used (In Kg)	2.104	4.641	0.916	5.102	11.835	48.431	12.767	54.739
Rooting Conditions : Mainly Non-Soil	0.003	0.056	0.004	0.064	0.004	0.065	0.000	0.013
Rooting Conditions : Moderate Constraint	0.324	0.468	0.138	0.345	0.184	0.388	0.193	0.395
Rooting Conditions : No Or Slight Constraint	0.466	0.499	0.503	0.500	0.541	0.498	0.618	0.486
Rooting Conditions : Severe Constraint	0.084	0.278	0.202	0.401	0.146	0.353	0.059	0.236
Rooting Conditions : Very Severe Constraint	0.123	0.329	0.153	0.360	0.125	0.330	0.130	0.337
Rural Household	0.960	0.197	0.970	0.171	0.997	0.053	0.985	0.120
Fraction of Households Using Credit Services	0.112	0.315	0.131	0.336	0.280	0.444	0.259	0.434
Fraction of Households Using Extension Programs	0.251	0.433	0.063	0.242	0.800	0.392	0.714	0.448
Uses Irrigation	0.025	0.136	0.029	0.142	0.027	0.105	0.026	0.099
Variations In Greenness	45.215	7.021	45.538	10.094	48.546	8.266	48.560	9.903

Table 1. **Clusters' Descriptive Statistics**

REFERENCES

[1] Abu-Mostafa, Yaser S., Malik Magdon-Ismail, Hsuan-Tien Lin. 2012. *Learning From Data*. AMLBook.

[2] Achlioptas, Dimitris, Frank McSherry. 2005. On spectral learning of mixtures of distributions. *Learning Theory*. Springer, 458–469.

[3] Anandkumar, Animashree, Rong Ge, Daniel Hsu, Sham M Kakade, Matus Telgarsky. 2014. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research* **15**(1) 2773–2832.

[4] Assael, Henry. 1970. Segmenting markets by group purchasing behavior: an application of the aid technique. *Journal of Marketing Research* 153–158.

[5] Bell, Robert M, Yehuda Koren. 2007. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter* **9**(2) 75–79.

[6] Brovman, Yuri M., Marie Jacob, Natraj Srinivasan, Stephen Neola, Daniel Galron, Ryan Snyder, Paul Wang. 2016. Optimizing similar item recommendations in a semi-structured marketplace to maximize conversion. *Proceedings of the 10th ACM Conference on Recommender Systems*. RecSys '16, ACM, 199–202.

[7] Comaniciu, Dorin, Peter Meer. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence* **24**(5) 603–619.

[8] de Hoon, Michiel JL, Seiya Imoto, John Nolan, Satoru Miyano. 2004. Open source clustering software. *Bioinformatics* **20**(9) 1453–1454.

[9] DeSarbo, Wayne S, Ajay K Manrai, Lalita A Manrai. 1994. Latent class multidimensional scaling. a review of recent developments in the marketing and psychometric literature. *Advanced Methods of Marketing Research, R. Bagozzi (Ed.), Blackwell Pub* 190–222.

[10] Dhillon, Inderjit S, Yuqiang Guan, Brian Kulis. 2004. Kernel k-means: spectral clustering and normalized cuts. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 551–556.

[11] Filippone, Maurizio, Francesco Camastra, Francesco Masulli, Stefano Rovetta. 2008. A survey of kernel and spectral methods for clustering. *Pattern recognition* **41**(1) 176–190.

[12] Gupta, Sachin, Pradeep K Chintagunta. 1994. On using demographic variables to determine segment membership in logit mixture models. *Journal of Marketing Research* 128–136.

[13] Herlocker, Jonathan L, Joseph A Konstan, Al Borchers, John Riedl. 1999. An algorithmic framework for performing collaborative filtering. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 230–237.

[14] Hsu, Daniel, Sham M Kakade. 2013. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*. ACM, 11–20.

[15] Jain, Anil K. 2010. Data clustering: 50 years beyond k-means. *Pattern recognition letters* **31**(8) 651–666.

[16] Kamakura, Wagner A. 1988. A least squares procedure for benefit segmentation with conjoint experiments. *Journal of Marketing Research* **25** 157–67.

[17] Kamakura, Wagner A, Byung-Do Kim, Jonathan Lee. 1996. Modeling preference and structural heterogeneity in consumer choice. *Marketing Science* **15**(2) 152–172.

[18] Kamakura, Wagner A, Gary Russell. 1989. A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research* **26** 379–390.

[19] Kannan, Ravindran, Hadi Salmasian, Santosh Vempala. 2005. The spectral method for general mixture models. *Learning Theory*. Springer, 444–457.

[20] Koren, Yehuda, Robert Bell, Chris Volinsky, et al. 2009. Matrix factorization techniques for recommender systems. *Computer* **42**(8) 30–37.

[21] Fraley, Chris, Adrian E Raftery. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association* **97**(458) 611–631.

[22] Lin, Jovian, Kazunari Sugiyama, Min-Yen Kan, Tat-Seng Chua. 2013. Addressing cold-start in app recommendation: latent user models constructed from twitter followers. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 283–292.

[23] Maclachlan, Douglas L, Johny K Johansson. 1981. Market segmentation with multivariate aid. *The Journal of Marketing* 74–84.

[24] Mazumder, Rahul, Trevor Hastie, Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research* **11**(Aug) 2287–2322.

[25] McLachlan, Geoffrey, David Peel. 2004. *Finite mixture models*. John Wiley & Sons.

[26] Ng, Andrew Y, et al. 2002. On spectral clustering: Analysis and an algorithm .

[27] Ogawa, Kohsuke. 1987. An approach to simultaneous estimation and segmentation in conjoint analysis. *Marketing Science* **6**(1) 66–81.

[28] Park, Seung-Taek, Wei Chu. 2009. Pairwise preference regression for cold-start recommendation. *Proceedings of the third ACM conference on Recommender systems*. ACM, 21–28.

[29] Rokach, Lior, Oded Maimon. 2005. Clustering methods. *Data mining and knowledge discovery handbook*. Springer, 321–352.

[30] Rossi, Peter E, Greg M Allenby, Rob McCulloch. 2005. *Bayesian statistics and marketing*. John Wiley & Sons.

[31] Schein, Andrew I, Alexandrin Popescul, Lyle H Ungar, David M Pennock. 2002. Methods and metrics for cold-start recommendations. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 253–260.

[32] Sedhain, Suvash, Scott Sanner, Darius Braziunas, Lexing Xie, Jordan Christensen. 2014. Social collaborative filtering for cold-start recommendations. *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 345–348.

[33] Shani, Guy, Asela Gunawardana. 2011. Evaluating recommendation systems. *Recommender systems handbook*. Springer, 257–297.

[34] Shi, Jianbo, Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* **22**(8) 888–905.

[35] Smith, Wendell R. 1956. Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing* **21**(1) 3–8.

[36] Strehl, Alexander, Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research* **3**(Dec) 583–617.

[37] Takács, Gábor, István Pilászy, Bottyán Németh, Domonkos Tikk. 2009. Scalable collaborative filtering approaches for large recommender systems. *Journal of machine learning research* **10**(Mar) 623–656.

[38] Von Luxburg, Ulrike. 2007. A tutorial on spectral clustering. *Statistics and computing* **17**(4) 395–416.

[39] Wainwright, Martin J, Michael I Jordan, et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* **1**(1–2) 1–305.

[40] Wang, Junhui. 2010. Consistent selection of the number of clusters via crossvalidation. *Biometrika* **97**(4) 893–904.

[41] Wedel, Michel, Wayne S DeSarbo. 1994. A review of recent developments in latent class regression models. *Advanced methods of marketing research* 352–388.

[42] Wedel, Michel, Wagner A Kamakura. 2000. *Market segmentation: Conceptual and methodological foundations*, vol. 8. Springer Science & Business Media.

[43] Wedel, Michel, Cor Kistemaker. 1989. Consumer benefit segmentation using clusterwise linear regression. *International Journal of Research in Marketing* **6**(1) 45–59.

[44] Wedel, Michel, Jan-Benedict EM Steenkamp. 1989. A fuzzy clusterwise regression approach to benefit segmentation. *International Journal of Research in Marketing* **6**(4) 241–258.

[45] Wright, John, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, Shuicheng Yan. 2010. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE* **98**(6) 1031–1044.

[46] Wu, Sen, Xiaodong Feng, Wenjun Zhou. 2014. Spectral clustering of high-dimensional data exploiting sparse representation vectors. *Neurocomputing* **135** 229–239.

[47] Xu, Rui, Donald Wunsch. 2005. Survey of clustering algorithms. *IEEE Transactions on neural networks* **16**(3) 645–678.

[48] Zhang, Mi, Jie Tang, Xuchen Zhang, Xiangyang Xue. 2014. Addressing cold start in recommender systems: A semi-supervised co-training algorithm. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 73–82.

[49] Zhong, Shi, Joydeep Ghosh. 2003. A unified framework for model-based clustering. *The Journal of Machine Learning Research* **4** 1001–1037.

[50] J. R. Hershey, Z. Chen, J. Le Roux and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 31-35, doi: 10.1109/ICASSP.2016.7471631.

[51] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16). JMLR.org, 478âĂŞ487.

[52] Caron M., Bojanowski P., Joulin A., Douze M. (2018) Deep Clustering for Unsupervised Learning of Visual Features. In: Ferrari V., Hebert M., Sminchisescu C., Weiss Y. (eds) Computer Vision âĂŞ ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11218. Springer, Cham.