Using community-level data to understand child labour risk in cocoa-growing areas in Côte d'Ivoire

Research Report

May 2019

International COCOA Initiative

UCW Understanding Children's Work
An Inter-Agency Research Cooperation Project

JACOBS FOUNDATION
Our Promise to Youth

ACKNOWLEDGEMENTS

This research study was prepared by Lorenzo Guarcello and Furio Rosati of the Understanding Children's Work (UCW) Programme, a joint research initiative by the International Labour Organization (ILO), the United Nations Children's Fund (UNICEF), and the World Bank.

Data collection in Côte d'Ivoire was conducted by the *Institut National de la Statistique du Côte d'Ivoire (INS)* and funded by the Jacobs Foundation.

This is one of two studies commissioned by the International Cocoa Initiative to develop country-specific child labour risk models for Côte d'Ivoire and Ghana, based on community-level data.
An executive summary of the findings of both studies is available on the ICI Knowledge Hub.

CONTENTS

EXECUTIVE SUMMARY 4

1 INTRODUCTION 6

2 DATA AND DESCRIPTIVE STATISTICS........ 7

2.1 Child employment and child labour in the study communities........8

3 ECONOMETRIC APPROACH 14

3.1 A Risk class approach........14

3.2 Empirical strategy........16

3.3 Variable selection 18

3.4 Community characteristics........21

4 IDENTIFYING RISK CLASSES 22

4.1 Child employment........ 22

4.2 Child employment: children aged 5-13 26

4.3 Child employment: children aged 14-17 27

4.4 Cross validation, child employment: randomly excluded communities........ 28

4.5 Child labour........30

4.6 Child labour: children aged 14-17 34

4.7 Cross validation, child labour: randomly excluded communities 35

5 RESULTS BASED ON INDIVIDUAL AND HOUSEHOLD LEVEL DATA........37

6 CONCLUSION........43

EXTENDED REFERENCES AND BIBLIOGRAPHY........45

APPENDIX........47

1. Additional cross validation results........ 47

2. Questions used to define children's employment and child labour 49

3. Finite mixture model specification........51

EXECUTIVE SUMMARY

Monitoring the child labour situation at community level and identifying the communities more exposed to the risk of child labour is a complex task. On the one hand disaggregated and timely information at community level are essential for effective targeting of interventions. On the other hand, obtaining reliable estimates of child labour at the community level would require one to sample a relatively large sample of individuals in each community. This would be complex in terms of sample design, costly and time consuming. Information on potentially relevant community characteristics, however, can be collected relatively easily, quickly and utilised to build a child labour "risk" indicator that can be applied and updated with relative ease.

We propose a "risk" indicator for the presence of child labour at community level in Côte d'Ivoire. We have used a concomitant variable mixture model that allows inferring information about the variable of interest, child labour, in cases in which data on this variable is not available. The proposed model, therefore, uses easily collected characteristics at community level to predict child labour at community level. In particular, the model allows one to not only to "predict" the risk of child labour in each community, but also to classify communities according to different classes of risk of child labour. This is an important advantage of the model because it allows identifying the different class of risk on the basis of a data driven process, without making an ad hoc assumption.

Using data from the Enquête de Base sur le Travail des Enfants en Côte d'Ivoire pour Développer le Cadre pour les Communautés Cacaoyères Protectrices (CCCP-2017) and data from the Protective Cocoa Community Framework (PCCF 2017), the model identifies three classes of child labour risk: about 79% of communities are in the low risk class[1], 12% of communities are in the medium risk class and 9% are in the high risk class. We find that the most statistically significant community characteristics influencing child labour risk classification are: women's education, availability of adult casual work, household involvement in cocoa production, access to basic services (such as scholarships available for secondary school) and availability of infrastructure (such as access to electricity, connection to mobile network and presence of a primary school).

Finally, in order to test the predictive power of the model, we randomly exclude some communities for which the individual child labour variable is supposed to not be observed (but community

[1] It is important to note that the term "low risk" has been used in this report for simplicity, however this is only low in comparison to the medium and high risk classes in relative terms, as it is still high overall.

characteristics are supposed to be observed). The excluded communities are classified into one of the three classes, identified using the full sample and the complete data on the bases of the parameters estimated in the full sample. Performing several tests, the model is able to correctly classify the excluded communities for most of the cases.

1 INTRODUCTION

Monitoring the child labour situation at community level and identifying the communities more exposed to the risk of child labour is a complex task. On the one hand disaggregated and timely information at community level are essential for effective targeting of intervention. On the other hand, obtaining reliable estimates of child labour at community level would require one to sample a relatively large sample of individuals in each community. This would be complex in terms of sample design, costly and time consuming. Information on potentially relevant community characteristics, however, can be collected relatively easily and quickly and utilized to build a child labour “risk” indicator that can be applied and updated with relative ease.

Moreover, as we shall discuss in more detail in the text, it is often more relevant to classify the communities in different risk groups rather than try to “predict” the specific incidence rate for each community.

In the former case, the issue becomes one of multiclass classification. Several techniques are available to this aim and the problem is closely linked to that of statistical learning. “Training” data are used to identify criteria for class membership, criteria that are subsequently used for the classification of “incomplete” data.

In the case considered here, the problem lies in the development of a methodology allowing the inference of child labour risk at the community level in Côte d’Ivoire from a set of information collected at community level, *without* the direct observation of involvement of children in child labour. We use as “training” data the information directly collected on children’s activities through an ad-hoc survey and link this to a set of indicators observed at the community level. In this way we are able to build a risk indicator that can then be applied to “incomplete” data, namely to data without the direct observation of involvement of children in child labour.

Different techniques are available to support the identification of class membership. Most of these techniques do not establish a priori the number of classes, but rather let the data identify them on the basis of the maximization of some objective function.

In this work we employ the so called Latent Class or Finite Mixture model that considers observations of the variable of interest as belonging to different “classes” whose number is to be identified by the data.

The rest of the paper is organized as follows. Section 2 illustrates the data used and presents the descriptive statistics. Section 3 discusses the risk class approach and introduces the econometric methodology. The results, including the cross validation test, are discussed in section 4.

2 DATA AND DESCRIPTIVE STATISTICS

The present study makes use of two primary data sources : (i) the Enquête de Base sur le Travail des Enfants en Côte d'Ivoire pour Développer le Cadre pour les Communautés Cacaoyères Protectrices, 2017 (CCCP-2017) ; and (ii) the Protective Cocoa Community Framework 2017 (PCCF 2017).

The 2017 Enquête de Base sur le Travail des Enfants en Côte d'Ivoire was carried out in February 2017 and was specifically designed to measure incidence and the characteristics of child labour in the cocoa-growing areas of Côte d'Ivoire. It collected information at the individual and household levels.

The data collection process included: (i) the development of a sampling strategy; (ii) the development of the data collection instruments (questionnaires); and (iii) and the data collection phase, conducted by the Institut National de la Statistique du Côte d'Ivoire (INS).

The sample was drawn following a two-stage stratified sample design approach. A sampling frame of 4,702 cocoa-growing localities covering the entire country was constructed on the basis of the list of localities provided by the Institut National de la Statistique du Côte d'Ivoire (INS) and supplemented with information on cocoa production obtained from ANADER, the national agency supporting rural development (Agence Nationale d'Appui au Développement Rural). ANADER identified 4 groups of cocoa production areas on the basis of three criteria: average yield, average acres per household and percentage of cocoa producing households. The groups are categorized as: no production, low level of production, mid-level of production and high-level of production.

The sampling frame excluded localities with no cocoa production and covered the 9 districts where cocoa is cultivated: Bas-Sassandra, Comoe, Gôh-Djiboua, Lacs, Lagunes, Montagnes, Sassandra-Marahoue, Yamoussoukro and Zanzan.[2]

Prior to the first-stage sampling, the localities were stratified by district and within districts by level of cocoa production. Then, in the first-stage, a sample of localities was extracted with unequal

[2] Côte d'Ivoire is divided in 12 ordinary district and 2 autonomous districts.

probabilities depending on the level of cocoa production in the locality, leading to a selection of 130 localities.

In the second-stage, a fixed number of households were selected within each sampled locality by systematic sampling with equal probabilities, leading to a final sample of 5,200 households, belonging to 130 communities (localities).

A survey based on two separate questionnaires was administered to the selected households. The household questionnaire, aimed at collecting background information on the household economic activities (agricultural and non-agricultural), occurrence of shocks and dwelling conditions. The individual level questionnaire aimed at collecting information on education, employment, unemployment and decisional power within the household.

At the same time, the ICI Protective Cocoa Community Framework Questionnaire (PCCF) was fielded. The PCCF is a community assessment tool, incorporating key indicators and proxies related to community development, community empowerment, education, child protection, gender and livelihoods in cocoa-growing communities in Ghana and Côte d'Ivoire. The PCCF survey in Côte d'Ivoire was conducted in 2017 in the same communities selected for the PCCF survey. It was administered to key actors in the community depending on the specific sections of the questionnaire: Community leaders, Community Child Protection Committee members, Community women's group, Children, Farmers and other organizations, School teachers and directors.

The data from the household survey, averaged at community level, were matched with the PCCF data to identify the community characteristics most relevant to the construction of the risk indicator.

On the basis of the individual data, we computed the average incidence of children's employment and children's labour at the community level. The definitions of child employment and child labour are provided in the next section, where we present also the community level characteristics used for the empirical analysis.

2.1 Child employment and child labour in the study communities

In this section we present the main characteristics of child employment and child labour in the 130 communities on the basis of the individual questionnaire. A child is considered to be in employment if during the week prior to the survey he/she has worked for at least one hour in any economic activity for pay or without pay, for profit, or in a family business.[3] A child is also considered to be in employment if he/she was not working during the week prior to the survey but had a job to go back to. We also present children's involvement in child labour. Following the national legislation, children are classified in child labour on the basis of the following criteria: children aged 5-13 years in employment and children aged 14-17 years working in hazardous occupations or working more than 40 hours per week or working at night. Hazardous occupations in turn include: children working in designated hazardous industries, namely mining and quarrying and construction; children involved in hazardous occupations, as detailed in national legislation (i.e., *Arrêté n°009MEMEASS/CAB du 19 janvier 2012,* which determines the list of hazardous work prohibited to children under 18 years); and exposure to dangerous factors.[4, 5]

2.1.1. Child employment

The main characteristics of child employment, disaggregated by gender and by age groups, are reported in Table 1.

As Table 1 shows, 22.4% of children aged 5-17 are in employment. While there are no differences in employment rates between urban and rural areas, children's involvement in employment is markedly higher among older children aged 14-17, about 40%, compared to younger children aged 5-13, about 17%. Child employment is differentiated also according to gender: 25% of boys, aged 5-17, are in employment as compared to 18% of girls in the same age range. The gender gap increases with age and is larger in urban areas[6].

As reported in Table 2, the majority of employed children are involved in non-wage activities (87%) The share of employed children who are employees (2%) or apprentices (3%) is very small.

[3] Economic activity covers all market production and certain types of non-market production (principally the production of goods and services for own use). It includes forms of work in both the formal and informal economies; inside and outside family settings; work for pay or profit (in cash or in kind, part-time or full-time), or as a domestic worker outside the child's own household for an employer (with or without pay).

[4] Exposure to dangerous factors includes: exposure to dust, fumes, gas (oxygen, ammonia), noisy environment, extreme temperatures or humidity, sharp/dangerous tools, work underground, work at heights, insufficient lighting, chemicals, paint, carry heavy loads, fire, explosive substances, operating cranes, machinery, insufficient ventilation.

[5] Additional detail on the survey questions used to define child employment and child labour are reported in Appendix 2.

[6] Each community includes both urban and rural areas.

Table 1. **Children in employment by region, age and sex**

	Residence		Age group		
	Urban	Rural	5-13	14-17	5-17
All	118,972	685,363	519,571	284,764	804,335
Boys	78,825	418,179	310,029	186,976	497,005
Girls	40,147	267,183	209,543	97,787	307,329
		%			
All	22.2	21.5	17.3	39.7	22.4
Boys	27.5	24.0	19.1	46.2	24.5
Girls	16.1	18.6	15.2	31.4	18.2

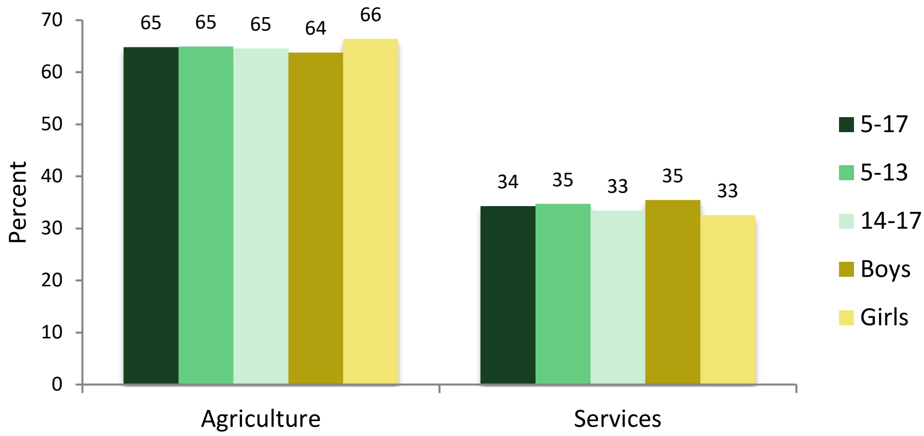
Table 2. **Child employment status, by age and sex**

	Employee	Self-employed	Contributing family worker	Apprentice	Other
5-17	2.5	21.8	65.0	2.6	8.2
5-13	1.1	20.1	67.4	1.7	9.7
14-17	5.0	24.8	60.7	4.2	5.3
Boys	3.3	21.0	65.4	2.4	8.0
Girls	1.1	23.0	64.5	2.9	8.5

Figure 1 reports the distribution of working children by sector of employment[7], age and sex. About two-thirds (65%) of children aged 5-17 are employed in the agricultural sector and 34% are in the service sector. There are no significant differences in terms of sector of employment between boys and girls or between younger and older children.

[7] The number of children working in the industry sector is negligible. Therefore, the percentages in this sector are not reported.

Figure 1. **Sector of employment by age and sex**



As far as the time intensity of work is concerned, the average number of weekly hours worked is lower in the agricultural sector (23 hours) compared to the service sector (29 hours). Older children work about 10 hours per week longer than younger children and boys put in about two more hours on average each week than their female peers (Table 3).

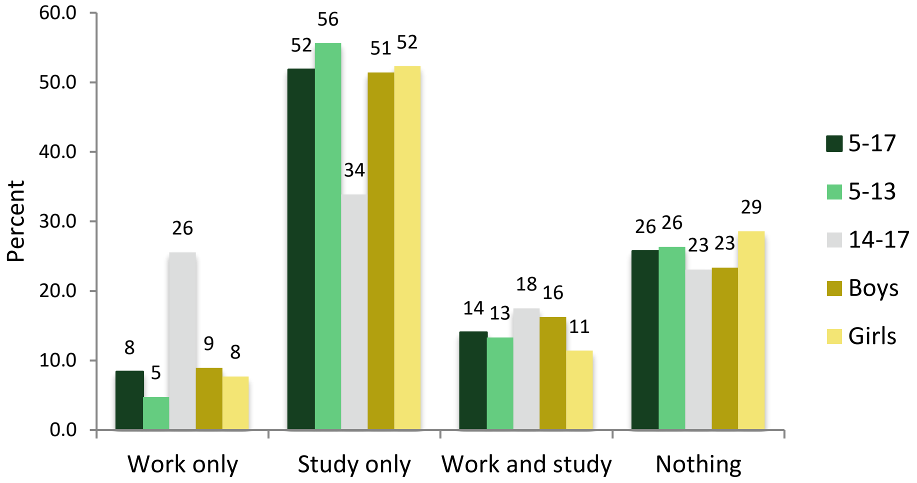
Table 3. **Hours worked by age, sex and sector of employment**

	Any sector	Agriculture		Services
5-17	25.2	22.8	N.A.	28.7
5-13	21.7	19.6	N.A.	24.9
14-17	31.6	28.8	N.A.	35.8
Boys	25.9	23.5	N.A.	29.5
Girls	23.9	21.9	N.A.	27.4

Children's distribution across four mutually exclusive activity categories (i.e., work only, study only, work and study, nothing) is reported in Figure 2. As shown, the distribution across the four categories differs considerably by age group: 26% of children aged 14-17 work without attending school, compared to 5% of children aged 5-13. On the other hand, younger children are more likely to attend school without working (56%) than their older counterparts (34%). About 13% of the younger children and 18% of older children combine school and work. The share of children neither attending school nor working is quite high for both age ranges and for both boys and girls[8]

[8] For a discussion on children neither working nor studying see Biggeri M., Rosati F., Lyos S. Guarcello, L. (2003). "The puzzle of 'idle' children: neither in school nor performing economic activity: evidence from six countries" UCW working paper series, at www.ucw-project.org.

Figure 2. **Type of activity by age and sex**



2.1.2. Child labour

Similar results are obtained if we consider child labour, as defined above. As shown in Table 4, incidence rates for child labour are very similar to those for child employment. In fact, given the definition of child labour, the only difference is observed for children aged 14 to 17, but even in this case the difference is relatively small and the characteristics of child labour are not very different from the ones described for child employment. Therefore, in order to reduce measurement errors, we use child employment in the analysis. All results are also replicated for child labour and, as we shall see, the results are substantially unchanged.

Since child employment and child labour coincide for children aged 5-13 years by definition, in this section we focus the discussion only on child labour for children aged 14-17 years.

As shown in Table 4, boys (41%) are more likely to be involved in child labour than girls (26%).Table 5 shows that the majority of children in child labour are involved in non-wage activities (85%). The share of children who are employees (5%) or apprentices (5%) is relatively small.

Table 4. **Children in child labour by region, age and sex**

	Residence		Age group		
	Urban	Rural	5-13	14-17	5-17
All	117,184	649,627	519,571	247,239	766,811
Boys	77,465	398,546	310,029	165,982	476,011
Girls	39,719	251,081	209,543	81,258	290,800
		%			
All	21.9	20.4	17.3	34.6	20.6
Boys	27.1	22.8	19.1	41.1	23.4
Girls	15.9	17.5	15.2	26.2	17.2

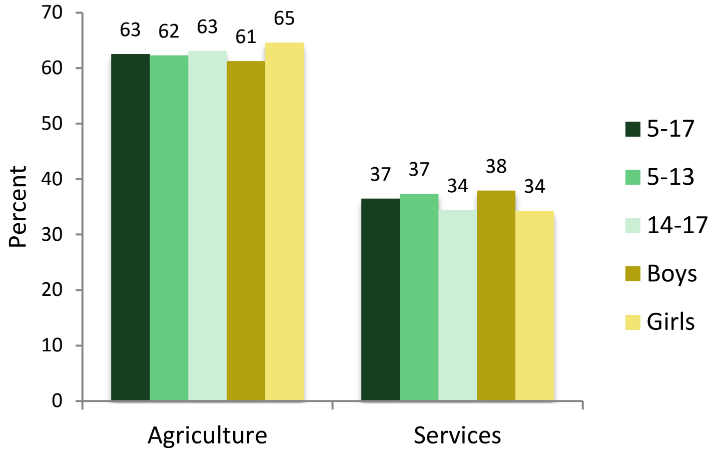
Table 5. **Child labour, by status in employment, age and sex**

	Employee	Self-Employed	Contributing family worker	Apprentice	Other
5-17	2.45	21.78	64.84	2.66	8.28
5-13	1.10	20.10	67.39	1.67	9.74
14-17	5.29	25.30	59.48	4.73	5.20
Boys	3.21	21.14	65.33	2.44	7.89
Girls	1.21	22.83	64.03	3.01	8.92

Figure 3 reports the distribution of child labourers by sector of employment,[9] age and sex. About 63% of children aged 14-17 involved in child labour are found in the agricultural sector and 34% in the service sector. There are no significant differences in terms of sector of employment between boys and girls or between younger and older children.

[9] The number of children working in the industry sector is negligible. Therefore, the percentages in this sector are not reported.

Figure 3. **Sector of child labour by age and sex**



Looking at the time intensity of work (Table 6), the average number of weekly hours worked by children aged 14-17 is lower in the agricultural sector (24 hours) compared to the service sector (29 hours).

Table 6. **Child labour. Average weekly working hours, by age, sex and sector of child labour**

	Any sector	Agriculture	Industry	Services
5-17	25.44	22.96	N.A.	29.38
5-13	21.65	19.612	N.A	24.89
14-17	33.47	30.02	N.A	39.19
Boys	26.26	23.57	N.A	30.17
Girls	24.15	22.02	N.A	27.94

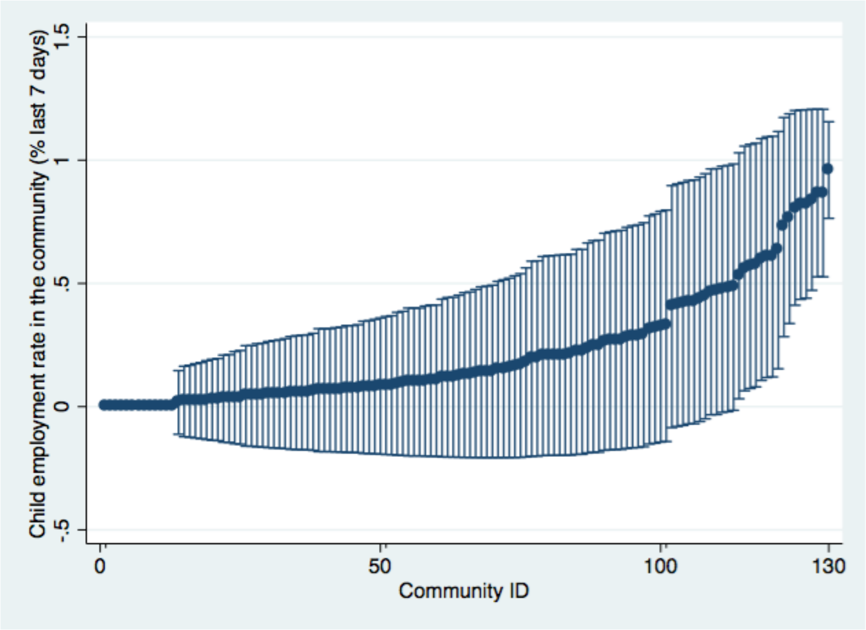
3 ECONOMETRIC APPROACH

3.1 A Risk class approach

Figure 4, which plots the average child employment rate by community together with the standard errors, and highlights two important characteristics of the incidence of child employment across communities. First, the range of variation is very large; there are communities where child employment is practically absent and others where most children work. This is a clear indication that communities face very different risks of child labour ranging from very low to very high. Second, there are nonetheless many communities that have very similar level of child employment incidence that are statistically indistinguishable from each other.

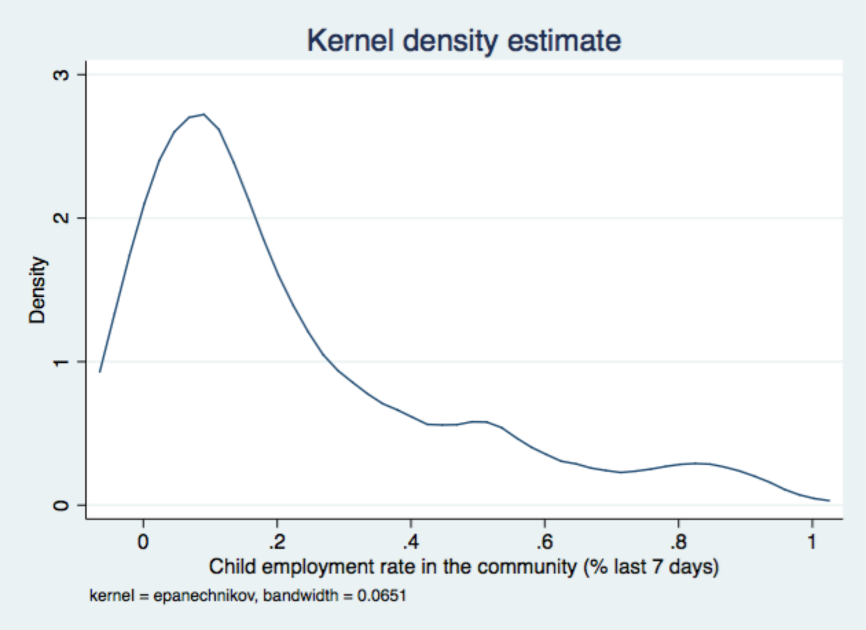
Owing to these characteristics, and also on the basis of some preliminary testing, we decided to identify different classes of child labour risk for the communities rather than to predict the expected incidence rate by single community. In other words, our approach seeks to identify the different classes of risk to which the communities belong and develop an econometric model able to predict the class membership (e.g. high risk, medium risk, low risk) of each community, without any a-priori assumptions on cut off points among classes (see above).

Figure 4. **Mean and SD of child employment in the study communities**



The existence of different classes of risk is also supported by looking at the density of the child employment incidence, as approximated by the kernel density presented in Figure 5. It is easy to see that the density has three peaks, indicating the presence of substantial heterogeneity in the distribution of child employment incidence and supporting the idea that different "risk" groups can be identified.

Figure 5. **Figure 4: Kernel density of child employment**



3.2 Empirical strategy

As discussed in the previous section, we have some initial evidence of the existence of heterogeneous groups within the communities in terms of child employment incidence. It is not possible a priori to establish the number of the different groups nor their boundaries. Rather than follow ad hoc criteria identifying the different risk groups in an arbitrary way, we follow a data driven approach. In particular, we use a flexible semi-parametric approach based on the so called finite mixture model (McLachlan and Peel 2000).

The finite mixture model assumes that data are heterogeneous and belong to a finite set of different groups. The number and characteristics of the groups is not assumed a priori, but is determined on the basis of the available information. The estimation procedure identifies, on the basis of some goodness of fit criteria, the number of classes and the probability of belonging to the different classes for each observation. In this way, we are able to predict class membership (risk class) for each community.

In what follows, we briefly present a heuristic outline of the model we use. Appendix 3 includes a more detailed description as well as the details on the estimation procedure.

Consider a random variable y_i observed on a random sample of subjects $i = 1, \dots, n$, (child labour incidence by community in our case). A finite mixture model for y_i assumes that its mass distribution function $f(y_i)$ is defined by a finite mixture of conditional distributions $f(y_i|u_i)$, where u_i is a categorical latent variable taking values $k = 1, \dots, K,$ with prior probabilities $\pi_k = P(u_i = k),$ where $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. We can think of the u_i as different "risk" class to whom the communities belong.

The distribution function can hence be written as:

$$f(y_i) = \sum_{k=1}^{K} \pi_k f(y_i|u_i = k) \quad (1)$$

A common interpretation of the latent variable u_i is in terms of latent classes, namely the population is assumed to be partitioned into K latent classes, where $u_i = k$ for subject i belonging to the k-th latent class. Thus, the prior probability π_k corresponds to the proportion of subjects in the k-th latent class (class size).

In principle both the probability of y_i and of belonging to a latent class k can be conditional on a set of covariates. In the present case we use a version of the so called Concomitant Variable Latent Class model (Dayton and MacReady 1988, Wedel 2002) and consider the probabilities of class membership as conditional on a set of community level covariates.

In particular, we assume that probabilities of belonging to a given class in the finite mixture vary across communities according to a vector of covariates $z_{i.}$

$$f(y_i|z_i) = \sum_{k=1}^{K} \pi_{k|z_i} f(y_i|u_i = k) \quad (2)$$

where $\pi_{k|z_i} = \Pr(u_i = k|z_i)$, with $\pi_{k|z_i} > 0$ and $\sum_{k=1}^{K} \pi_{k|z_i} = 1$ for each subject i. The probabilities of belonging to the k-th class are conditional on the covariate vector and are estimated using a multinomial logit model:

$$\pi_{k|z_i} = \frac{\exp(\beta_{0k} + z_i' \beta_{1k})}{\sum_h \exp(\beta_{0h} + z_i' \beta_{1h})} \quad (3)$$

Once the prior probabilities are derived, $\hat{\pi}_{k|z_i}$, we predict the child employment incidence for each community by plugging the estimated parameter in the following equation:

$$\hat{y}_i = \sum_{k=1}^{K} \hat{\pi}_{k|z_i} f(y_i|u_i = k) \quad (4)$$

Finally, in our model we assume that the distribution of child employment rate, $f(y_i)$, is a log normal distribution.

3.3 Variable selection

The PCCF survey contains information on a very large set of variables. The first step, therefore, in order to estimate the model, outlined in the previous section, is to identify a subset of relevant variables. In doing that, we need to take into consideration the fact that the sample size of the household survey was selected in order to allow for a set of around 10 explanatory variables. We have considered all the sections of the PCCF and divided the available variables in four categories: infrastructure, farming, education and women empowerment and child protection. Given the large number of variables, it is difficult to choose a priori which one should be included in order to predict the risk of child employment. Therefore, we use the stepwise regression model to gather information on which variable is more informative with respect to the child employment rate in each community. The stepwise procedure was carried out separately for each of the four categories described above.

Table 7 present the results of the procedure indicating the variables considered and the one that resulted significant (detailed results are available upon request).

Table 7. **Variable selection: stepwise regression analysis**

Variable	Description	Significant	Variable	Description	Significant
Infrastructure					
Kindergarten	Dummy variable, kindergarten is in the community		Kindergarten No.	Number of kindergartens in the community	
Primary school	Dummy variable, primary school is in the community	Yes	Primary school No.	Number of primary schools in the community	
Health centre	Dummy variable, primary health centre is in the community		Health centre No.	Number of health centre in the community	
Electricity	Dummy variable, connection to electricity network in the community	Yes	Kindergarten distance	Distance from kindergarten	
Mobile	Dummy variable, connection to mobile network in the community	Yes	Primary school distance	Distance from primary school	
Internet	Dummy variable, connection to internet network in the community		Health centre distance	Distance from health centre	

Table 7. **Variable selection: stepwise regression analysis (cont'd)**

Variable	**Description**	**Significant**	**Variable**	**Description**	**Significant**
Infrastructure					
Road	Dummy variable, community reachable by road		Junior secondary distance	Distance from junior secondary	
Road surface	Dummy variable, community road surface		Vocational distance	Distance from vocational school	
Road accessible	Dummy variable, road accessible all year		Birth certificate	Percentage of children with birth certificate (0-40, 41-69, 70-100)	
Farming					
Buying company	Dummy variable, licence buying company in the community		Cocoa land size	Cocoa farm size per farmer in the community (acres)	
Cocoa organization	Dummy variable, cocoa farmer organization in the community		Cocoa farmers	Number of cocoa farmers in the community	
Extension services	Dummy variable, extension services in the community		Cocoa production	Cocoa production per year in the community (ton)	
Input available	Dummy variable, farming inputs available in the community		Farmers trained by Ext. Serv.	Number of farmers trained by ext. services in the community	
Casual work available	Dummy variable, adult casual work available in the community	Yes	Share of households cultivating cocoa	Percentage of households cultivating cocoa in the community	
Input affordable	Dummy variable, farming inputs affordable in the community				
Agr. Services	Dummy variable, agricultural services in the community				

Table 7. **Variable selection: stepwise regression analysis (cont'd)**

Variable	Description	Significant	Variable	Description	Significant
Education					
Toilet facilities in primary	Dummy variable, toilet facilities in primary school		Enrolment rate	Percentage of children 5-17 enrolled in school	
Scholarship in secondary	Dummy variable, scholarship in secondary school	Yes	Children enrolled in kindergarten	Number of children enrolled in kindergarten	
School Feeding program in primary school	Dummy variable, feeding programme in primary school		Children enrolled in primary	Number of children enrolled in primary school	
			Children enrolled in junior secondary	Number of children enrolled in junior secondary school	
			Children enrolled in senior secondary	Number of children enrolled in senior secondary school	
Women empowerment and child protection					
Community Action Plan	Dummy variable, Community Action Plan in the community		Female lead farmers	Number of female lead farmers in the community	
Community Child Protection Committee	Dummy variable, Community Child Protection Committee in the community		Female leadership positions	Number of leadership positions occupied by females in the community	
Regulations to protect children	Dummy variable, regulations to protect children in the community		Women education	Main education level reached by women in the community	Yes
Remediation services	Dummy variable, remediation services for children in the community				

A limited number of variables are statistically significant in explaining rates of children's involvement in employment. According to the results of the stepwise regression, the following variables are included in the estimation of model: presence of primary school, access to electricity and mobile network (from the infrastructure section); availability of casual work (from the farming section); scholarships for secondary education (from the education section); women's education (from women empowerment and child protection section). Even if not identified as significant in the stepwise regression, we include some additional variables that are potentially of interest in predicting child employment, as suggested by ICI, i.e. presence of kindergarten (from the infrastructure section), the average school attendance at community level of children aged 5-17 (from the education section), inputs available, share of households cultivating cocoa, cocoa organization and cocoa production (from the farming section). The inclusion of these additional variables does not invalidate the estimation procedure.

3.4 Community characteristics

Table 8 shows the descriptive statistics for the community characteristics included in the estimate of the empirical model. Access to education is proxied by three variables: two dummy variables indicating, respectively, whether there is a primary school in the community and whether children at secondary school receive any scholarship, as well as the attendance rate at community level. As reported in the table, 88% of the communities have a primary school but only 13% have children in the community receiving secondary school scholarships and on average 65% of children attended school at the time of the survey. We then consider whether there is a kindergarten in the community and a categorical variable representing the education level reached by the majority of the women in the community. About 12% of the communities have a kindergarten, while in the majority of communities women attended primary school (28% primary grades1-3 and 36% primary grades 4-6). Women attended secondary school in only 14% of communities, while in 22% of communities women have no education at all.

As measure of community infrastructure, two dummy variables are selected, indicating whether the community has access to the electricity network and to the mobile network. About 52% of communities have access to electricity and 63% have access to mobile services. The farming background is characterized by several variables: availability of adult casual work (68%), availability of farming inputs (82%) and presence of cocoa organizations (28%). Finally, we consider the share of households involved in the production of cocoa obtained from the household questionnaire and aggregated at the community level. On average, 60% of households are involved in cocoa production. The average cocoa production per year by communities (measured in tons) is about 300 tons.

Table 8. **Community characteristics**

Variable	Obs.	Mean	SD	Min	Max
Child employment in the community (No.)	130	6186.77	9676.45	0.00	63687.00
Children 5-17 (No.)	130	28670.32	32538.16	1720.00	241557.00
Child employment in the community (%)	130	0.22	0.23	0.00	0.96
Kindergarten	130	0.12	0.32	0.00	1.00
Primary school	130	0.88	0.33	0.00	1.00
Scholarship in Secondary school	130	0.13	0.34	0.00	1.00
Attendance rate	130	0.65	0.15	0.17	1.00
Women education: No school	130	0.22	0.41	0.00	1.00
Women education: Primary 1-3	130	0.28	0.45	0.00	1.00
Women education: Primary 4-6	130	0.36	0.48	0.00	1.00
Women education: Junior High School	130	0.12	0.33	0.00	1.00
Women education: Senior High School	130	0.02	0.15	0.00	1.00
Electricity	130	0.52	0.50	0.00	1.00
Mobile	130	0.63	0.48	0.00	1.00
Adult casual work available	130	0.68	0.47	0.00	1.00
Inputs available	130	0.82	0.39	0.00	1.00
Share of households cultivating cocoa	130	0.60	0.34	0.00	1.00
Cocoa organization	130	0.28	0.45	0.00	1.00
Cocoa production	130	299.82	414.89	3.00	3084.00

4 IDENTIFYING RISK CLASSES

4.1 Child employment

As detailed in Appendix 3, the model in equation (2) is estimated through the Expectation-Maximization (EM) algorithm with a fixed number of latent classes. The selection of the number of latent classes (u_i) is guided by a goodness of fit criteria. In particular, we employ the so-called AIC criterion and we present the results in Table 9.

Table 9. **Selection of number of latent classes**

Number of Latent Classes	**Log-likelihood**	**AIC**	**BIC**
1	42.71	-81.42	-75.69
2	80.10	-128.20	-82.32
3	95.08	130.16	-44.13
4	Not converged	-	-

AIC: Akaike information criterion; BIC: Bayesian information criterion

Following the AIC criterion, we choose the concomitant mixture model with three latent classes and we present the estimates for the model (2) assuming three latent classes. The three latent classes have the following location points: $\hat{u}_1 - 1.93$; $\hat{u}_2 = -0.58$; $\hat{u}_3 = 0.06$. The location points represent deviation of the average percentage of working children from the overall intercept in class one, two and three, respectively. The values of the estimated location points indicate that the first class includes communities with a relatively low level of child employment, the second class includes communities with a medium level of child employment, and communities in the third class are characterized by a high level of child employment. Therefore, we can consider the three latent classes as three different classes of risk of child employment: low, medium and high risk respectively in class one, two and three.

Table 10 shows the distribution of communities across the three classes. About 80% of the communities belong to the "low" risk group and have an average child employment incidence rate of 17 percent. Just over 12 percent of communities belong to the medium risk class with an average incidence of child employment of about 34 percent. Finally, 10 percent of the communities belong to the high risk group where the average incidence of children's employment is over 50 percent. [10]

Table 10. **Latent Classes based on prior probabilities and average child employment rate**

Class	Obs.	%	Average Employment Rate
1-low risk	102	78.46	0.17
2-medium risk	16	12.31	0.34
3-high risk	12	9.23	0.53
Total	130	100.0	

[10] Note that the adjective "low" should be interpreted in relative terms with respect to child employment in the other communities.

Figure 6 presents graphically these results indicating the range of variation around the mean.

Figure 6. **Box plot of child employment by class of risk**

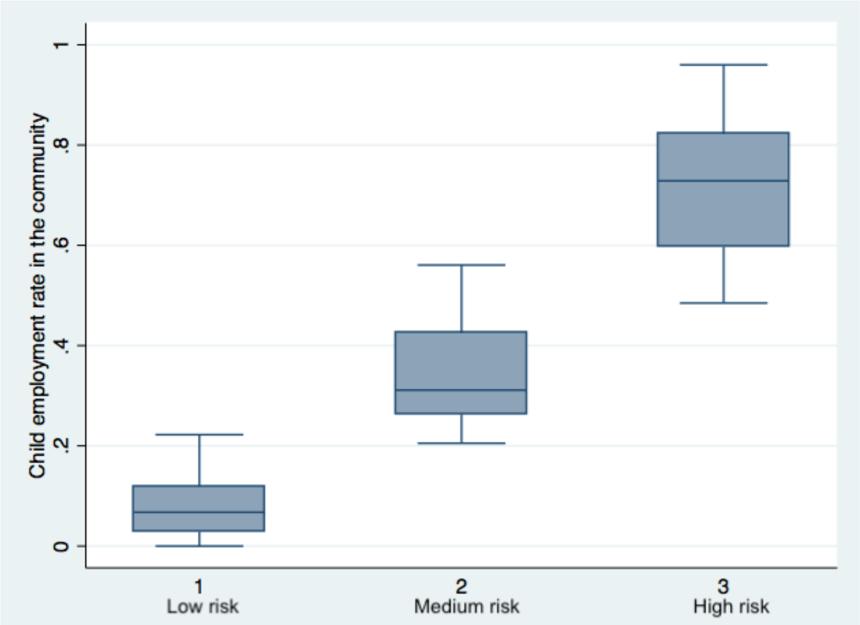


Table 11 shows estimation results of the multinomial logit model from equation (2). The coefficients in Table 11 are reported in terms of log odds ratio with respect to the base category, which in this case is the third component, i.e. the group with highest risk of child employment. The marginal effects derived from the multinomial logit model are reported in Table 12. Table 11 indicates that the presence of primary school in the community, higher women's education and connection to the mobile network are statistically significant and positively related with the probability of belonging to the low risk class relatively to the high risk class. Surprisingly, connection to the electricity network is statistically significant but negatively correlated with the probability of belonging to the low risk class with respect to the base category. Availability of occasional adult labour supply significantly increases the probability of belonging to the low and medium risk class with respect to the high risk class. Finally, the higher the number of households involved in cocoa production the lower is the probability of belonging to the low and medium risk class with respect to the high risk class.

Table 11. **Concomitant mixture model - parameter estimates**

	Classes		
VARIABLES	1	2	3
Kindergarten	-1.14	-1.69	-
	(1.257)	(1.541)	-
Primary school	2.11*	2.24	-
	(1.217)	(1.428)	-
Scholarship in Secondary school	17.07	15.69	-
	(1,896.462)	(1,896.462)	-
Attendance rate	1.00	-1.77	-
	(2.799)	(2.880)	-
Women education	1.03**	0.73	-
	(0.446)	(0.476)	-
Connected to mobile network	2.69**	1.06	-
	(1.063)	(1.170)	-
Connected to electricity network	-3.79***	-1.97	-
	(1.320)	(1.440)	-
Adult casual work available	2.71***	2.22**	-
	(0.945)	(1.014)	-
Farming inputs available	1.15	1.05	-
	(1.069)	(1.138)	-
Share of households cultivating cocoa	-3.78**	-4.02**	-
	(1.796)	(1.902)	-
Cocoa farmer organization	-1.47	-1.07	-
	(1.023)	(1.071)	-
Cocoa production	0.56	0.48	
	(0.393)	(0.428)	
Constant	-5.09*	-2.52	-
	(2.685)	(2.748)	-
Observations	130	130	130

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 12. **Concomitant mixture model - marginal effects**

	Classes		
VARIABLES	1	2	3
Kindergarten	0.02	-0.12	0.10
Primary school	0.07	0.08	-0.16
Scholarship in Secondary school	0.93	0.27	-1.20
Attendance rate	0.42	-0.42	0.01
Women education	0.09	-0.02	-0.07
Connected to mobile network	0.34	-0.19	-0.15
Connected to electricity network	-0.41	0.19	0.22
Adult casual work available	0.18	-0.00	-0.18
Farming inputs available	0.06	0.02	-0.08
Share of households cultivating cocoa	-0.13	-0.15	0.28
Cocoa farmer organization	-0.12	0.02	0.10
Cocoa production	0.04	0.00	-0.04
Observations	130	130	130

4.2 Child employment: children aged 5-13

In this section we show the results obtained by analysing involvement in employment of children aged 5-13.

Although the smallest AIC value corresponds to the model with three latent classes, we choose the model with two latent classes, since in the former model the medium risk class is composed only of one community.

Table 13. **Selection of number of latent classes - Child (5-13) Employment**

Number of Latent Classes	Log-likelihood	AIC	BIC
1	40.10	-76.19	-70.46
2	87.07	-142.16	-96.28
3	108.87	-157.74	-71.71
4	Not converged	-	-

AIC: Akaike information criterion; BIC: Bayesian information criterion

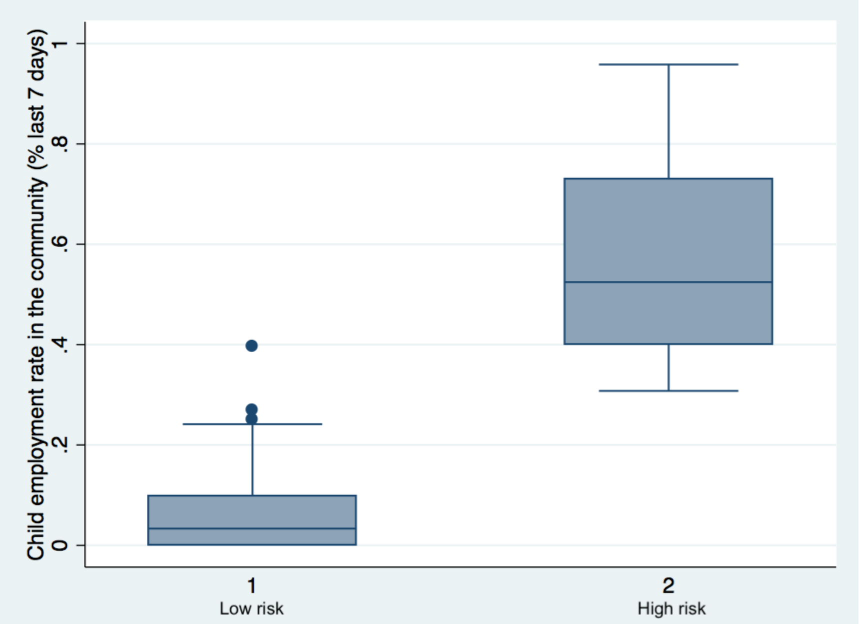
The distribution of communities over the two classes of risk based on the prior probabilities is reported in Table 14. The low risk class includes 89% of communities, with an average child employment rate of 14%, and the high risk class includes 11% of communities reporting an average child employment rate of 47%.

Table 14. **Latent Classes based on prior probabilities and average child (5-13) employment rate**

Class	Obs.	%	Average Child Employment Rate
1-low risk	116	89.23	0.14
3-high risk	14	10.77	0.47
Total	130	100.0	

The following figure shows the graphical distribution around the mean of child employment, using the posterior probabilities (Figure 7).

Figure 7. **Box plot of child (5-13) employment by class**



4.3 Child employment: children aged 14-17

Considering only children aged 14-17 leads to the classification of communities in two classes of risk, as obtained also with children aged 5-13 (shown in the previous section). In fact, the smallest AIC value is obtained with the model with two latent classes (Table 15).

Table 15. **Selection of number of latent classes - Child (14-17) Employment**

Number of Latent Classes	Log-likelihood	AIC	BIC
1	18.63	-33.27	-27.53
2	38.95	-45.90	-0.02
3	Not converged	-	-

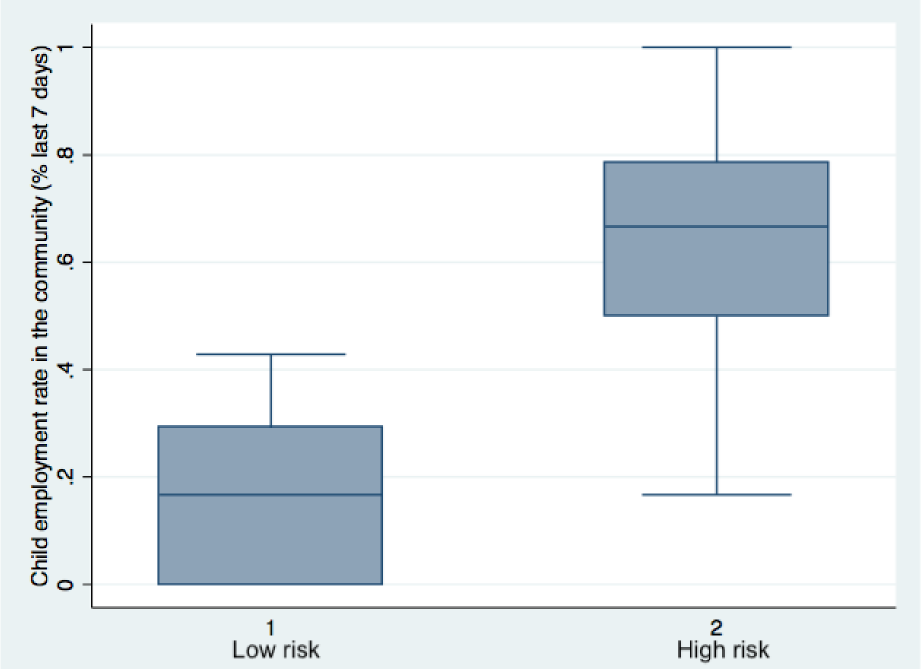
AIC: Akaike information criterion; BIC: Bayesian information criterion

However, differently from younger children, in this case the distribution of communities between the two risk classes is characterized by higher concentration in the high risk class. In fact, 59% of communities are in the high risk class, with an average child employment rate of 56%, and 41% of communities are in the low risk class, with an average child employment rate of 32% (Table 16).

Table 16. **Latent Classes based on prior probabilities and average child (14-17) employment rate**

Class	Obs.	%	Average Child Employment Rate
1-low risk	53	40.77	0.32
2-high risk	77	59.23	0.56
Total	130	100.0	

Figure 8. **Box plot of child (14-17) employment by class**



4.4 Cross validation, child employment: randomly excluded communities

Turning to the model with all children aged 5-17, we perform several tests, by randomly excluding some communities from the model estimation, to check whether the predicted prior probabilities for the excluded communities correspond to the “true” prior probabilities obtained from the full sample and shown in the previous section. Consequently, if the predicted prior probabilities correspond to the “true” prior probabilities, the predicted class membership is also equal to the “true” class membership. Moreover, we also analyse the predicted child employment incidence for each excluded communities making use of equation (4). In order to examine whether the prediction of child labour is satisfactory, we consider the relative ranking of the actuals and fitted values. Namely, we check whether communities lie in the same ranking positions both according to the observed and the predicted values through the Spearman's rho index. The results on the predicted child employment level in each community and the Spearman's rho index are shown in Appendix 1.

In Test 1 we randomly exclude 11 communities: 8 from the low risk class, 1 from the medium risk class and 2 from the high risk class. The predicted risk class membership and the predicted child employment are reported in Table 17.

Overall, the tests show a reasonably good "predictive" power of the model with respect to its ability to class communities in the different group of risks identified.

Table 17. Test 1

	Predicted class (excluded communities)			
Class (full sample)	1	2	3	Total
1	8	0	0	8
2	0	1	0	1
3	0	0	2	2
Total	8	1	2	11

In Test 2 we randomly exclude 12 communities: 10 from the low risk class, 1 from the medium risk class and 1 from the high risk class. As shown in Table 18, all the class memberships are perfectly predicted except two communities. One community is predicted in class 2 rather than in class 1, and another community is predicted in class 1 rather than in class 2.

Table 18. Test 2

	Predicted class (excluded communities)			
Class (full sample)	1	2	3	Total
1	9	1	0	10
2	1	0	0	1
3	0	0	1	1
Total	10	1	1	12

In Test 3 we randomly exclude 4 communities: 3 from the first risk class and 1 from the third risk class. For test 3, the predicted class memberships correspond to the "true" class membership, as shown in Table 19.

Table 19. Test 3

	Predicted class (excluded communities)		
Class (full sample)	1	3	Total
1	3	0	3
3	0	1	1
Total	3	1	4

In Test 4 we randomly exclude 19 communities: 14 from the first risk class 2 from the second risk class and 3 from the third risk class (Table 20). Also for test 4, the prediction of the class membership is highly satisfactory.

Table 20. Test 4

	Predicted class (excluded communities)			
Class (full sample)	1	2	3	Total
1	14	0	0	14
2	1	1	0	2
3	0	0	3	3
Total	15	1	3	19

As expected, Tables A.1-A.4 in the Appendix show that the ranking of communities according to the predicted child employment incidence does not always coincide with the ranking based on the observed child labour incidence. In fact, the values of the Spearman's rho index are quite low (in the ideal situation of equal ranking the index should be equal to 1) and the child employment incidence is predicted with some margin of errors.

4.5 Child labour

We repeated the same estimation approach, detailed above, considering child labour rate as the dependent variable. We obtain estimation results very close to the results obtained with child employment.

As shown in Table 21, the AIC criterion leads to the choice of three latent classes, identifying communities with relatively low level of child labour risk in the first, medium level in the second, and high level in the third.

Table 21. **Selection of number of latent classes**

Number of Latent Classes	Log-likelihood	AIC	BIC
1	42.76	-81.53	-75.79
2	82.23	-132.46	-86.58
3	99.13	-138.26	-52.24
4	Not converged	--	--

AIC: Akaike information criterion; BIC: Bayesian information criterion

Table 22 shows the distribution of communities across the three classes. About 87% of the communities belong to the low risk class and have an average incidence of child labour of 17 percent. About 5 percent of communities belong to the medium risk class with an average incidence of child labour of 30 percent. Finally, 8 percent of the communities belong to the high risk class where the average incidence of child labour is 54 percent.

Table 22. **Latent Classes based on prior probabilities and average child labour rate**

Class	Obs.	%	Average Child labour Rate
1-low risk	113	86.92	0.17
2-medium risk	6	4.62	0.30
3-high risk	11	8.46	0.54
Total	130	100.0	

We check to what extent the classes of risk of child employment correspond to the classes of risk of child labour. Table 23 shows that the majority of communities are classified in the same class of risk of both child employment and child labour. However, there are some differences: 11 communities that were classified as medium or high risk in terms of child employment are now classified as low risk in terms of child labour.

Table 23. **Classes of risk, child employment and child labour**

	Classes of child employment			
Classes of child labour	1	2	3	Total
1	102	9	2	113
2	0	6	0	6
3	0	1	10	11
Total	102	16	12	130

Figure 9 presents graphically these results indicating the range of variation around the mean.

Figure 9. **Box plot of child labour by class**

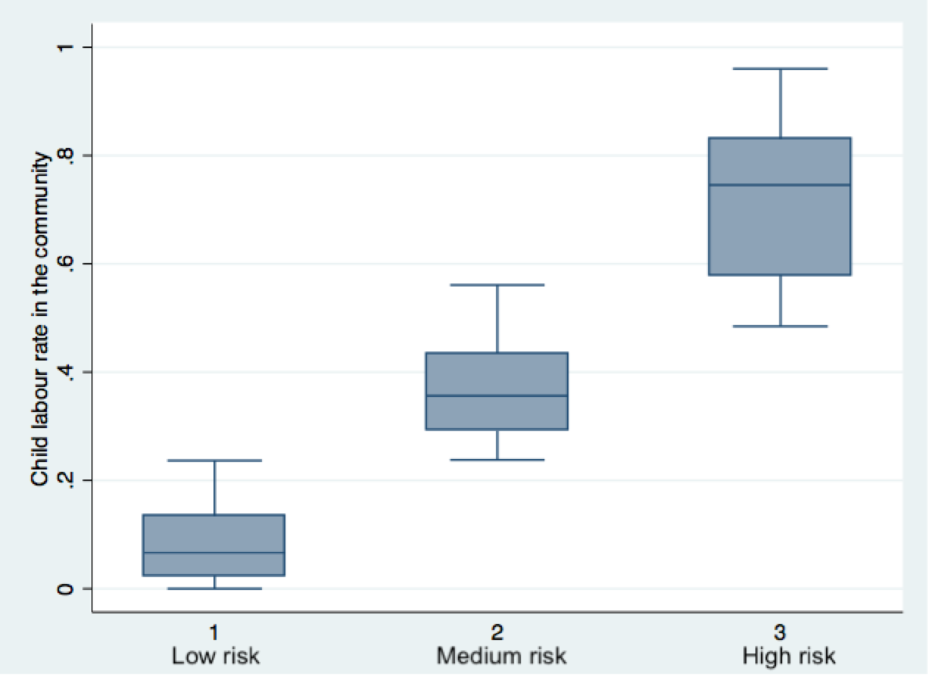


Table 24 shows estimation results of the multinomial logit model from equation (3). The coefficients in Table 24 are reported in terms of log odds ratio with respect to the base category, which in this case is the third class, i.e. the class with higher risk of child labour. Table 25 reports the corresponding marginal effects.

The results obtained using the definition of child labour are very similar to the results on child employment. Table 24 indicates that higher women education and connection to the mobile network are statistically significant and positively related with the probability of belonging to the low risk class of child labour relatively to the high risk class. The presence of primary school in the community is marginally significant and positively related with the probability of belonging to the low and medium risk class with respect to the high-risk class. Also when child labour is taken into consideration, connection to the electricity network is statistically significant but negatively correlated with the probability of belonging to the low risk class with respect to the high-risk class. Availability of occasional adult labour supply significantly increases the probability of belonging to the low and medium risk class with respect to the high risk class.

Table 24. **Concomitant mixture model - parameter estimates**

	Classes		
VARIABLES	1	2	3
Kindergarten	-0.83	-0.44	-
	(1.422)	(1.531)	-
Primary school	2.51*	3.11*	-
	(1.298)	(1.647)	-
Scholarship in Secondary school	16.50	15.35	-
	(1,510.852)	(1,510.852)	-
Attendance rate	0.55	-2.99	-
	(3.051)	(3.199)	-
Women education	1.07**	0.94*	-
	(0.487)	(0.539)	-
Connected to mobile network	3.02***	1.74	-
	(1.131)	(1.236)	-
Connected to electricity network	-4.01***	-2.62*	-
	(1.340)	(1.452)	-
Adult casual work available	3.02***	2.23**	-
	(1.026)	(1.116)	-
Farming inputs available	1.72	1.50	-
	(1.181)	(1.270)	-
Share of households cultivating cocoa	-5.23**	-5.13**	-
	(2.076)	(2.214)	-
Cocoa farmer organization	-1.72*	-0.92	-
	(1.042)	(1.127)	-
Cocoa production	0.63	0.43	
	(0.435)	(0.482)	
Constant	-4.78	-2.50	-
	(2.992)	(3.206)	-
Observations	130	130	130

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 25. **Concomitant mixture model - marginal effects**

	Classes		
VARIABLES	1	2	3
Kindergarten	-0.09	0.042	.047
Primary school	0.02	0.18	-0.20
Scholarship in Secondary school	0.83	0.30	-1.14
Attendance rate	0.52	-0.59	0.07
Women education	0.06	0.01	-0.07
Connected to mobile network	0.30	-0.13	-0.18
Connected to electricity network	-0.36	0.12	0.24
Adult casual work available	0.23	-0.04	-0.19
Farming inputs available	0.10	0.01	-0.12
Share of households cultivating cocoa	-0.23	-0.14	0.37
Cocoa farmer organization	-0.18	0.09	0.10
Cocoa production	0.05	-0.01	-0.04
Observations	130	130	130

4.6 Child labour: children aged 14-17

Also for child labour, we disaggregate child labour rate by age. To be noted that, according to the definition of child labour, child labour of children aged 5-13 corresponds to child employment of children aged 5-13, for which the results have been already shown. Therefore, we present the results obtained considering child labour of children aged 14-17.

Table 26. **Selection of number of latent classes - Child (14-17) Labour**

Number of Latent Classes	**Log-likelihood**	**AIC**	**BIC**
1	16.26	-28.52	-22.79
2	33.69	-35.39	10.49
3	Not converged	-	-

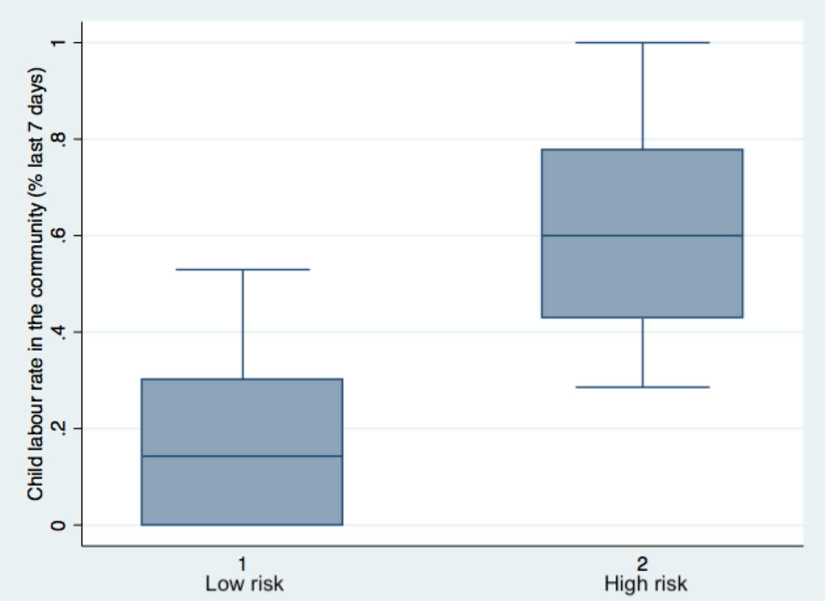
AIC: Akaike information criterion; BIC: Bayesian information criterion

Also for child labour of older children, we obtain that the best model according to the AIC criteria, is the one with two latent classes. About 56% of communities are in the low risk class, with on average 28% of children in child labour, and 44% of communities are in the high risk class, with on average 50% of children in child labour (Table 27). The graphical distribution in the box plot, based on the posterior probabilities, is shown in Figure 10.

Table 27. **Latent Classes based on prior probabilities and average child (14-17) labour rate**

Class	Obs.	%	Average Child labour Rate
1-low risk	73	56.15	0.28
2-high risk	57	43.85	0.50
Total	130	100.0	

Figure 10. **Box plot of child (14-17) labour by class**



4.7 Cross validation, child labour: randomly excluded communities

Considering all children (aged 5-17), also for classes of risk of child labour, we perform several tests, by randomly excluding some communities from the model estimation, to check whether the predicted prior probabilities for the excluded communities correspond to the “true” prior probabilities obtained from the full sample.

In Test 1 we randomly exclude 11 communities: 7 from the low risk class, 1 from the medium risk class and 3 from the high risk class. In Test 2 we randomly exclude 12 communities from the low risk class. In Test 3 we randomly exclude 3 communities from the low risk class and 1 community from the medium risk class. In Test 4, 18 communities are randomly excluded from the low risk class and 1 from the high risk class.

The predicted risk class memberships are reported in Table 28-Table 31. Overall, the tests show a reasonably good “predictive” power of the model with respect to its ability to class communities in the different group of risks identified.

Table 28. Test 1

	Predicted class (excluded communities)			
Class (full sample)	1	2	3	Total
1	6	0	1	7
2	0	0	1	1
3	0	0	3	3
Total	6	0	5	11

Table 29. Test 2

	Predicted class (excluded communities)			
Class (full sample)	1	2	3	Total
1	12	0	0	12
2	0	0	0	0
3	0	0	0	0
Total	12	0	0	12

Table 30. Test 3

	Predicted class (excluded communities)			
Class (full sample)	1	2	3	Total
1	3	0	0	3
2	0	1	0	1
3	0	0	0	0
Total	3	1	0	4

Table 31. Test 4

	Predicted class (excluded communities)			
Class (full sample)	1	2	3	Total
1	17	1	0	18
2	0	0	0	0
3	0	0	1	1
Total	17	1	1	19

5 RESULTS BASED ON INDIVIDUAL AND HOUSEHOLD LEVEL DATA

Given the availability of individual and household level data, we analyse the determinants of child employment, exploiting this information, in order to check the consistency with results obtained using community level data.

We consider the following individual level covariates: sex and age of the child and a dummy variable indicating whether the child attends a school with free supplies. At the household level, we include: sex and education of the household head, a dummy variable indicating whether the household is in an urban or rural area, a dummy variable indicating whether the household cultivates cocoa (which was also included in the community level analysis and averaged at the community level) and monthly household income. Finally, we include also covariates at the community level (from the PCCF questionnaire) used also for the classification of communities in classes of risk. In fact, we consider the number of primary schools in the community, dummy variables indicating whether the primary schools have toilet facility and feeding programme, a dummy variable indicating whether a secondary school scholarship is provided, dummy indicators for adult casual work availability and farming input availability in the community.

We first consider a multinomial logit based on four mutually exclusive categories: study only, work only, work and study, nothing. The results of the multinomial logit are shown in Table 32 in terms of marginal effects. Then we estimate a probit model for the probability to be in employment and the results are shown in Table 33 in terms of marginal effects.

Both the multinomial logit and the probit model estimates show that at the individual level gender and age are important determinants of child employment. In fact, boys and older children are more likely to be involved in employment. Moreover, if the child attends a school with free supply he is more likely to study and less likely to work. Among the household level characteristics, head education and monthly income are negatively related with child employment, while household cocoa production is positively related with child employment. Consistently with results obtained in the previous sections, the community level covariates on availability of adult casual work, primary schools and of secondary school scholarships significantly reduce the probability of child employment.

Table 32. **Determinants of child activities - individual level**

	Study only	Work only	Work and study	Nothing
Male	-0.01	0.01**	0.04*	-0.05*
	(0.01)	(0.00)	(0.01)	(0.01)
Age	0.16*	-0.04*	0.07*	-0.19*
	(0.01)	(0.00)	(0.01)	(0.01)
Age^2	-0.01*	0.00*	-0.00*	0.01*
	(0.00)	(0.00)	(0.00)	(0.00)
Urban	0.02	0	0.03**	-0.05*
	(0.01)	(0.01)	(0.01)	(0.01)
Male head	0	0	-0.02**	0.03*
	(0.02)	(0.01)	(0.01)	(0.01)
Head education	0.01*	-0.01*	-0.00**	-0.01*
	(0.00)	(0.00)	(0.00)	(0.00)
HH in cocoa	0.02	0	0.02**	-0.03**
	(0.01)	(0.01)	(0.01)	(0.01)
Monthly income (log)	0.01*	-0.01**	-0.01**	0
	(0.00)	(0.00)	(0.00)	(0.00)
Free school supplies	0.46*	-0.12*	0.15*	-0.49*
	(0.03)	(0.01)	(0.01)	(0.03)
Primary in the community (No.)	0.02**	-0.01**	-0.01**	0
	(0.01)	(0.00)	(0.01)	(0.01)
Toilet facilities at primary school	0.04**	-0.02*	-0.04*	0.02**
	(0.01)	(0.01)	(0.01)	(0.01)
Scholarship at Secondary school	0.18*	-0.09*	-0.14*	0.05*
	(0.02)	(0.01)	(0.01)	(0.01)
Feeding programme at Primary school	-0.02	0.01	0	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Adult casual work available in the community	0.10*	-0.05*	-0.09*	0.05*
	(0.01)	(0.00)	(0.01)	(0.01)
Farming inputs available in the community	0.02	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)

* 0.10 ** 0.05 * 0.001; SE in parenthesis; marginal effects from multinomial logit estimation

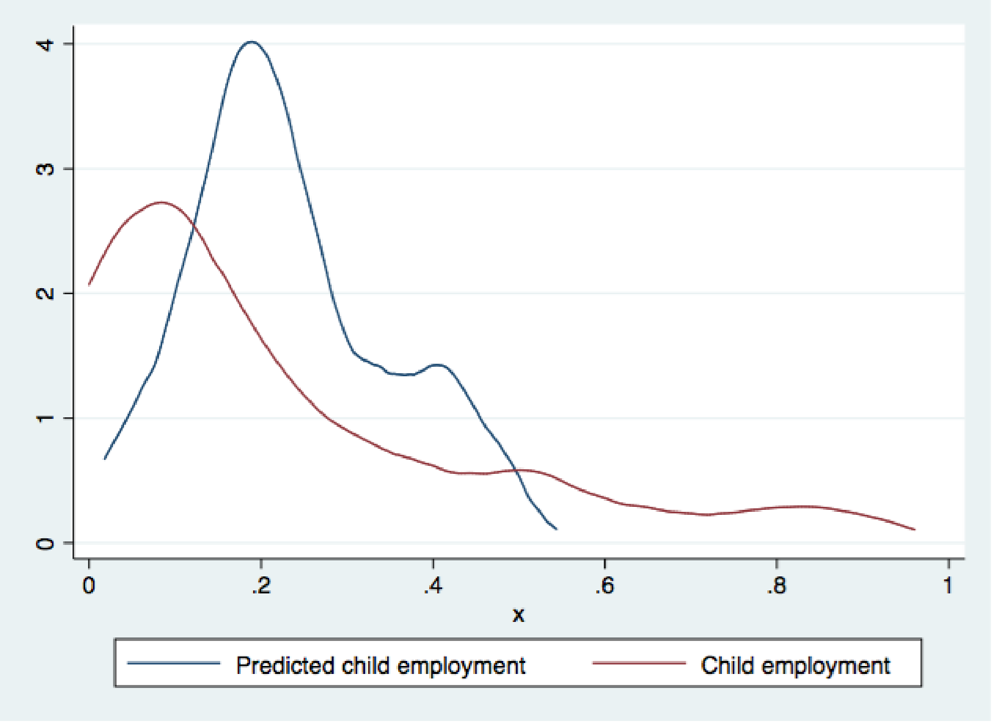
Table 33. **Determinants of child employment - individual level**

	Child employment
Male	0.06*
	(0.01)
Age	0.03**
	(0.01)
Age^2	0
	(0.00)
Urban	0.02*
	(0.01)
Male head	-0.02
	(0.01)
Head education	-0.01*
	(0.00)
HH in cocoa	0.02
	(0.01)
Monthly income (log)	-0.05**
	(0.02)
Free school supplies	0.03**
	(0.01)
Primary in the community (No.)	-0.02**
	(0.01)
Toilet facilities at primary school	-0.06*
	(0.01)
Scholarship at Secondary school	-0.22*
	(0.02)
Feeding programme at Primary school	0.01
	(0.01)
Adult casual work available in the community	-0.14*
	(0.01)
Farming inputs available in the community	-0.02
	(0.01)

* 0.10 ** 0.05 * 0.001; SE in parenthesis; marginal effects from probit estimation.

The following figure shows the predicted child employment obtained with the probit model (at the individual level) and averaged at the community level. We compare the predicted values with the observed child employment at the community level looking at the kernel densities. As can be noticed, the probit model does not allow taking into account the heterogeneity across communities, as the two peaks on the right of the distribution of the observed child employment are not reflected in the distribution of the predicted child employment.

Figure 11. **Predicted child employment from individual level probit estimation**



The probit model is estimated also using a child labour as dependent variables and the same set of covariates as independent variables. The results, reported in Table 34 are in line with the results obtained considering child employment. In fact, we observe that the same sub-set of covariates are statistically significant with a similar magnitude, with the additional statistically significant dummy variable indicating whether the household cultivates cocoa.

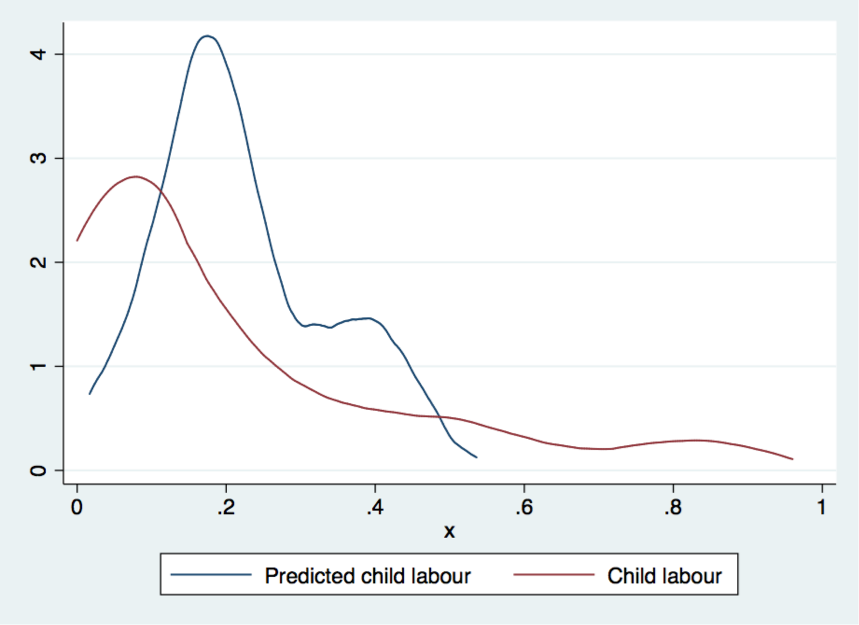
Also, as shown in Figure 12, the comparison between the predicted value of child labour and the observed child labour, averaged at community level, leads to a similar picture with respect to what observed in the case of child employment (see **Error! Reference source not found.**).

Table 34. **Determinants of child labour - individual level**

	Child labour
Male	0.06*
	(0.01)
Age	0.04*
	(0.01)
Age^2	0.00
	(0.00)
Urban	0.03**
	(0.01)
Male head	-0.02
	(0.01)
Head education	-0.01*
	(0.00)
HH in cocoa	0.03**
	(0.01)
Monthly income (log)	-0.01**
	(0.00)
Free school supplies	0.03**
	(0.01)
Primary in the community (No.)	-0.02**
	(0.01)
Toilet facilities at primary school	-0.06*
	(0.01)
Scholarship at Secondary school	-0.22*
	(0.02)
Feeding programme at Primary school	0.01
	(0.01)
Adult casual work available in the community	-0.15*
	(0.01)
Farming inputs available in the community	-0.02
	(0.01)

* 0.10 ** 0.05 * 0.001; SE in parenthesis; marginal effects from probit estimation.

Figure 12. **Predicted child labour from individual level probit estimation**



6 CONCLUSION

In this paper we have built a “risk” indicator for the presence of child labour at community level in Côte d’Ivoire. We have used a concomitant variable mixture model that allows inferring information about the variable of interest, child labour, in cases in which data on this variable are not available. In fact, individual data on child labour are difficult and expensive to collect through survey tools. The proposed model, therefore, uses easily collected characteristics at community level to predict child labour at community level. In particular, the model allows not only to “predict” the risk of child labour in each community, but also to classify communities according to different classes of risk of child labour. This is an important advantage of the model because it allows to identify the different class of risk on the basis of a data driven process, without making use of ad hoc assumption.

We apply the concomitant mixture model to study child labour in Côte d’Ivoire, on the basis of data from the Enquête de Base sur le Travail des Enfants en Côte d’Ivoire pour Développer le Cadre pour les Communautés Cacaoyères Protectrices (CCCP-2017), that provides information at individual and community level.

The concomitant mixture model assumes that the population of interest can be classified in different classes that reflect some unobserved heterogeneity. In our case we interpret the unobserved heterogeneity as risk of child labour. Therefore, the classes, identified by the model, correspond to different risk of child labour. The class membership of each community is hence conditioned on relevant community characteristics.

We first consider the full sample and the individual child employment information averaged at community level to estimate the model and classify the communities in classes of risk. According to the AIC criteria, we identify three classes on the full sample: low risk, medium risk and high risk, representing respectively 78.5%, 12.3% and 9.1% of communities. We find that the most statistically significant community characteristics influencing child labour risk classification are: availability of infrastructure, women’s education, availability of adult casual work and household involvement in cocoa production.

Then, in order to test the predictive power of the model, we randomly exclude some communities for which the individual child labour variable is supposed to not be observed (but community characteristics are supposed to be observed). The excluded communities are classified into one of the three classes, identified using the full sample and the complete data, on the basis of the parameters estimated in the full sample. Performing several tests, the model is able to correctly classify the excluded communities for most of the cases. We also predict child labour at the community level for the excluded communities, finding that the statistical precision of prediction worsen with respect to the prediction of class membership.

EXTENDED REFERENCES AND BIBLIOGRAPHY

Abou, Edouard Pokou (2014). A Re-examination of the determinants of child labour in Côte d'Ivoire. Nairobi: African Economic Research Consortium.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. Biometrics 55, 117–28.

Aitkin, M. (2005). Statistical modelling in GLIM 4. Oxford University Press.

Alfò, M., and Trovato, G. (2004). Semiparametric mixture models for multivariate count data, with application. Econometrics Journal, Royal Economic Society, vol. 7(2), pages 426-454, December.

Boas, Morten; Huser, Anne (2006). Child labour and Cocoa Production in West Africa: The Case of Côte d'Ivoire and Ghana. FAFO Report 522, 2006.

Dayton, C. M., Macready, G. B. (1988). Concomitant-Variable Latent-Class Models. Journal of the American Statistical Association, 83: 173-178.

Grunn, B. and F. Leish (2008). Fexmix Version 2: Finite mixture Model with Concomitant Variable and varying and constant parameter. Journal of statistical software, 28, 1-35.

ICI 2016. *Researching the Impact of Increased Cocoa Yields on the Labour Market and Child Labour Risk in Ghana and Côte d'Ivoire*. ICI Labour Market Research Study, 2016

Institut National de la Statistique (INS). (2014). Enquête Nationale sur la Situation de l'Emploi et du Travail des Enfants (ENSETE 2013). Rapport descriptif sur le travail des enfants.

Institut National de la Statistique (INS) & ICF International (2012). Enquête Démographique et de Santé et à Indicateurs Multiples de Côte d'Ivoire 2011-2012. Calverton, Maryland, USA: INS et ICF International

Kolavalli, S., & Vigneri, M. (2011). Cocoa in Ghana: Shaping the success of an economy. Yes, Africa can: success stories from a dynamic continent, 2011

Laird NM (1978). Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association, 73, 805-11.

Lindsay, B.G (1983a). The geometry of mixture likelihoods, part ii: the exponential family. Annals of Statistics, 11:783-792, 1983.

Lindsay, B.G. and. Lesperance, M.L (1983b). A review of semiparametric mixture models. Journal of statistical planning and inference, 47:29-39, 1995.

McLachlan, G., Peel, D. (2000). Finite Mixture Models. New York: Wiley.

Nkamleu, Guy Blaise (2009). Determinants of Child Labour and Schooling in the Native Cocoa Households of Côte d'Ivoire. AERC Research Paper 190 African Economic Research Consortium, Nairobi October 2009

Nkamleu, Guy Blaise, Kielland, Anne (2006). Modeling farmers' decisions on child labor and schooling in the cocoa sector: a multinomial logit analysis in Côte d'Ivoire. Agricultural economics, Volume 35, Issue 3, November 2006, Pages 319–333

Schrage and Anthony P. Ewing (2005). The Cocoa Industry and Child Labour. The Journal of Corporate Citizenship, No. 18, Corporate Citizenship in Africa (Summer 2005), pp. 99-112

Tulane University Payson Centre for International Development (2015). 2013/14 Survey Research on Child Labor in the West African Cocoa Sector. available at: http://www.childlaborcocoa.org/index.php/2013-14-final-report .

Understanding Children's Work (UCW) Programme (2016). Not just cocoa. Child labour in the agricultural sector in Ghana, Understanding Children's Work Programme Country Report Series (Rome)

Understanding Children's Work (UCW) Programme (2014). Le double défi du travail des enfants et de la marginalisation scolaire dans la région de la CEDEAO. Understanding Children's Work Programme Country Report Series (Rome)

Wang, P., Puterman, M.L., Cockburn, I. and Le N (1996). Mixed poisson regression models with covariate dependent rates. Biometrics, 52:381-400.

Wedel, M. (2002). Concomitant variables in finite mixture models. Statistica Neerlandica 2002, Vol. 56, nr. 3, p. 362-375.

APPENDIX

1. Additional cross validation results

Table A1. Test 1

Class (full sample)	Predicted class (excluded communities)	Ranking Observed child employment (full sample)	Observed child employment (full sample)	Ranking Predicted child employment (excluded communities)	Predicted child employment (excluded communities)
1	1	1	0.00	8	0.15
1	1	2	0.02	7	0.14
1	1	3	0.10	2	0.09
1	1	4	0.10	4	0.11
1	1	5	0.13	5	0.11
1	1	6	0.13	3	0.10
1	1	7	0.22	6	0.13
2	2	8	0.24	9	0.31
3	3	9	0.31	10	0.41
1	1	10	0.42	1	0.09
3	3	11	0.61	11	0.46
Spearman's rho=0.23					

Table A2. Test 2

Class (full sample)	Predicted Class (excluded communities)	Ranking Observed child employment (full sample)	Observed child employment (full sample)	Ranking Predicted child employment (excluded communities)	Predicted child employment (excluded communities)
1	1	1	0.00	10	0.23
1	1	2	0.02	5	0.14
1	1	3	0.05	1	0.09
1	1	4	0.05	3	0.12
1	1	5	0.07	9	0.20
1	1	6	0.16	7	0.14
1	2	7	0.19	11	0.28
1	1	8	0.21	6	0.14
1	1	9	0.25	2	0.12
2	1	10	0.32	8	0.18
1	1	11	0.47	4	0.13
3	3	12	0.61	12	0.37
Spearman's rho=0.18					

Table A3. Test 3

Class (full sample)	Predicted class (excluded communities)	Ranking Observed child employment (full sample)	Observed child employment (full sample)	Ranking Predicted child employment (excluded communities)	Predicted child employment (excluded communities)
1	1	1	0.09	3	0.18
1	1	2	0.11	1	0.11
1	1	3	0.17	2	0.16
3	3	4	0.76	4	0.51
Spearman's rho=0.40					

Table A4. Test 4

Class (full sample)	Predicted class (excluded communities)	Ranking Observed child employment (full sample)	Observed child employment (full sample)	Ranking Predicted child employment (excluded communities)	Predicted child employment (excluded communities)
1	1	1	0	11	0.17
1	1	2	0	12	0.18
1	1	3	0	6	0.12
1	1	4	0	8	0.13
1	1	5	0.03	4	0.12
1	1	6	0.07	7	0.12
1	1	7	0.07	13	0.18
3	3	8	0.07	19	0.51
1	1	9	0.09	2	0.10
1	1	10	0.14	5	0.12
1	1	11	0.14	1	0.10
1	1	12	0.16	10	0.14
3	3	13	0.21	18	0.49
1	1	14	0.27	9	0.13
2	1	15	0.29	15	0.24
1	1	16	0.32	3	0.11
1	1	17	0.33	14	0.22
3	3	18	0.87	17	0.33
2	2	19	0.87	16	0.28
Spearman's rho=0.31					

2. Questions used to define children's employment and child labour

In what follows we detail the questions of the survey instrument used to define child employment and child labour.

Children's Employment. We define a child, aged 5 to 17 years, to be in employment if the following questions were affirmatively answered:

- Durant les 7 derniers jours, avez-vous travaillé pour quelqu'un qui n'est pas un membre de votre ménage, par exemple, pour une entreprise, une société, le gouvernement, un voisin ou n'importe quelle autre personne ?
- Durant les 7 derniers jours, avez-vous travaillé dans une ferme possédée ou louée par un membre de votre ménage, que ce soit dans la culture de céréales ou dans d'autres taches agricoles, ou vous êtes-vous occupé de bétail vous appartenant, à vous ou à un membre de votre ménage ?
- Durant les 7 derniers jours, avez-vous travaillé à votre propre compte ou pour un commerce vous appartenant, à vous ou à un membre de votre ménage, par exemple, comme vendeur de rue, vendeur dans une boutique, en préparant de la nourriture pour la vendre ou pour tout autre commerce ?
- Bien que ..[NOM].. n'ait pas travaillé durant les 7 derniers jours, ..[NOM].. possède-t-il un emploi duquel / une activité de laquelle il était temporairement absent ? (par ex. : absent pour cause de congés ou d'une maladie)

Child Labour. Following the national legislation, children are classified in child labour on the basis of the following criteria:

- children aged 5-13 years in employment; and
- children aged 14-17 years working in (i) hazardous occupations or (ii) working more than 40 hours per week or (iii) working at night;

(i) **Hazardous occupation.**

Working children aged 14 to 17 years were considered to be involved in hazardous occupations if working in:

- **hazardous industries:** include children working in the "mining and quarrying" and "construction" sectors, identified using the following question: *Dans quel secteur est cette activité principale ?*

- **hazardous activities.** Hazardous activities comprise the involvement of working children in: (1) a set of activities identified as hazardous by the national legislation; (2) the exposure to dangerous factors.

(1) *Hazardous activities,* identified using the following questions: *Durant les 7 derniers jours, quelles sont les tâches que tu as exercées ?*

Abattre et découper des arbres ; Brûler les champs ; Pulvériser des insecticides ; Epandre des engrais ; Epandre des fongicides / herbicides / et autres produits chimiques ; Vente /Transport des produits agro-pharmaceutiques (insecticides, herbicides, fongicide, engrais chimiques) ; Chasse; Produire du charbon de bois ; 12. Aller ou revenir du travail seul ou

travailler entre 18h et 6h; Travail à des hauteurs dangereuses à (en haut sur les arbres, escalade;
(2) *Exposure to dangerous factors,* identified using the following question: *Les conditions suivantes s'appliquent-elles à votre environnement de travail ? 1 Oui 2 Non*
Dangerous factors includes: Poussières, vapeurs, gaz (oxygène, ammoniaque); Environnement bruyant; Températures extrêmes ou humidité; Outils coupants / dangereux; Travail souterrain; Travail en hauteur; Éclairage insuffisant; Produits chimiques, peintures; Transport de charges Lourdes; Feu, substances explosives; Pilotage de grue, de machinesVentilation insuffisante.

(ii) working more than 40 hours per week.

Average weekly working hours. We consider the number of days and the number of hours per days worked during the week prior to the survey reported under the following questions: "*Durant les sept derniers jours, combien de jours et d'heures avez-vous travaillé à cet emploi ?*

Quand pendant la semaine:
1. Jours de la semaine (Lundi-Vendredi)
2. Weekends (Samedi-Dimanche)
3. Les week-ends et les jours de semaine

(iii) Night work. Children working at night where identified using the information collected through the following question: *Durant les sept derniers jours,à quel moment travaillez-vous habituellement ?*

1. Toute la journée (du matin au soir)
2. Le matin (avant l'école)
3. Le matin (pendant les heures de cours)
4. L'après-midi (pendant les heures de cours)
5. Apres l'école
6. Les weekends
7. Durant les vacances solaires
8. Ne sait pas

3. Finite mixture model specification

Our outcome of interest, y_i, is the number of children in employment in each community. This is a count variable, but we assume that observed counts y_i, $i = 1, \dots, n$, are realizations of independent Gaussian random variables, since the Gaussian distribution is the superior limit of a Poisson when the number of events tend to infinity. In fact the number of event (children in employment in a community) is larger than 1000 and we can suppose that its distribution approaches that of a Gaussian. We suppose then that y_i is modelled, in a regression context, by defining a generalized linear model (GLM) for the analysed response. Formally, it is modelled as a function of a set of p covariates $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^T$, as follows:

$$\log(y_i) = \eta_i = \beta_0 + \sum_{l=1}^{p} x_{il}\beta_l = \boldsymbol{x}_i^T \boldsymbol{\beta} \quad (1)$$

where a canonical link has been adopted and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ represents the $p + 1$-dimensional vector of regression parameters. Failure of the adopted model to fit the data could be due to misspecification of any of the elements defining the GLM: a simple way to unify these possibilities, is through omitted variables (for a detailed discussion of this topic see Aitkin et al., 2003). We assume that some fundamental covariates were not considered in the model specification and that their joint effect can be summarized by adding a set of unobserved variables u_i, $i = 1, \dots, n$, to the linear predictor, (in theory each observation has a proper unobservable):

$$\log(y_i) = \eta_i = \beta_{0i} + \sum_{l=1}^{p} x_{il}\beta_l = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i \quad (2)$$

In this context, the term $\beta_{0i} = \beta_0 + u_i$ is the overall random intercept where u_i represents a mean zero random deviation from β_0. We have imposed that u_i appears additively in the model, but this assumption can be easily relaxed by associating random parameters to some elements of the adopted covariates set (see Alfò and Trovato, 2004).

The observed responses are assumed independent given the random vector u_i. Treating the u_i's as nuisance parameters and integrating them out, we obtain for the likelihood function the following expression:

$$L(.) = \prod_{i=1}^{n} \left\{ \int_B f(y_i | \boldsymbol{x}_i, u_i) dG(u_i) \right\} \quad (3)$$

where B represents the support for $G(u_i)$, the distribution function of u_i.

Rather than using a parametric specification for $G(.)$, we leave it unspecified and provide a nonparametric maximum likleihood estimator of it according to Laird (1978) and Lindsay (1983a, 1983b).

That is, the integral in equation (3) may be approximated by the following weighted sum:

$$L(.) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} f(y_i | \boldsymbol{x}_i, u_k) \pi_k \right\} = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} [f_{ik} \pi_k] \right\} \quad (4)$$

where $f(y_i | \boldsymbol{x}_i, u_k) = f_{ik}$ denotes the response distribution in the k-th component of the finite mixture. Locations $\boldsymbol{u}_k$ and corresponding masses π_k represent unknown parameters, as well as K, which is treated as fixed and estimated via formal model selection techniques. Denoting with δ the $complete$ parameter vector and proceeding as in Aitkin (1999), we obtain:

$$\frac{\partial \log [L(\boldsymbol{\delta})]}{\partial \boldsymbol{\delta}} = \frac{\partial \log\ l(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\frac{\pi_k f_{ik}}{\sum_{k=1}^{K} \pi_k f_{ik}} \right) \frac{\partial \log f_{ik}}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \frac{\partial \log f_{ik}}{\partial \boldsymbol{\delta}} \quad (5)$$

where w_{ik} represents the posterior probability that the $i-th$ unit comes from the $k-th$ component of the mixture. Equating the derivatives to zero gives the corresponding likelihood equations, which are weighted sums of those for an ordinary GLM with weights w_{ik}. Solving these equations for a given set of weights, and updating the weights from the current parameter estimates defines an Expectation-Maximization (EM) algorithm. From a computational perspective, the EM algorithm is quite simple to implement, and will be sketched in the following section.

The mixture model with concomitant variables

In the previous section we have shown that finite mixture models could be a proper tool to model unobserved heterogeneity under a semi-parametric approach. Moreover, the side result of mixture models is the classification of units in components with homogeneous unobserved characteristics, based on the posterior probability estimates $\widehat{w}_{ik}$. According to a simple mapping rule, in fact, the $i-th$ community can be classified in the $l-th$ component if $\widehat{w}_{il} = \max(\widehat{w}_{i1}, \dots, \widehat{w}_{ik})$. It is worth noticing that each component is characterized by homogeneous values of the estimated latent effects, i.e. conditionally on the observed covariates, communities from that group show a similar structure, at least in the steady state. In equation (5) the w_{ik} weights are estimated in an unconditional way. In the following we specify that the weights could depend on different factors. Strictly speaking, we allow each component of the mixture to have an assigned weight depending on further variables (i.e. concomitant variables). As stressed by Grunn and Leish (2008) and by Dayton and Macready (1988) the parameters of the concomitant variables (i.e. those that could model the probability weight) are simultaneously estimated in the EM process. In our specification the relationship between child labour and the socio-economic variables is then modelled through concomitant variable model where group sizes (i.e. the weights of the mixture) depend on the socio-economic variables. In particular, by assuming that $\pi_k = f(\alpha, c)$ where c is the concomitant variable and α its parameter, we can modify the likelihood in equation (4) obtaining:

$$L(.) = \prod_{i=1}^{n}\left\{\sum_{k=1}^{K} f_{ik}\pi_k(\alpha, c)\right\} = \prod_{i=1}^{n}\left\{\sum_{k=1}^{K} f_{ik}(y_i|\boldsymbol{x}_i, u_k)\pi_k(\alpha, c)\right\} \quad (6)$$

where

$$\forall c \sum_{k=1}^{K} \pi_k(\alpha, c) = 1\,, \pi_k(\alpha, c) > 0 \quad (7)$$

Following Dayton and Mcready (1988), we use the multinomial logit model to estimate the posterior probability:

$$\pi(\alpha, c) = \frac{e^{cT}\alpha_k}{\sum_{k=1}^{K} e^{cT}\alpha_k} \quad (8)$$

As a consequence the estimated posterior probability is:

$$\hat{\pi}_{ik} = \frac{\pi(c_k, \alpha) f_{ik}}{\sum_{k=1}^{K} \pi_k\,(c_k, \alpha) f_{ik}} \quad (9)$$

Summing up, the above model allows us to account more properly for the role of socio-economic characteristics on child labour among communities and to test whether the initial level of socio-economic characteristics affects the probability of belonging to a specific cluster.

Computational details

As it is well known (see, among others, Aitkin, 1999 and Wang et al., 1996), the EM algorithm is designed to maximize the complete data likelihood in expression (3). Let us start denoting with $\boldsymbol{z}_i = (z_{i1}, \dots, z_{ik})$ the unobservable vector of component indicators, where $z_{ik} = 1$, if the community has been sampled from the component of the mixture, and 0 otherwise. Since the component labels in $\boldsymbol{z}$ are unobservable, they have to be treated as missing data. We therefore denote the complete-data with $\boldsymbol{y}_c = \{\boldsymbol{y}, \boldsymbol{z}\}$. The likelihood for the $complete$ data is defined by the following expression:

$$L(.) = \prod_{i=1}^{n}\prod_{k=1}^{K}\{\pi_k f(y_i|x_i, u_k)\}^{z_{ik}} \quad (10)$$

while the corresponding log-likelihood function is given by:

$$l_c(.) = \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_{ik}\left[\log(\pi_k) + \sum\nolimits_i \log\{f(y_i|u_k)\}\right] \quad (11)$$

Since the z_{ik} are treated as missing data, in the $r-th$ iteration of the E-step, we take the expectation of the log-likelihood for $complete$ data over the unobservable component indicator vector z_i given the observed data y and the current parameter estimates, say $\delta^{(r)} = (u^{(r)}, \pi^{(r)})$.

$E - step$: given the current parameter estimates, $\delta^{(r)}$, in the $r - th$ iteration, replace the missing data z_{ik} by the estimated conditional expectation

$$\hat{z}_{ik}\left(\boldsymbol{\delta}^{(r)}\right) = w_{ik}^{(r)} = \frac{\pi(c_k, \alpha^r) f(y_i|x_i, u_k^r)}{\sum_{k=1}^{K}\pi_k\,(c_k, \alpha^r) f(y_i|x_i, u_k^r)} \quad (12)$$

where $\hat{z}_{ik}\left(\boldsymbol{\delta}^{(r)}\right) = w_{ik}^{(r)}$ is the posterior probability that the $i - th$ unit belongs to the $k - th$ component of the mixture.

$M - step$: new $\delta^{(r+1)}$ are given maximizing the function

$$Q\left(\delta^{(r+1)}\middle|\delta^{(r)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}\{\log f(y_i|x_i, u_k^r)\} + \sum_{i=1}^{n}\sum_{k=1}^{K}\log\pi_k(c_i, \alpha^{r+1}) \quad (13)$$

The M-step aims at maximizing the expected value of the complete data likelihood given the observed data and the current parameter estimates. The estimated parameters are the solution of the following M-step equations:

$$\frac{\partial Q}{\partial \pi_k} = \sum_{i=1}^{n} \left\{ \frac{w_{ik}^{(r)}}{\hat{\pi}_k} - \frac{w_{iK}^{(r)}}{\hat{\pi}_K} \right\} = 0 \quad (14)$$

$$\frac{\partial Q}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(r)} \frac{\partial}{\partial \boldsymbol{\delta}} \log(f_{ik}) = 0 \quad (15)$$

To obtain updated estimates of the unconditional probability π_k we replace each z_{ik} by $\hat{z}_{ik}\left(\boldsymbol{\delta}^{(r)}\right)$, and, solving equation (15), we obtain:

$$\hat{\pi}_k^{(r)} = \sum_{i=1}^{n} \frac{w_{ik}^{(r)}}{n} \quad (16)$$

which represents a well known result from ML in finite mixtures. Solutions of equation (15) can be obtained through a Iteratively Weighted Least Squares (IWLS) algortihm.

If the adopted criterium is based on the sequence of likelihood values $l^{(r)}$, $r = 1, \ldots$, the E and M-steps are alternatively repeated until the following relative difference

$$\frac{|l^{(r+1)} - l^{(r)}|}{l^{(r)}} < \epsilon, \qquad \epsilon > 0 \quad (17)$$

changes by an arbitrarily small amount. Since $l^{(r+1)} \geq l^{(r)}$, convergence is obtained with a sequence of likelihood values which are upward bounded. Penalized likelihood criteria (such as AIC, CAIC or BIC) have been used to estimate the number of mixture components.

The use of finite mixtures has some significant advantages over parametric mixture models. First, it allows us to classify communities in clusters characterized by homogeneous values of the latent effects, where this kind of classification is possible only if community heterogeneity does exist. Second, since locations and corresponding probabilities are completely free to vary over the corresponding supports, the proposed approach can readily accommodate extreme and/or strongly asymmetric departures from the Gaussian assumption.