

Transforming Conditional Density Estimation Into a Single Nonparametric Regression Task

Alexander G. Reisach¹

CNRS, MAP5
Université Paris Cité
F-75006 Paris, France

Olivier Collier

MODAL'X, UPL
Université Paris Nanterre
F-92000 Nanterre, France

Alex Luedtke

Department of Health Care Policy
Harvard University
Boston, MA 02115, USA

Antoine Chambaz

CNRS, MAP5
Université Paris Cité
F-75006 Paris, France

Abstract

We propose a way of transforming the problem of conditional density estimation into a single nonparametric regression task via the introduction of auxiliary samples. This allows leveraging regression methods that work well in high dimensions, such as neural networks and decision trees. Our main theoretical result characterizes and establishes the convergence of our estimator to the true conditional density in the data limit. We develop *condensité*, a method that implements this approach. We demonstrate the benefit of the auxiliary samples on synthetic data and showcase that *condensité* can achieve good out-of-the-box results. We evaluate our method on a large population survey dataset and on a satellite imaging dataset. In both cases, we find that *condensité* matches or outperforms the state of the art and yields conditional densities in line with established findings in the literature on each dataset. Our contribution opens up new possibilities for regression-based conditional density estimation and the empirical results indicate strong promise for applied research.

1 Introduction

Conditional density estimation is concerned with estimating the density of a random variable given a set of other covariate random variables. It is one of the fundamental problems of statistics and of interest in applications relying on predictive models where summary statistics like the conditional mean or quantiles are insufficient.

Applications of conditional density estimation include many scenarios in economics, causal inference, and machine learning. In the study of economic inequality, analyses often aim to decompose distributional differences in order to explain them. They can benefit from access to the conditional density for specific covariates to do so on a fine-grained level (Kneib et al. 2023). In finance, conditional densities are useful for credit risk assessment, asset return estimation, and options pricing. In causal inference, the predominant statistical approaches (Pearl 2009; Imbens and Rubin 2015) follow an interventionist account of causation (see Woodward 2005), in which causal effects manifest as changes in distributions. Hence, that describing conditional distributions is of central interest in causality and motivates several works on conditional density estimation, notably Muñoz and van der Laan (2011) and Cevic et al. (2022). In machine learning, conditional density estimation can aid a range of methodologies, including anomaly detection, probabilistic forecasting, and risk-sensitive policy learning (see Nachman and Shih 2020; Gneiting and Katzfuss 2014; Dabney et al. 2018, respectively).

¹Correspondence to alexander.reisach@math.cnrs.fr

On a rudimentary level, conditional density estimation targets the conditional density $y \mapsto f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x)$ of target Y and covariates X with joint density $(x,y) \mapsto f_{X,Y}(x,y)$ and marginal density $x \mapsto f_X(x)$ of the covariates. We assume Y to be univariate, and X to be multivariate. The challenge is to estimate $f_{Y|X}$ from independently and identically distributed observations $(X_i, Y_i)_{1 \leq i \leq n}$ following $f_{X,Y}$. General ideas for approaching this problem center around either identifying values of X with similar conditional densities and then performing estimation of the corresponding marginal, or on estimating $f_{X,Y}$ and f_X separately and then combining them. Both approaches face difficulties if X is high-dimensional and data points therefore sparse, a phenomenon known as the “curse of dimensionality” (Bellman 1961). The curse of dimensionality makes density estimation a difficult problem in high dimensions (see e.g. Wasserman 2006, Section 4.5). Conditional density estimation is usually thought to inherit this challenge, which would render it infeasible in many settings of interest, without prior dimensionality reduction. This contrasts with the spectacular success of high-dimensional regression for estimating summary statistics, in particular the mean, using methods based on neural networks and decision trees. Transforming conditional density estimation into a regression problem that allows for leveraging such methods would open the possibility of transferring their success in high dimensions. This goal motivates our contribution.

Contribution. We propose a way of estimating the conditional density directly and simultaneously for all data points via a single nonparametric regression using auxiliary samples derived from the observations. This allows for the use of any flexible function approximator such as neural networks or boosted decision trees, while mitigating the overfitting problem common to other flexible conditional density estimators. In Section 2, we outline connections of our contribution to the existing literature. We formally introduce our approach in Section 3 and state our theoretical result regarding the convergence of our estimator. The proof can be found in Appendix B. We propose *condensité*², a method that implements our approach, and describe the algorithm and implementation in Section 4. We provide a proof of concept and analyze the influence of hyperparameters on synthetic data in Section 5. In Section 6 we provide a detailed analysis of the performance of our method on a large population survey dataset, and an evaluation and comparison on a satellite imaging dataset. We discuss our findings in Section 7 and provide our conclusion in Section 8.

2 Related Literature

One fundamental approach to conditional density estimation is the parametrization of distributional families using either specific functions (e.g. linear as in Gelman et al. 2013, Chapter 14), or flexible function approximators (e.g. Bishop 1994). Parametric approaches can suffer from model misspecification and have limited flexibility to adapt to different local structures, especially in high-dimensional settings. In our review of the literature, we therefore focus on nonparametric approaches. Among these, we draw the following distinction with the goal of connecting our own contribution. We first cover what we refer to as *density-based* approaches, which estimate the conditional density as a combination of other densities. Then, we give an overview of *regression-based* approaches, which leverage regression for conditional density estimation. Other than through conditional densities, conditional laws can also be characterized through conditional quantiles and conditional cumulative distribution functions, giving rise to quantile regression and distribution regression, which are closely related to conditional density estimation. For details on these tasks we refer to Koenker et al. (2017) and Kneib et al. (2023), respectively.

Density-based approaches. A conditional density $f_{Y|X}$ can be understood as the ratio of $f_{X,Y}(\cdot, \cdot)$, the joint density of target and covariates, and $f_X(\cdot)$, the density of the covariates. A simple way of estimating it is by estimating each component separately, e.g. using kernel density estimators as in Rosenblatt (1969), and then performing division. In high dimensions, the resulting sparsity of data points means that the density estimation may become unstable, for instance if the bandwidth is not selected suitably, which necessitates the use of complex bandwidth selection procedures (e.g. Hyndman et al. 1996; Hall et al. 2004). An alternative is proposed by Efromovich (2007), who use an orthogonal series

²Pronounced *kon-dahn-see-tay*, like the French word for “density”.

estimator. Similarly to the kernel-based methods, this method also struggles in high dimensions due to the curse of dimensionality. This prompts Efromovich (2010) to add adaptive dimension reduction to the orthogonal series approach, which helps in reducing the impact of irrelevant covariates, but the problem remains if many covariates are relevant. Izbicki and Lee (2016) develop and study an orthogonal series estimator that achieves good performance for many relevant covariates with low intrinsic dimensionality, but struggles with irrelevant covariates. Another way of approaching conditional density estimation is to apply unconditional density estimation to observations with similar conditional densities. Cevic et al. (2022) utilize random forests as nearest neighbors method (see Lin and Jeon 2006) in order to construct weighted empirical distributions that can be smoothed into densities. Gao and Hastie (2022) partition the covariate space using boosted decision trees and perform unconditional density estimation using Lindsey’s method (Lindsey 1974) for each partition. Both of Cevic et al. (2022) and Gao and Hastie (2022) benefit from leveraging machine learning methods that perform well in high-dimensional settings.

Regression-based approaches. Viewing each observation (X_i, Y_i) as a sample from the conditional law $f_{Y|X}(\cdot|X_i)$, finding the conditional density at a given point X_i can be seen as a prediction problem. Since there are usually few or no samples at the given point, utilizing regression requires some form of smoothing or localization. The “double kernel” (Hall et al. 1999) approach by Fan et al. (1996) lays the foundation for this idea. It can be understood as performing a localized regression for inference at a given point (x, y) . The targets are evaluations of a first kernel $K_1(Y_i - y)$ for every data point $1 \leq i \leq n$, and the covariates are the corresponding X_i . In the regression, the samples are weighted by another kernel evaluation $K_2(X_i - x)$. The resulting prediction at x gives the conditional density at (x, y) , and by scanning over the range of Y one can construct the entire conditional density for x . The semiparametric estimator of Hjort and Jones (1996) also weighs in X , but does not perform the same kind of target smoothing. Instead of smoothing targets, Muñoz and van der Laan (2011) localize the dependent variable through binning, and estimate sequential conditional bin probabilities with a super learner. Sugiyama et al. (2010) estimate the conditional density directly via a least-squares density ratio objective developed in Kanamori et al. (2009). This method struggles in high dimensions, so Shiga et al. (2015) add a penalty on irrelevant covariates, but the problem remains for settings with many relevant covariates. In order to address settings with many relevant and many irrelevant covariates, Izbicki and Lee (2017) propose an orthogonal series method that leans heavily on regression. Whereas previous orthogonal series methods expand $f_{Y|X}$ in X and Y , they expand only in Y and estimate the coefficients as a function of X through regression. This enables the use of regression methods that perform well in high-dimensions.

Placing our contribution. Our approach builds upon ideas from the regression-based conditional density estimation literature. At its core is a single regression, akin to Sugiyama et al. (2010). However, they target the conditional density directly, which limits the approach to low dimensions, since in high dimensions the data points are too sparse for the regression to perform well. By contrast, we target an auxiliary approximation of the conditional density, which is defined using a probability kernel. Fan et al. (1996) introduce the same kind of target, but perform a covariate-specific regression that is localized by a kernel similarity of the covariates, which does not work well in high dimensions for the same reason as the other kernel-based methods. Thus, our approach – implemented as the method *condensité* – can be seen as a combination of the targets from Fan et al. (1996) with the regression framing in Sugiyama et al. (2010), allowing it to overcome the limitations of either method alone. Moreover, unlike these earlier works, *condensité* uses powerful nonparametric regression methods that have since become widely available. Muñoz and van der Laan (2011) and Izbicki and Lee (2017) also use such regression methods, but in an indirect fashion that offloads much of the complexity to binning and basis expansion, respectively. By utilizing regression directly, *condensité* opens up the possibility of transferring the success of high-dimensional regression methods such as those in Ke et al. (2017); Paszke et al. (2019) to high-dimensional conditional density estimation.

3 Method

This section contains two parts. First, we provide an overview of the approach underlying our method, including a formal description of the setting and key idea, and a numerical illustration of how the auxiliary targets for the regression problem are derived. Second, we outline in what sense our estimator converges, illustrate the intermediate steps of the transformation into a regression problem, and state our theoretical main result.

3.1 Overview

Setting and objective. Suppose there are independent observations $(X_1, Y_1), \dots, (X_n, Y_n)$ following the law P^* on $\mathcal{X} \times [0, 1]$ of a generic couple of random variables (X, Y) . We assume P^* has a joint density $f_{X,Y}^*$ with respect to a dominating product measure $\mu \otimes \text{Unif}[0, 1]$ for some measure μ on a measurable space $(\mathcal{X}, \mathcal{B})$. Our goal is to learn the collection of conditional densities of Y given X , $\{y \mapsto f^*(y|x) : x \in \text{Supp}(P_X^*)\}$, where P_X^* is the marginal law of X under P^* (note that here, and moving forward, we write conditional laws without the subscript $Y|X$ for notational ease).

Key idea. In order to transform conditional density estimation into a regression task, we start by choosing an approximate identity $\{K_h : h > 0\}$. We will rely on the properties of the approximate identity, as stated in Appendix C, to show that the transformation into a regression task recovers f^* as h goes to zero. For instance, K_h could be the density of the centered Gaussian law with variance h^2 .

For our estimator, we choose a closed and convex class \mathcal{F}_h of functions $\mathcal{X} \times [0, 1] \rightarrow \mathbb{R}_+$, $(x, y) \mapsto f(y|x)$. We ensure that there exists a constant $c > 0$ such that, for all $h > 0$, $h\|K_h\|_\infty \leq c$ and $h\|f\|_\infty \leq c$ for every $f \in \mathcal{F}_h$ (hence the h in the index of \mathcal{F}_h).

We set up our regression as follows. For each $1 \leq i \leq n$, independently of the observations, we draw M auxiliary samples Y'_{i1}, \dots, Y'_{iM} independently from $\text{Unif}[0, 1]$ and compute every $K_h(Y_i - Y'_{im})$ – the closer Y'_{im} is to Y_i , the larger the result. We then find the best $f \in \mathcal{F}_h$ to approximate the $K_h(Y_i - Y'_{im})$ by $f(Y'_{im}|X_i)$ across all data points with $1 \leq i \leq n$ and $1 \leq m \leq M$ with respect to the least-squares criterion

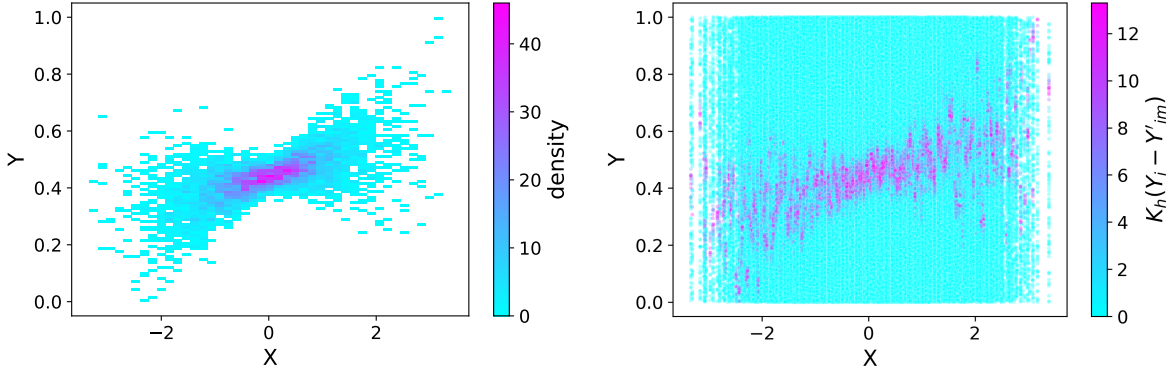
$$\hat{f} := \arg \min_{f \in \mathcal{F}_h} \sum_{i=1}^n \sum_{m=1}^M [K_h(Y_i - Y'_{im}) - f(Y'_{im}|X_i)]^2. \quad (1)$$

We analyze in what sense \hat{f} approximates the true conditional density f^* . For this we rely on the theoretical counterpart to Equation (1) given by

$$f_{\mathcal{F}_h}^* := \arg \min_{f \in \mathcal{F}_h} \int_{x \in \mathcal{X}} \int_{y \in [0, 1]} \int_{y' \in [0, 1]^M} \sum_{m=1}^M [K_h(y - y'_m) - f(y'_m|x)]^2 f_{X,Y}^*(x, y) \prod_{m=1}^M dy'_m dy \mu(dx). \quad (2)$$

Note that M affects the estimation of \hat{f} in Equation (1), but the inner integral in Equation (2) is the same regardless of the value of the hyperparameter M , so $f_{\mathcal{F}_h}^*$ is invariant to the choice of M .

Illustration. For illustration, we provide a simple example using the two-dimensional data-generating mechanism with $X \sim \mathcal{N}(0, 1)$ and $Y|X \sim \mathcal{N}(X, 0.25 + X^2)$. Figure 1 shows how the joint law of (X, Y) gives rise to our regression targets (note that we rescale Y to $[0, 1]$ by min-max scaling). We sample $n = 5000$ observations independently, use $M = 100$ auxiliary samples per observation, and set a sharpness of $h = 0.03$. The joint density shown in Figure 1a is concentrated around the point $(0, 0)$, and one can recognize the linear slope and heteroskedasticity of the conditional density. In Figure 1b one can see how the transformations $K_h(Y_i - Y'_{im})$ map samples from the joint law into our regression targets mimicking the conditional law. The concentration around $(0, 0)$ is gone, and instead each vertical slice along the x-axis of the right-hand plot approximates a renormalized version of the corresponding slice of the left-hand plot.



(a) Empirical joint density of (X, Y) .

(b) Scatterplot of the regression targets.

Figure 1: Illustration of the key idea.

3.2 Theoretical Main Result

What follows is a formal definition of the steps involved in defining our estimator, and the result of our theoretical analysis of its convergence. For reference, we provide glossaries of central objects in Appendix A. Definitions of random variables are given in Table 4, functions in Table 5, and laws in Table 6. The full proofs can be found in Appendices B and C. We use the following notational convention: for P a measure on some measurable space (S, \mathcal{B}) and $\varphi: S \rightarrow \mathbb{R}$ a measurable function, $P\varphi$ denotes the integral $\int \varphi dP$; in particular, $P\varphi^2 = \|\varphi\|_{L^2(P)}^2$.

Let ℓ be the loss function mapping any $f \in \mathcal{F}_h$ to the function $\ell[f]$ characterized by

$$\ell[f](x, \bar{y}', \bar{z}) = \frac{1}{M} \sum_{m=1}^M (f(y'_m | x) - z_m)^2,$$

with $x \in \mathcal{X}$, $\bar{y}' = (y'_1, \dots, y'_M) \in [0, 1]^M$, and $\bar{z} = (z_1, \dots, z_M) \in \mathbb{R}_+^M$. This corresponds to $1/M$ times the inner summand of Equation (1) and the integrand of Equation (2).

Let \bar{P}_h^* be the joint law of (X, \bar{Y}', \bar{Z}_h) , where (X, Y) is drawn from P^* , $\bar{Y}' = (Y'_1, \dots, Y'_M)$ is drawn from $(\text{Unif}[0, 1])^{\otimes M}$ independently of (X, Y) , and $\bar{Z}_h = (Z_{h,1}, \dots, Z_{h,M})$ with $Z_{h,m} = K_h(Y - Y'_m)$ for each $1 \leq m \leq M$. We denote the minimizer of the risk function induced by ℓ under the law $Q^* = P_X^* \otimes \text{Unif}[0, 1]$ over all square integrable functions as

$$f_h^* := \arg \min_{f \in L^2(Q^*)} \{ \bar{P}_h^* \ell[f] \}.$$

Since $\{K_h: h > 0\}$ is an approximate identity, for h approaching zero the convolution $K_h * \varphi$ converges to φ in $L^p(\mathbb{R})$ for any function $\varphi \in L^p(\mathbb{R})$ with $p \geq 1$, and converges pointwise to φ for any bounded and uniformly continuous φ (see Proposition 2 in Appendix C). By Lemma 2 it follows that f_h^* converges to f^* as h goes to zero (cf. Fan et al. 1996, Equation 2.1).

The function f_{h, \mathcal{F}_h}^* defined by Equation (2) is the projection of f_h^* onto \mathcal{F}_h and can be written as

$$f_{\mathcal{F}_h}^* := \arg \min_{g \in \mathcal{F}_h} \{ Q^*(g - f_h^*)^2 \}.$$

This is the function approximated by the empirical risk minimizer defined in Equation (1) and below as

$$\hat{f} := \arg \min_{f \in \mathcal{F}_h} \{ \bar{P}_{h,n} \ell[f] \},$$

where $\bar{P}_{h,n}$ is the empirical law that puts mass $1/n$ on every $(X_i, \bar{Y}'_i, \bar{Z}_{hi})$. A schematic of the relationship between f^* , f_h^* , $f_{\mathcal{F}_h}^*$, \hat{f} can be seen in Figure 2.

We assess the proximity of $f \in \mathcal{F}_h$ to f_h^* through the excess risk induced by loss ℓ with respect to $f_{\mathcal{F}_h}^*$ which we denote as

$$\begin{aligned}\mathcal{E}_h(f) &:= Q^*(f - f_h^*)^2 - \inf_{g \in \mathcal{F}_h} Q^*(g - f_h^*)^2 \\ &= Q^*(f - f_h^*)^2 - Q^*(f_{\mathcal{F}_h}^* - f_h^*)^2 \\ &= \bar{P}_h^*(\ell[f] - \ell[f_{\mathcal{F}_h}^*]).\end{aligned}\tag{3}$$

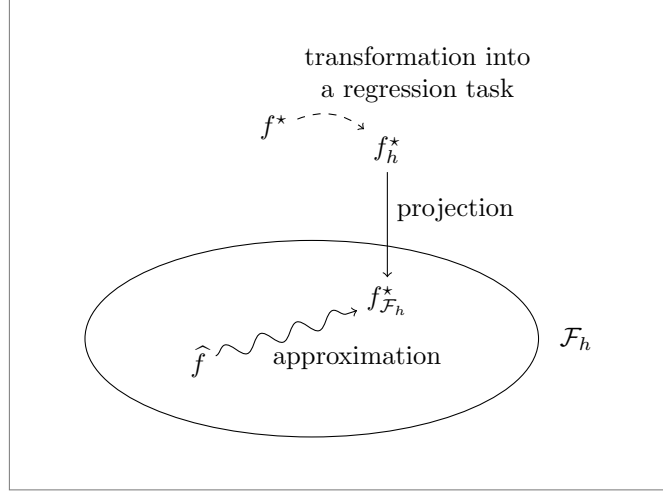


Figure 2: Illustration of the transformation of the conditional density estimation problem into a regression task. We use an approximate identity with bandwidth h to define the regression task, choose a function class \mathcal{F}_h for the regression, and optimize the fit on a set of observations.

Assumption 1. Suppose that \mathcal{F}_h is chosen such that its complexity in terms of covering numbers is controlled by the following condition (van der Vaart and Wellner 1996, Section 2.5.1): there exist a measurable envelope F_h , constants $A > 0$ and $\rho \in [0, 1)$ such that for every probability measure Q on $\mathcal{X} \times [0, 1]$ with finite support, for all $\varepsilon > 0$,

$$\log N(\varepsilon, \mathcal{F}_h, L^2(Q)) \leq \left(\frac{A \|F_h\|_{L^2(Q)}}{\varepsilon} \right)^{2\rho}.\tag{4}$$

This implies the finiteness of the uniform entropy integral $\sup_Q \int_0^1 \sqrt{1 + \log N(\varepsilon, \mathcal{F}_h, L^2(Q))} d\varepsilon$. The excess risk of our estimator, as defined in Equation (3), is bounded by the following concentration inequality.

Theorem 1 (Excess risk bound). Under Assumption 1, there exists a constant $C > 0$ such that, for all tuning parameters $M \geq 1$ and $h > 0$, for all $t > 0$,

$$\bar{P}_h^* \left(\mathcal{E}_h(\hat{f}) \geq C \left[\frac{c^2 2^{1-\rho}}{h^2 \sqrt{n} (1-\rho)} + \frac{t}{n} \left(\sqrt{\left(\frac{2c}{h} \right)^2 + \frac{1}{8}} - \frac{2c}{h} \right)^{-2} \right] \right) \leq e^{1-t}.\tag{5}$$

The proof of the theorem is an adaptation of that of Koltchinskii (2011, Theorem 4.3) using Maurer (2016, Corollary 1), and is deferred to Appendix B. For simplicity, we use the constant envelope function c/h which, without loss of generality, upper bounds the envelope F_h used in Assumption 1. The result is improved when using F_h directly. A function with a larger (constant) envelope may lead to $f_{\mathcal{F}_h}^*$ being

closer to f_h^* , but would also lead to a looser bound. Observe further that, as h goes to zero, the second summand is of the form $\frac{tc^2}{nh^2}(1024 + o(1))$ and goes to infinity. Yet, h must go to zero for f_h^* to converge to f^* by Lemma 2, so there is a bias-variance trade-off. Finally, note that Theorem 1 does not capture an effect of the choice of the hyperparameter M . The forthcoming simulation study in Section 5.2 investigates its impact empirically.

4 Implementation

Evaluation. We evaluate the discrepancy between our estimator \hat{f} and the true conditional density f^* using the integrated squared error (ISE). The ISE is given by

$$\iint \left(\hat{f}(y|x) - f^*(y|x) \right)^2 P_X^*(dx) dy = \iint \hat{f}^2(y|x) P_X^*(dx) dy - 2 \iint \hat{f}(y|x) f^*(y|x) P_X^*(dx) dy + C, \quad (6)$$

where C is a constant that does not depend on \hat{f} . We estimate the ISE up to C (hence the negative values in our experiments). In our experiments we approximate the ISE over a grid using trapezoidal integration.

Algorithm. We provide a modular implementation that allows for using a variety of predictors for conditional density estimation. It consists of the following two parts.

1. `Condensite`
2. `CondensitePredictor`

The `Condensite` class implements the functionality for transforming samples from the joint law into targets for conditional density estimation, and allows for fitting a `CondensitePredictor`. In doing so, it takes care of appropriate variable scaling and evaluates the fit on a validation set. `CondensitePredictor` is an abstract class that acts as an interface that predictors passed to `Condensite` have to implement. In principle, any regression model can serve as the basis for a `CondensitePredictor`. This modularity allows our approach to take advantage of a wide range of different implementations. Alongside our Python implementation we provide examples using predictors from the `PyTorch` (Paszke et al. 2019), `LightGBM` (Ke et al. 2017), and `scikit-learn` (Pedregosa et al. 2011) libraries. The following pseudocode algorithm captures the essential part of the data fitting functionality of `Condensite`.

Algorithm 1: Fitting *condensité*.

Input: data $(X_i, Y_i)_{1 \leq i \leq n}$, number of auxiliary samples M , sharpness h .

Output: estimator \hat{f} , validation ISE.

1. Split data into training and validation sets.
 2. Repeat training samples of X a total of M times.
 3. Min-max-scale the training samples of Y to $[0, 1]$.
 4. Generate auxiliary samples \bar{Y}' by sampling M points uniformly in $[0, 1]$ per training data point.
 5. Compute the targets as $K_h(Y_i - Y'_{im})$ per training and auxiliary sample point.
 6. Column-stack the repeated training samples of X and auxiliary samples \bar{Y}' into a feature matrix.
 7. Standardize each feature and the targets. *// for training stability*
 8. Find \hat{f} by minimizing Equation (1).
 9. Compute ISE of \hat{f} on validation data. *// includes inverse scaling of predicted targets*
 10. **Return:** \hat{f} , ISE.
-

Post-processing. To make sure we obtain a density, we set $\hat{f}(\cdot) := \max\{0, \hat{f}(\cdot)\}$, and then redefine $\hat{f}(\cdot) := \hat{f}(\cdot) / \int_a^b \hat{f}(y) dy$.

5 Proof of Concept and Hyperparameter Analysis

We test *condensité* on synthetic data from the following three synthetic data-generating mechanisms inspired by Izbicki and Lee (2016, Section 4.1). The primary goal of our synthetic data experiments is to provide a proof of concept of *condensité* in a simple and controlled setting, and to examine the impact of the hyperparameters h and M . For each mechanism, we use 20 independently standard Gaussian distributed covariates X .

1. **Single relevant covariate:** the single covariate $X^{(1)}$ determines the conditional density via $Y|X \sim \mathcal{N}(X^{(1)}, 0.25 + (X^{(1)})^2)$.
2. **Data on manifold:** the conditional density depends on the angle θ mapped to $[0, 2\pi]$ between $X^{(1)}$ and $X^{(2)}$ via $Y|X \sim \mathcal{N}(\theta, 0.5)$.
3. **Non-sparse data:** the conditional density depends on all covariates via $Y|X \sim \mathcal{N}(\text{mean}(X), 0.5)$.

Note that the first setting is heteroskedastic to make it more challenging; the relevant covariate affects Y as shown in our illustration in Figure 1. We evaluate two different versions of *condensité*:

- *condensité (NN)*: a version using a neural network as predictor.
- *condensité (tree)*: a version using a gradient boosted decision tree as predictor.

5.1 Proof of Concept

We train the methods on 10000 samples and evaluate on another 1000 samples for each data-generating mechanism. We choose $M = 100$ and $h = 10^{-2}$ for the *condensité* methods; for details on the hyperparameters, see Appendix D.1. For comparison, we run the following methods from the literature:

- *LinCDE* (Gao and Hastie 2022),
- *DRF* (Cevic et al. 2022),
- *FlexCode* (Izbicki and Lee 2016),
- *condensier* (based on Muñoz and van der Laan 2011).

LinCDE and *DRF* employ decision tree variants that have proven successful in various regression and classification contexts. *FlexCode* and *condensier* utilize flexible regression models. We choose the decision tree predictor provided in the *FlexCode* implementation and call this version *FlexCode (tree)*, for *condensier* we keep the inbuilt default predictor. All four methods promise good performance in high dimensions. We provide details on the implementation and hyperparameters in Appendix D.2.

Proof of concept results. The results in terms of ISE (lower is better) are shown in the table below. The estimates by *condensité* are on par with those of the other methods from the literature. The landmark (specific covariates) analysis in Appendix E.1 shows that all methods successfully learn a density close to the ground truth, with *condensité (NN)* and *LinCDE* providing the smoothest fits. Note that the same hyperparameter configuration is used across datasets for each method, including the *condensité* variants. This suggests that *condensité* can achieve good out-of-the-box results using generic choices for the hyperparameters M and h . The following analysis explores their influence in greater detail.

	Single relevant covariate	Data on manifold	Non-sparse data
<i>condensier</i>	-0.2865	-0.3561	-0.2651
<i>condensité (NN)</i>	-0.3311	-0.3853	-0.2800
<i>FlexCode (knn)</i>	-0.2561	-0.2637	-0.2694
<i>FlexCode (tree)</i>	-0.3068	-0.3701	-0.2584
<i>DRF</i>	-0.3453	-0.3973	-0.2815
<i>LinCDE</i>	-0.3364	-0.3964	-0.2832

5.2 Hyperparameter Analysis: The Effect of M and Its Interaction With h

The hyperparameters M and h are at the core of our method. We investigate their interplay on our synthetic data. We use the same hyperparameters as in the proof of concept experiment (see Appendix D.1), which yield good results for $h = 10^{-2}$ and $M = 100$ on all of our data-generating mechanisms. To investigate the role of M and its interplay with h , we run a grid search over $h \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and $M \in \{1, 10, 20, 40, 80, 160, 320\}$ for each mechanism. Figure 3 shows the results in terms of ISE for *condensité* (NN) on each of the data-generating mechanisms. Note that we limit the training to 20 epochs, so it is possible that some of the results would improve with further training and our findings here may reflect the speed of convergence as well as the best attainable performance.

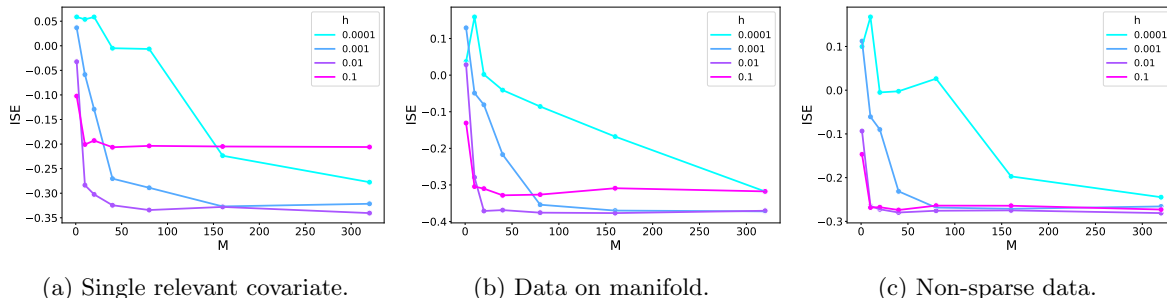


Figure 3: ISE for *condensité* (NN) with different M and h on the synthetic data-generating mechanisms.

Across the three settings we find that larger values of M tend to give better results, and that lower h require larger M to work well. For well-chosen h , here 0.01 and 0.001, even a moderate number of M seems to improve the performance substantially. If h is chosen too high, the value of M seems to make little difference, and the method does not perform as well. Since h must go to zero for the transformation into a regression to recover the true conditional density, this is not surprising. For M large enough, performance stabilizes at different levels depending on h . The results for *condensité* (tree) are shown in Figure 8 in Appendix E.2. They exhibit the same qualitative trends, although the performance benefit of larger M materializes and also levels out faster than for *condensité* (NN). In our following experiments we use $h = 10^{-2}$ and $M = 100$ throughout for both of our estimators, since this combination works well on our synthetic data and appears to provide sufficient flexibility in \mathcal{F}_h without requiring an excessively high number of auxiliary samples.

6 Evaluation on Real-World Data

To assess the promise of our approach under realistic conditions, we evaluate the performance of the *condensité* methods and compare them to other methods from the literature on two challenging real-world datasets.

6.1 IPUMS Current Population Survey Data

We evaluate *condensité* (NN), *condensité* (tree), and the other methods from the literature on a dataset compiled from the IPUMS Current Population Survey (CPS) database³ (Flood et al. 2024). The IPUMS-CPS is a monthly household survey of more than 65000 households from the United States of America. It collects data for social science research and has been conducted since the year 1962. It is conducted jointly by the US Census Bureau and the US Bureau of Labor Statistics. The IPUMS database has been used particularly widely in economics research. For a recent example, see Blanchet et al. (2022), who propose a real-time measure of inequality by estimating income distributions conditional on different economic and demographic variables. We use a selection of variables measuring geographic, demographic, work-related,

³<https://cps.ipums.org/cps/>

and education-related characteristics of individuals. Given these covariates, we estimate the conditional density of yearly total personal income. The hyperparameters used for the different methods are listed in Appendix D, and we describe our data preprocessing steps in Appendix F.1.

Our evaluation consists of a quantitative and a qualitative part. First, we compare the ISE of our methods and the other methods from the literature introduced previously in Section 5. Second, we visually inspect the estimates of the different methods at two landmarks that correspond to typical questions in the economic literature on income inequality. Since the true conditional density is unknown, we manually construct a local empirical density as a reference point for the plausibility of the estimates and discuss the consistency of the estimates with established findings in the economics literature.

6.1.1 Empirical Comparison

Our CPS dataset comprises 113,104 observations of 26 covariates, of which 6 are multi-valued, with the remaining ones being binary. We swap *FlexCode (knn)* for *FlexCode (tree)*, a tree-based version better suited to tabular data. For the other methods we keep the same hyperparameters as before (see Appendix D for details on the hyperparameters). We train on 80% of the data and evaluate the ISE on the test set given by the remaining 20%. For *DRF*, we limit the number of training samples to 50000 due to the large memory requirements of the implementation. The results in terms of ISE (lower is better) can be seen in Table 1.

	CPS
<i>condensité (NN)</i>	−0.0414
<i>condensité (tree)</i>	−0.0241
<i>FlexCode (tree)</i>	−0.0389
<i>LinCDE</i>	−0.0297
<i>DRF</i>	−0.0317
<i>condensier</i>	−0.0249

Table 1: ISE results on the IPUMS-CPS dataset.

We find that *condensité (NN)* achieves the best result, with *FlexCode (tree)* a close second. *DRF* also performs well, especially given that it has been trained only on about half of the dataset. *LinCDE* is still close to *DRF*, but the other methods from the literature and *condensité (tree)* perform substantially worse. The difference between *condensité (NN)* and *condensité (tree)* points to the choice of regressor as a decisive factor in our approach. The *condensité* methods are the only ones in this group that use regression directly for conditional density estimation and thus depend strongly on the inductive biases of different regression methods. As can be seen, a well-chosen regressor allows for highly competitive performance. Although *condensité (NN)* and *FlexCode (tree)* appear to perform best, it is important to note that all methods offer a high degree of flexibility and different architecture choices, hyperparameters, or preprocessing steps may affect the results. Moreover, we caution that good summary performance may not suffice for an estimator to be useful in practice. In real-world applications, high-variance fits and artifacts may distort downstream analyses. The following landmark analyses therefore explore qualitative aspects of the results.

6.1.2 Landmark Analysis for Realistic Use-Cases

We choose two landmarks corresponding to realistic use-cases in the study of economic income inequality, and visualize the estimates of the different methods as well as a local empirical density for comparison. In this section we focus on the results of the *condensité* methods, and of *FlexCode (tree)* as the second-best performing method (see Table 1). A visualization and analysis of the remaining methods can be found in Appendix F.2.

Skill premium. We inspect the conditional income densities estimated by our methods for two landmarks that differ only in years of education. The income gap that arises between people with similar characteristics but different education is referred to as *skill premium* and is traditionally analyzed by comparing mean wages between broadly defined groups of similar education (see e.g. Autor et al. 2008; Acemoglu and Autor 2011). However, analyzing wage distributions for specific subgroups and beyond simple summary statistics may be necessary to gain a more complete picture (Firpo et al. 2018). Conditional density estimation allows for doing so on a fine-grained level. We choose the characteristics in the table below and vary the education level (see boldfaced row). For reference we construct a local empirical density by including observations within 10 years of age and within 10 weekly work hours. This yields 115 observations for the 12 year education landmark, and 175 observations for the 16 year education landmark.

<i>characteristic</i>	<i>landmark</i>
personal	40 years, male, no children, white (race)
geography	metropolitan (5+ million)
nativity	US-born (self and parents)
work hours	40 (weekly, constant), private sector wage or salary
education	12 years 16 years

Figure 4 shows the methods’ estimates and the local empirical density. Overall, the methods’ estimates are closely aligned with the local empirical density and show a strong and heterogeneous skill premium. For more years of education total incomes are not only higher, they are also substantially more dispersed. This trend is in line with the long-term developments of increasing skill premiums and job polarization described in Acemoglu and Autor (2011, Sections 2.4, 2.5). Both *condensité* methods provide smooth fits, with the shape of the *condensité* (NN) estimate being slightly closer to the local empirical density for the 16 years of education. Although the *FlexCode* (tree) estimate exhibits the correct trend, it is notably more bumpy, which is presumably an artifact resulting from the cosine basis expansion. The other methods from the literature (shown in Figure 9) also capture the correct trend, yet either yield undesirably high-variance estimates akin to *FlexCode* (tree), or over-smooth the more concentrated density for 12 years of education.

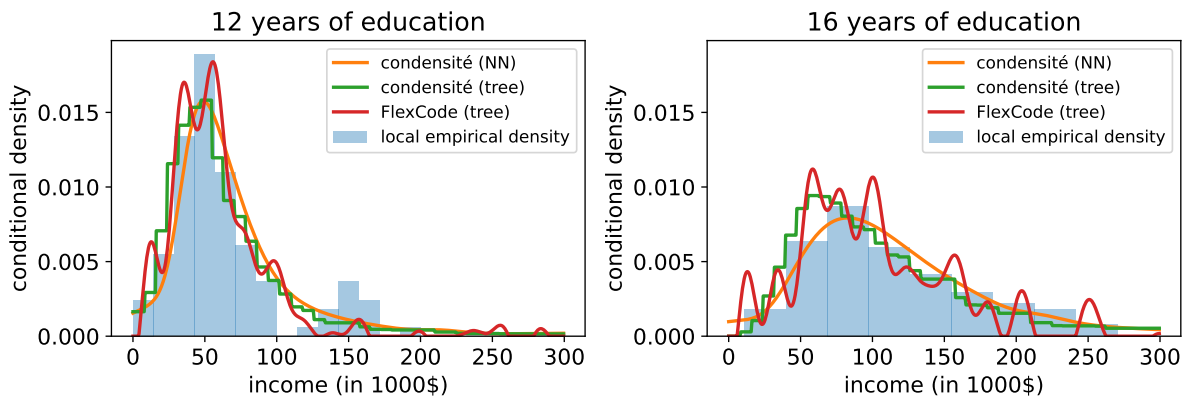


Figure 4: Conditional income density for 12 and 16 years of education with otherwise identical covariates.

Geographic income dispersion. Another frequent subject of analysis in inequality research is geographic income dispersion. In this context, different groups, e.g. with different levels of education are usually analyzed separately (see e.g. Baum-Snow and Pavan 2012). Conditional density estimation allows for a fine-grained distinction of groups on the level of specific covariate sets. For illustration we choose a landmark with the characteristics specified in the table below. For reference, we estimate a local empirical density for observations within 10 years of age and within 10 weekly work hours. This yields 84 observations for the non-metropolitan landmark, and 118 observations for the metropolitan landmark

<i>characteristic</i>	<i>landmark</i>
personal	40 years, female, two children, white (race)
geography	non-metropolitan metropolitan (5+ million)
nativity	US-born (self and parents)
work hours	40 (weekly, constant), private sector wage or salary
education	16 years

The results are shown in Figure 5. In the methods’ estimates, as well as the local empirical density, we observe higher wages and a greater wage dispersion for the metropolitan covariate set compared to the non-metropolitan one. This is in line with Baum-Snow and Pavan (2012), who report strong urban wage premiums for highly skilled workers such as those specified by our covariates. The *FlexCode (tree)* estimate is closer to the local empirical non-metropolitan density than the *condensité* methods, yet again suffers from pronounced spurious bumps. The other methods from the literature (see Figure 10) capture the same trend as the ones presented here but either yield high-variance estimates akin to *FlexCode (tree)*, or over-smooth the more concentrated non-metropolitan density.

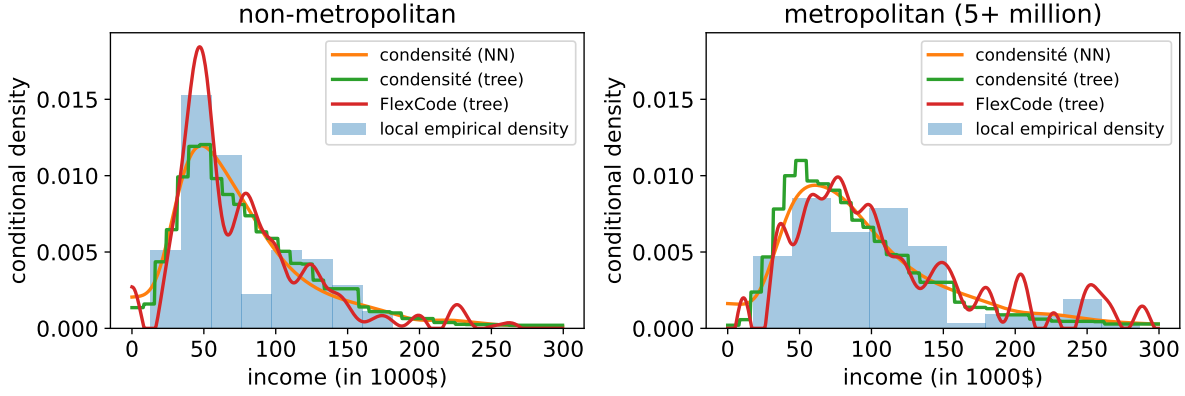


Figure 5: Conditional income density for (non-)metropolitan inhabitants with otherwise identical covariates.

6.2 ESA ICC Satellite Imaging Data

For an evaluation and comparison of our methods on unstructured data in a realistic setting, we compile a dataset on above-ground biomass (AGB) estimation. We use AGB labels for the year 2020 provided by the European space agency’s (ESA) climate change initiative (CCI)⁴ as described in Santoro and Cartus (2025), and corresponding Sentinel-1A satellite images⁵ as features. AGB is an “essential climate variable” (Penman et al. 2003), and as such of crucial interest to climate modelling. Uncertainty quantification in AGB modelling is listed as a desirable target in Santoro and Cartus (2025, Table 2-1), making this an application of conditional density estimation with high potential impact (see also Araza et al. 2022). We construct a dataset by taking the average AGB in tons per square kilometer, and corresponding 100×100 two-channel satellite image patches. We choose the geographic region, weather, and season to ensure the sensor data used for imaging is informative about AGB. Our training and test dataset are constructed from two disjoint regions of tropical savannah in the Northern Territory of Australia, and contain 14653 and 14683 observations respectively. We train and evaluate all algorithms on the training region (chosen to be the one with the greater range of AGB values) using 80% of the data for training, and the remaining 20% as a hold-out set for final performance evaluation. Then, we additionally evaluate the methods on the test region. We refer to Appendix G.1 for more details on the data, a description of the preprocessing steps, and a visualization of the satellite images from which the features patches are extracted.

⁴<https://archive.ceda.ac.uk/>

⁵<https://geodes-portal.cnes.fr>

6.2.1 Empirical Comparison.

For this application on AGB satellite image data, we use *condensité (CNN)*, a version using a convolutional neural network (CNN) and a fully connected neural network head with skip connections between them. To take full advantage of the end-to-end training enabled by *condensité*, we provide the auxiliary target coordinates Y'_{im} as a third image channel to the convolutional layers, and again as a separate feature to the head. We use the same architecture, save for the auxiliary sample coordinates, as basis coefficient estimator in *FlexCode (CNN)*. To conduct as fair a comparison as possible, we use a neural network of the same architecture trained to predict the labels as a feature extractor for the other methods. The features are the CNN activations of the network. For further details and hyperparameters, see Appendix D.

Table 2 shows the methods’ performance on the held-out 20% of the training region dataset, and Table 3 shows the performance on the complete test region dataset. Performance is measured in terms of ISE (lower is better). We find that the *condensité* methods are among the best-performing methods in both evaluations, with the two CNN-based methods *condensité (CNN)* and *FlexCode (CNN)* performing best of all. We observe further that the ISE results are consistently worse on the held-out 20% of the training region data than on the test region data. This can be explained by the training region being closer to the coast and to the equator and thus having greater biomass. As a result, it presents a more complex estimation task with fewer of the near-zero AGB areas that characterize the test data region (compare the label panel of Figure 6 and Figure 13). Although this complicates the quantitative interpretation of the test region performance, we note that the *condensité* methods and *FlexCode (CNN)* are the top-performing methods on both datasets. The performance difference between the training and test data is particularly pronounced for *condensier*; we treat this point in further detail in the following qualitative assessment.

	ESA ICC AGB
<i>condensité (CNN)</i>	−0.9224
<i>condensité (tree)</i>	−0.8785
<i>FlexCode (CNN)</i>	−0.8929
<i>LinCDE</i>	−0.8329
<i>DRF</i>	−0.7775
<i>condensier</i>	−0.5565

Table 2: **Training region.** ISE on held-out 20% of data.

	ESA ICC AGB
<i>condensité (CNN)</i>	−2.1245
<i>condensité (tree)</i>	−1.9646
<i>FlexCode (CNN)</i>	−2.2726
<i>LinCDE</i>	−1.5222
<i>DRF</i>	−1.3861
<i>condensier</i>	−1.8709

Table 3: **Test region.** ISE on all data points.

6.2.2 Visualization and Qualitative Assessment

The ISE results in Table 2 and Table 3 give a rough idea about comparative model performance, but other factors such as smoothness or variance of the estimates may be relevant for downstream tasks. To obtain a qualitative understanding of the methods’ performances, we visualize the tail width, skew, and mode of the per-pixel estimate for each method. We measure tail width as the difference between the ninth decile d_9 and first decile d_1 . To measure skewness, we compute the Bowely-type (Bowley 1920) statistic

$$\frac{(d_9 - m) - (m - d_1)}{d_9 - d_1}, \quad (7)$$

where m is the median; we upscale the value to the AGB range for greater visual contrast. For comparison, we also show the AGB labels. In Figure 6 below we show visualizations of the test region data estimates for all methods. A visualization of all methods on the complete training region dataset can be found in Appendix G.2.

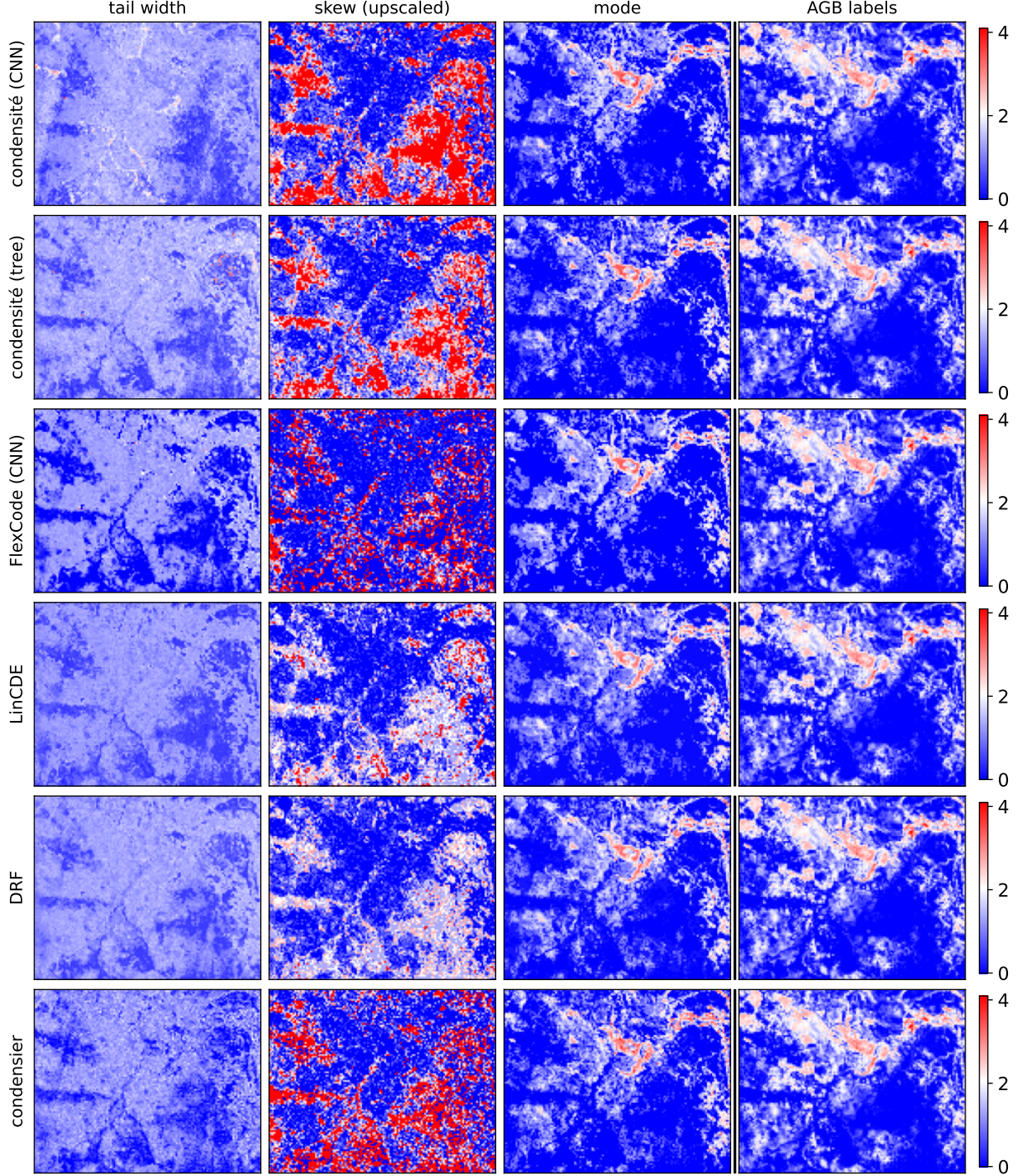


Figure 6: **Test region.** Visualization of method estimate summary statistics and labels.

We find that the mode closely resembles the labels for all methods, indicating a stable dominant peak in the conditional density. Tail width is lowest in low-AGB areas, indicating narrow densities. The test region has many such areas, making the density estimation task comparatively less complex than for the training region (cf. Figure 13). The two best-performing methods *condensité (CNN)* and *FlexCode (CNN)* appear to have the greatest contrast in tail width between high- and low-AGB areas, although the latter generally has narrower tails. Araza et al. (2022, Section 4) assert a positive association between AGB estimates

and uncertainty across AGB datasets, hence we may expect such a contrast. The two worst-performing methods on the test region, *DRF* and *LinCDE*, have more uniform tail widths, indicating possible underfitting. We observe the greatest difference between the model estimates is in their skew. Both *condensité* methods have high skewness in contiguous areas corresponding to low-AGB areas. Skewness is similarly high though more dispersed for *condensier*. For *FlexCode (CNN)*, skewness appears to be highest at the borders of low-AGB areas. To a lesser extent, this can also be seen for *LinCDE* and *DRF*, which are somewhere in between the *condensité* methods and *FlexCode (CNN)*, although again with less pronounced contrasts between high- and low-AGB areas. Overall, the experiment demonstrates that the *condensité* methods, like *FlexCode*, can match or outperform the other methods in the literature in terms of ISE. It also highlights the complementary nature of the qualitative aspects of the different methods, pointing to different strengths for different purposes.

7 Discussion and Limitations

Considerations for applying *condensité*. Applying *condensité* involves choosing a predictor as well as the hyperparameters M and h . The choice of predictor can be a major determinant for overall performance, and we recommend testing different methods and architectures. Choosing suitable values for M and h is essential for good performance, but doing so is not difficult. Higher values of M and lower values of h are better, and low values of h require high values of M . For optimal performance, it is advisable to evaluate different predictor and hyperparameter combinations, yet our results suggest that good performance can be achieved without extensive tuning.

Theoretical result. Our main theoretical result establishes the convergence of our estimator to the true conditional density in the data limit and as h goes to zero. The latter condition is inherited from the transformation into a regression task (see also Fan et al. 1996, Equation 2.1). There is a bias-variance trade-off between choosing h large for a tighter bound by Theorem 1, and choosing h small for the transformation to remain close to the true density by Lemma 2. Although we find clear empirical evidence corroborating the benefit of the auxiliary samples, our proof strategy does not capture an effect of M . Since the auxiliary samples are only block-wise independent, the influence of M cancels out in the symmetrization step. A proof strategy that overcomes this limitation and captures the influence of M would help outline the advantage gained by our approach and could guide the choice of M and h . We are not aware of any technique allowing this, and such a result would represent a substantial advancement.

Empirical evaluation. On synthetic data we find that increasing the number of auxiliary samples M greatly and consistently improves the performance of our estimators. This highlights the benefit of our approach compared to using the observations alone. In our evaluation on challenging real-world datasets, we find that the *condensité* estimators match or outperform the other methods in the literature. On a large population survey dataset we find that *condensité (NN)* outperforms the other methods and yields estimates with desirable smoothness properties. The latter point may be particularly interesting for applications that rely on comparing densities at different points, and underscores the ability to choose an inductive bias via the regressor as an advantage of *condensité*. On a satellite image dataset, we showcase its ability to integrate different machine learning methods and find that *condensité (CNN)* achieves state of the art performance. A qualitative assessment of the estimates highlights that methods with similar performance may provide estimates with very different characteristics. Hence, our approach complements the existing ones in the literature beyond the raw performance of the *condensité* methods.


Limitations. We do not investigate or compare the computational requirements of our method in this work, but there is no doubt that the computation time will tend to increase in the number of auxiliary samples M . Though we consider it unlikely for this to be prohibitive since *condensité* can benefit from the widely available infrastructure for large-scale regression, it may slow down prototyping and complicate parameter tuning. Unlike some other methods in the literature, *condensité* also does not yet allow for multivariate dependent variables. Such an extension would present a natural and exciting

opportunity for future research. Our experimental setup is aimed at mimicking realistic use-cases, but future research is needed to assess how these empirical results generalize to other datasets and data types, including multimodal data. In addition, our evaluation criteria are of a generic nature. For practitioners, domain-specific criteria or evaluations on downstream tasks may help better outline the benefit of our methods in practice.

8 Conclusion

We propose a way of transforming conditional density estimation into a single nonparametric regression task using auxiliary samples that encourage nearby points in the feature space to yield similar estimates. This acts as regularization and mitigates the overfitting common to flexible conditional density estimators. We develop *condensité*, a method that implements this approach. By interpreting regression outputs as density estimates directly, rather than viewing them as an input to another parametrization, it is able to leverage the full expressive power of high-dimensional regression methods and is trainable without modification using the same libraries and infrastructure. We establish convergence of the estimator for function classes of limited capacity and numerically study how the number of auxiliary samples affects performance. Our results show that more auxiliary samples generally improve performance, and that the hyperparameters are straightforward to tune. In our empirical evaluation, we find that *condensité* with a neural network as predictor outperforms the state of the art in the literature on a large real-world population survey dataset and matches it on a satellite imaging dataset. Qualitatively, we find that the *condensité* models provide different and complementary estimates to existing methods. Moreover, the ability to adapt the inductive bias via the choice of regressor may make our approach interesting to a wide range of applications. Conditional density estimation in high dimensions is often considered a challenging problem out of reach of current methods. Our findings demonstrate that it can be feasible and effective under realistic conditions. Overall, our results indicate that *condensité* in particular, and state of the art methods more generally, hold strong promise for applied research in domains requiring flexible conditional density estimates.

Acknowledgements

We thank Julia M. Schmidt for suggesting the population survey application, Rémy Abergel for his advice on satellite image datasets, and Myrto Limnios for feedback on the theoretical part of the work. AGR received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945332 .

References

- [1] Daron Acemoglu and David Autor. [Chapter 12 - Skills, Tasks and Technologies: Implications for Employment and Earnings](#). In: vol. 4. Handbook of Labor Economics. Elsevier, 2011, pp. 1043–1171 (cit. on p. 11).
- [2] Arnan Araza, Sytze De Bruin, Martin Herold, Shaun Quegan, Nicolas Labriere, Pedro Rodriguez-Veiga, Valerio Avitabile, Maurizio Santoro, Edward TA Mitchard, Casey M Ryan, et al. A comprehensive framework for assessing the accuracy and uncertainty of global above-ground biomass maps. In: *Remote Sensing of Environment* 272 (2022), p. 112917 (cit. on pp. 12, 14).
- [3] David H Autor, Lawrence F Katz, and Melissa S Kearney. [Trends in U.S. Wage Inequality: Revising the Revisionists](#). In: *The Review of economics and statistics* 90.2 (2008), pp. 300–323 (cit. on p. 11).
- [4] Nathaniel Baum-Snow and Ronni Pavan. [Understanding the City Size Wage Gap](#). In: *The Review of economic studies* 79.1 (2012), pp. 88–127 (cit. on pp. 11, 12).
- [5] R. E. Bellman. [Adaptive Control Processes - A Guided Tour](#). Princeton University Press, 1961 (cit. on p. 2).

- [6] Christopher M Bishop. [Mixture Density Networks](#). Tech. rep. 1994 (cit. on p. 2).
- [7] Thomas Blanchet, Emmanuel Saez, and Gabriel Zucman. [Real-Time Inequality](#). Tech. rep. National Bureau of Economic Research, 2022 (cit. on p. 9).
- [8] Olivier Bousquet. [A Bennett concentration inequality and its application to suprema of empirical processes](#). In: *Comptes Rendus Mathématique* 334.6 (2002), pp. 495–500 (cit. on p. 21).
- [9] Arthur Lyon Bowley. *Elements of Statistics*. 4th ed. 1920 (cit. on p. 13).
- [10] Domagoj Cevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. [Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression](#). In: *Journal of Machine Learning Research* 23.333 (2022), pp. 1–79 (cit. on pp. 1, 3, 8).
- [11] Tianqi Chen and Carlos Guestrin. [XGBoost: A Scalable Tree Boosting System](#). In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD. Association for Computing Machinery, 2016, pp. 785–794 (cit. on p. 28).
- [12] CNES. [GEODES – CNES Portal for Earth Observation Data Access](#). Accessed: 2025-11-16. 2025 (cit. on p. 32).
- [13] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. [Distributional Reinforcement Learning With Quantile Regression](#). In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018 (cit. on p. 1).
- [14] Richard M Dudley. [Uniform Central Limit Theorems](#). Vol. 142. Cambridge University Press, 2014 (cit. on p. 25).
- [15] Sam Efromovich. [Conditional density estimation in a regression setting](#). In: *The Annals of Statistics* 35.6 (2007), pp. 2504–2535 (cit. on p. 2).
- [16] Sam Efromovich. [Dimension Reduction and Adaptation in Conditional Density Estimation](#). In: *Journal of the American Statistical Association* 105.490 (2010), pp. 761–774 (cit. on p. 3).
- [17] Stefan Elfving, Eiji Uchibe, and Kenji Doya. [Sigmoid-weighted linear units for neural network function approximation in reinforcement learning](#). In: *Neural Networks* 107 (2018), pp. 3–11 (cit. on p. 27).
- [18] Jianqing Fan, Qiwei Yao, and Howell Tong. [Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamical Systems](#). In: *Biometrika* 83.1 (1996), pp. 189–206 (cit. on pp. 3, 5, 15).
- [19] Sergio P Firpo, Nicole M Fortin, and Thomas Lemieux. [Decomposing Wage Distributions Using Recentered Influence Function Regressions](#). In: *Econometrics* 6.2 (2018), p. 28 (cit. on p. 11).
- [20] Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Megan Schouweiler, and Michael Westberry. [IPUMS CPS: Version 12.0 \[dataset\]](#). Minneapolis, MN: IPUMS, 2024 (cit. on p. 9).
- [21] Zijun Gao and Trevor Hastie. [LinCDE: Conditional Density Estimation via Lindsey’s Method](#). In: *Journal of Machine Learning Research* 23.52 (2022), pp. 1–55 (cit. on pp. 3, 8).
- [22] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. [Bayesian Data Analysis \(3rd ed.\)](#) Chapman and Hall/CRC, 2013 (cit. on p. 2).
- [23] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. [Deep Sparse Rectifier Neural Networks](#). In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 315–323 (cit. on p. 27).
- [24] Tilmann Gneiting and Matthias Katzfuss. [Probabilistic Forecasting](#). In: *Annual Review of Statistics and Its Application* 1.1 (2014), pp. 125–151 (cit. on p. 1).
- [25] Peter Hall, Jeff Racine, and Qi Li. [Cross-Validation and the Estimation of Conditional Probability Densities](#). In: *Journal of the American Statistical Association* 99.468 (2004), pp. 1015–1026 (cit. on p. 2).

- [26] Peter Hall, Rodney CL Wolff, and Qiwei Yao. [Methods for Estimating a Conditional Distribution Function](#). In: *Journal of the American Statistical Association* 94.445 (1999), pp. 154–163 (cit. on p. 3).
- [27] Dan Hendrycks and Kevin Gimpel. [Gaussian Error Linear Units \(GELUs\)](#). In: *arXiv preprint* (2016) (cit. on p. 27).
- [28] Nils Lid Hjort and M Chris Jones. [Locally Parametric Nonparametric Density Estimation](#). In: *The Annals of Statistics* (1996), pp. 1619–1647 (cit. on p. 3).
- [29] Rob J Hyndman, David M Bashtannyk, and Gary K Grunwald. [Estimating and Visualizing Conditional Densities](#). In: *Journal of Computational and Graphical Statistics* 5.4 (1996), pp. 315–336 (cit. on p. 2).
- [30] Guido W. Imbens and Donald B. Rubin. [Causal Inference for Statistics, Social, and Biomedical Sciences](#). Cambridge University Press, 2015 (cit. on p. 1).
- [31] Sergey Ioffe and Christian Szegedy. [Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift](#). In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. PMLR, July 2015, pp. 448–456 (cit. on p. 27).
- [32] Rafael Izbicki and Ann B Lee. [Converting high-dimensional regression to high-dimensional conditional density estimation](#). In: *Electronic Journal of Statistics* 11 (2017), pp. 2800–2831 (cit. on p. 3).
- [33] Rafael Izbicki and Ann B Lee. [Nonparametric conditional density estimation in a high-dimensional regression setting](#). In: *Journal of Computational and Graphical Statistics* 25.4 (2016), pp. 1297–1316 (cit. on pp. 3, 8).
- [34] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. [A Least-squares Approach to Direct Importance Estimation](#). In: *The Journal of Machine Learning Research* 10 (2009), pp. 1391–1445 (cit. on p. 3).
- [35] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. [LightGBM: A Highly Efficient Gradient Boosting Decision Tree](#). In: *Advances in Neural Information Processing Systems* 30 (2017) (cit. on pp. 3, 7).
- [36] Diederik P Kingma and Jimmy Lei Ba. [Adam: A Method for Stochastic Optimization](#). In: *ICLR: International Conference on Learning Representations*. 2015, pp. 1–15 (cit. on p. 27).
- [37] Thomas Kneib, Alexander Silbersdorff, and Benjamin Säfken. [Rage Against the Mean – A Review of Distributional Regression Approaches](#). In: *Econometrics and Statistics* 26 (2023), pp. 99–123 (cit. on pp. 1, 2).
- [38] Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng. [Handbook of Quantile Regression](#). In: (2017) (cit. on p. 2).
- [39] Vladimir Koltchinskii. [Oracle inequalities in empirical risk minimization and sparse recovery problems: École D’Été de Probabilités de Saint-Flour XXXVIII-2008](#). Vol. 2033. Springer Science & Business Media, 2011 (cit. on pp. 6, 20–22, 24, 25).
- [40] Yi Lin and Yongho Jeon. [Random Forests and Adaptive Nearest Neighbors](#). In: *Journal of the American Statistical Association* 101.474 (2006), pp. 578–590 (cit. on p. 3).
- [41] JK Lindsey. [Comparison of probability distributions](#). In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.1 (1974), pp. 38–47 (cit. on p. 3).
- [42] Ilya Loshchilov and Frank Hutter. [Decoupled Weight Decay Regularization](#). In: *International Conference on Learning Representations*. 2019, pp. 1–11 (cit. on p. 27).
- [43] Andreas Maurer. [A Vector-Contraction Inequality for Rademacher Complexities](#). In: *Algorithmic Learning Theory*. Springer, 2016, pp. 3–17 (cit. on pp. 6, 21, 24).
- [44] Iván Díaz Muñoz and Mark J van der Laan. [Super learner based conditional density estimation with application to marginal structural models](#). In: *The International Journal of Biostatistics* 7.1 (2011) (cit. on pp. 1, 3, 8).

- [45] Benjamin Nachman and David Shih. [Anomaly detection with density estimation](#). In: *Physical Review D* 101.7 (2020), p. 075042 (cit. on p. 1).
- [46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. [PyTorch: An Imperative Style, High-Performance Deep Learning Library](#). In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on pp. 3, 7).
- [47] Judea Pearl. [Causality: Models, Reasoning, and Inference](#). Cambridge University Press, 2009 (cit. on p. 1).
- [48] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. [Scikit-learn: Machine learning in Python](#). In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 7).
- [49] Jim Penman, Michael Gytarsky, Taka Hiraishi, Thelma Krug, Dina Kruger, Riitta Pipatti, Leandro Buendia, Kyoko Miwa, Todd Ngara, Kiyoto Tanabe, et al. [Good Practice Guidance for Land Use, Land-Use Change and Forestry](#). Tech. rep. The Intergovernmental Panel on Climate Change (IPCC), 2003 (cit. on p. 12).
- [50] Prajit Ramachandran, Barret Zoph, and Quoc V Le. [Searching for Activation Functions](#). In: *arXiv preprint arXiv:1710.05941* (2017) (cit. on p. 27).
- [51] Murray Rosenblatt. Conditional probability density and regression estimators. In: *Multivariate analysis II* 25 (1969), p. 31 (cit. on p. 2).
- [52] Maurizio Santoro and Oliver Cartus. [ESA Biomass Climate Change Initiative \(Biomass_cci\): Global datasets of forest above-ground biomass for the years 2007, 2010, 2015, 2016, 2017, 2018, 2019, 2020, 2021 and 2022, v6.0](#). NERC EDS Centre for Environmental Data Analysis, 2025 (cit. on pp. 12, 32).
- [53] Motoki Shiga, Voot Tangkaratt, and Masashi Sugiyama. [Direct conditional probability density estimation with sparse feature selection](#). In: *Machine Learning* 100.2 (2015), pp. 161–182 (cit. on p. 3).
- [54] Bernard W Silverman. [Density Estimation for Statistics and Data Analysis](#). Routledge, 2018 (cit. on p. 28).
- [55] Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. [Conditional Density Estimation via Least-Squares Density Ratio Estimation](#). In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 781–788 (cit. on p. 3).
- [56] Michel Talagrand. [New concentration inequalities in product spaces](#). In: *Inventiones mathematicae* 126.3 (1996), pp. 505–563 (cit. on p. 21).
- [57] Alexandre B. Tsybakov. [Introduction to Nonparametric Estimation](#). Springer Series in Statistics. Revised and extended from the 2004 French original, translated by Vladimir Zaiats. Springer, 2009 (cit. on p. 27).
- [58] A van der Vaart and JA Wellner. [Weak convergence and empirical processes](#). Springer Series in Statistics. Springer, 1996 (cit. on p. 6).
- [59] Larry Wasserman. [All of Nonparametric Statistics](#). Springer, 2006 (cit. on p. 2).
- [60] James Woodward. [Making things happen: A theory of causal explanation](#). Oxford University Press, 2005 (cit. on p. 1).

A Glossary

(X, Y)	Random variable on $\mathcal{X} \times [0, 1]$ drawn from P^\star .
\bar{Y}'	$\bar{Y}' = (Y'_1, \dots, Y'_M)$ is drawn from $(\text{Unif}[0, 1])^{\otimes M}$ independently of (X, Y) .
\bar{Z}_h	$\bar{Z}_h = (Z_{h,1}, \dots, Z_{h,M})$ with $Z_{h,m} = K_h(Y - Y'_m)$ for each $1 \leq m \leq M$.

Table 4: Glossary of random variables related to (X, Y) .

$f_{X,Y}(\cdot, \cdot)$	Joint density of X and Y .
$f^\star(\cdot \cdot)$	Conditional density of Y given X .
$f_h^\star(\cdot \cdot)$	Conditional density of Y'_m given X .
$f_{\mathcal{F}_h}^\star(\cdot \cdot)$	Projection of f_h^\star onto \mathcal{F}_h .
$\hat{f}(\cdot \cdot)$	Approximation of $f_{\mathcal{F}_h}^\star$ on n samples of (X, Y) .

Table 5: Glossary of densities related to (X, Y) .

$\mathcal{X} \times [0, 1]$	Sample space of (X, Y) .
P^\star	Law of the experiment of interest on $\mathcal{X} \times [0, 1]$.
P_X^\star	Marginal law of X under P^\star .
\bar{P}_h^\star	Joint law of (X, \bar{Y}', \bar{Z}_h) .
$\bar{P}_{h,n}$	The empirical counterpart to \bar{P}_h^\star that puts mass $1/n$ on every $(X_i, \bar{Y}'_i, \bar{Z}_{hi})$, where $(X_1, \bar{Y}'_1, \bar{Z}_{h1}), \dots, (X_n, \bar{Y}'_n, \bar{Z}_{hn})$ are independently drawn from \bar{P}_h^\star .
Q^\star	Product of P_X^\star with the uniform law on $[0, 1]$.
Q_{Mn}	The empirical law that puts mass $1/(nM)$ on every (X_i, Y'_{im}) .

Table 6: Glossary of laws related to (X, Y) .

B Proof of the Main Result

We aim to bound the probability of the excess risk $\hat{\delta} := \mathcal{E}_h(\hat{f})$ of our estimator, as defined in Equation (3), exceeding a certain threshold. In a first step (Appendix B.1), we start with the insight that $\hat{\delta}$ can be expressed as the supremum over the empirical process $(\bar{P}_{h,n} - \bar{P}_h^\star)$ evaluated over the loss differences of pairs of functions from a class of functions $\mathcal{F}_h(\delta)$ with excess risk no greater than δ . The deviation of this expression from its expectation, denoted as $\phi_{h,n}(\hat{\delta})$, can be bounded probabilistically by a version of Talagrand's concentration inequality. The probabilistic bound is only informative if it bounds an excess risk that is lower than that of at least one function. We define σ_n^t as the lowest excess risk beyond which that is always the case. This allows us to use the concentration inequality in Koltchinskii (2011, Theorem 4.3) to bound the chance of $\hat{\delta}$ exceeding σ_n^t . In a second step (Appendix B.2), we aim to isolate the part of σ_n^t that depends on the unknown true law. We denote that part as $\phi_{h,n}^\sharp$ and refer to the remainder as s_n^t . In a third step (Appendix B.3), we aim to express the difference $(\bar{P}_{h,n} - \bar{P}_h^\star)$ that features in $\phi_{h,n}^\sharp$ solely in terms of the empirical law using Rademacher sums via a standard symmetrization argument and the use

of Maurer (2016, Corollary 1). In a fourth step (Appendix B.4), we use the assumption on the covering number integral to bound the expectation of the Rademacher sum. Finally (Appendix B.5), we combine our bound on $\phi_{h,n}^\sharp$ with a solution for the remainder s_n^t and plug them in for σ_n^t to obtain our result.

B.1 Controlling the excess risk

Our goal is to obtain a tail inequality on the excess risk $\hat{\delta}$ with respect to $f_{\mathcal{F}_h}^*$. To this end we apply Koltchinskii (2011), Theorem 4.3.

Defining an excess risk measure. For any $\delta > 0$, let the δ -minimal subset of functions be defined as

$$\mathcal{F}_h(\delta) := \{f \in \mathcal{F}_h : \mathcal{E}_h(f) \leq \delta\}.$$

In light of Equation (3), for any $\epsilon \in (0, \hat{\delta}]$, for any $f' \in \mathcal{F}_h(\epsilon)$,

$$\begin{aligned} \hat{\delta} &= \bar{P}_h^*(\ell[\hat{f}] - \ell[f']) + \bar{P}_h^*(\ell[f'] - \ell[f_{\mathcal{F}_h}^*]) \\ &\leq \bar{P}_h^*(\ell[\hat{f}] - \ell[f']) + \epsilon \\ &= \bar{P}_{h,n}(\ell[\hat{f}] - \ell[f']) + (\bar{P}_h^* - \bar{P}_{h,n})(\ell[\hat{f}] - \ell[f']) + \epsilon \\ &\leq (\bar{P}_h^* - \bar{P}_{h,n})(\ell[\hat{f}] - \ell[f']) + \epsilon \\ &\leq \sup_{f_1, f_2 \in \mathcal{F}_h(\hat{\delta})} |(\bar{P}_{h,n} - \bar{P}_h^*)(\ell[f_1] - \ell[f_2])| + \epsilon. \end{aligned} \tag{8}$$

This relates the excess risk $\hat{\delta}$ to the supremum of the empirical process $(\bar{P}_{h,n} - \bar{P}_h^*)$ over

$$\Lambda(\hat{\delta}) := \{\ell[f_1] - \ell[f_2] : f_1, f_2 \in \mathcal{F}_h(\hat{\delta})\},$$

which we denote as

$$\|\bar{P}_{h,n} - \bar{P}_h^*\|_{\Lambda(\hat{\delta})} := \sup_{\lambda \in \Lambda(\hat{\delta})} |(\bar{P}_{h,n} - \bar{P}_h^*)\lambda|. \tag{9}$$

The deviation of Equation (9) from its expectation

$$\phi_{h,n}(\hat{\delta}) := \mathbb{E}_{\bar{P}_h^*} \|\bar{P}_{h,n} - \bar{P}_h^*\|_{\Lambda(\hat{\delta})}, \tag{10}$$

can be bounded using the Bousquet (2002) version of Talagrand's inequality (Talagrand 1996).

A distribution-dependent upper bound. For said inequality, we introduce the squared diameter of $\mathcal{F}_h(\delta)$ as

$$D_h(\delta) := \sup_{f_1, f_2 \in \mathcal{F}_h(\delta)} \bar{P}_h^*(\ell[f_1] - \ell[f_2])^2.$$

Now, fix arbitrarily $t > 0$. For any $\delta > 0$ one can define a probabilistic upper bound $U_n^t(\delta)$ on Equation (9) such that, if $\delta \geq \sup\{\delta \in (0, 1] : \delta \leq U_n^t(\delta)\}$, then the bound is exceeded with probability at most e^{1-t} . Let the \flat -transform for any non-negatively valued function $\delta \mapsto \psi(\delta)$ be

$$\delta \mapsto \psi^\flat(\delta) := \sup_{\sigma \geq \delta} \frac{\psi(\sigma)}{\sigma}.$$

For any $\delta > 0$, let $\Lambda(\delta)$ and $\phi_{h,n}(\delta)$ be defined as in Equation (9) and Equation (10), with δ substituted for $\hat{\delta}$, and then let

$$V_n^t(\delta) := 4 \left[\phi_{h,n}^b(\delta) + \sqrt{D_h^b(\delta)} \sqrt{\frac{t}{n\delta}} + \frac{t}{n\delta} \right] \quad (11)$$

be an upper bound on $(U_n^t)^b(\delta)$. Note that $D_h^b(\delta)$ can be upper bounded by a constant (see the forthcoming Proposition 1), and thus every term of $V_n^t(\delta)$ is decreasing in δ . Let

$$\sigma_n^t := \inf\{\delta > 0: V_n^t(\delta) \leq 1\}. \quad (12)$$

In essence, $V_n^t(\sigma_n^t) \leq 1$, hence $U_n^t(\sigma_n^t) \leq \sigma_n^t$, which implies that $\sigma_n^t \geq \sup\{\delta \in (0, 1]: \delta \leq U_n^t(\delta)\}$. Therefore, σ_n^t is a probabilistic upper bound on Equation (9), which, in view of Equation (8), is itself an upper bound on $\mathcal{E}_h(\hat{f})$. In summary, Koltchinskii (2011), Theorem 4.3 yields the distribution-dependent concentration inequality

$$\bar{P}_h^* \left(\mathcal{E}_h(\hat{f}) \geq \sigma_n^t \right) \leq e^{1-t}. \quad (13)$$

B.2 Deriving a weaker but simpler control of the excess risk

In this step we aim to isolate the part of the risk bound from Equation (13) that depends on the unknown true distribution \bar{P}_h^* . The definition of σ_n^t in Equation (12) can be separated into two parts. Let the first part be given by

$$\phi_{h,n}^\sharp(1/8) := \inf \left\{ \delta > 0: \phi_{h,n}^b(\delta) \leq 1/8 \right\}. \quad (14)$$

For the second part we derive an upper bound on D_h^b from Equation (11) based on the following Lemma.

Lemma 1. *For all $\delta > 0$, $\mathcal{F}_h(\delta) \subseteq \{f \in \mathcal{F}_h: Q^*(f - f_{\mathcal{F}_h}^*)^2 \leq \delta\}$.*

Proof. Consider the functional

$$\begin{aligned} \psi: \mathcal{F}_h &\rightarrow \mathbb{R}_+, \\ f &\mapsto \frac{1}{2} Q^*(f_h^* - f)^2. \end{aligned}$$

It is convex and its gradient at any $f \in \mathcal{F}_h$ is given by the Riesz representer $\nabla\psi(f) = -(f_h^* - f)$. This means in particular that $\nabla\psi(f) \cdot g = \langle -(f_h^* - f), g \rangle_{Q^*}$. The projection of f_h^* onto \mathcal{F}_h is the minimizer $f_{\mathcal{F}_h}^* = \arg \min_{f \in \mathcal{F}_h} \psi(f)$. Since \mathcal{F}_h is closed and convex, it follows that for all $f \in \mathcal{F}_h$,

$$\langle -(f_h^* - f_{\mathcal{F}_h}^*), f - f_{\mathcal{F}_h}^* \rangle_{Q^*} \geq 0.$$

This implies the inequality below, from which the result follows:

$$\begin{aligned} Q^*(f - f_h^*)^2 &= Q^*(f - f_{\mathcal{F}_h}^* + f_{\mathcal{F}_h}^* - f_h^*)^2 \\ &= Q^*(f - f_{\mathcal{F}_h}^*)^2 + Q^*(f_{\mathcal{F}_h}^* - f_h^*)^2 + 2\langle f - f_{\mathcal{F}_h}^*, f_{\mathcal{F}_h}^* - f_h^* \rangle_{Q^*} \\ &\geq Q^*(f - f_{\mathcal{F}_h}^*)^2 + Q^*(f_{\mathcal{F}_h}^* - f_h^*)^2. \end{aligned}$$

□

Proposition 1 (Upper bound on $D_h^b(\delta)$). *For all $\delta > 0$, $D_h^b(\delta) \leq 16 \frac{c^2}{h^2}$.*

Proof. Note that for all $x \in \mathcal{X}$, $\bar{y} \in [0, 1]^M$, $\bar{z} \in [0, c/h]^M$, and $f_1, f_2 \in \mathcal{F}_h(\delta)$,

$$\begin{aligned} (\ell[f_1](x, \bar{y}, \bar{z}) - \ell[f_2](x, \bar{y}, \bar{z}))^2 &\leq \left(\frac{1}{M} \sum_{m=1}^M \left| (z_m - f_1(y_m|x))^2 - (z_m - f_2(y_m|x))^2 \right| \right)^2 \\ &\leq \left(2 \frac{c}{h} \times \frac{1}{M} \sum_{m=1}^M |(f_1 - f_2)(y_m|x)| \right)^2, \end{aligned}$$

since each of the squared differences in the absolute value of the first line is bounded by $(c/h)^2$. Note that the squared distance between any two functions in \mathcal{F}_h is at most 4δ by the triangle inequality, since $\mathcal{F}_h(\delta)$ is a subset of $\{f \in \mathcal{F}_h : Q^*(f - f_{\mathcal{F}_h}^*)^2 \leq \delta\}$ by Lemma 1. It follows that by convexity (second inequality) and the said squared triangle inequality in $\mathcal{F}_h(\delta)$ (third inequality),

$$\bar{P}_h^*(\ell[f_1] - \ell[f_2])^2 \leq 4 \frac{c^2}{h^2} Q^*(f_1 - f_2)^2 \leq 4 \frac{c^2}{h^2} \sup_{f_1, f_2 \in \mathcal{F}_h(\delta)} Q^*(f_1 - f_2)^2 \leq 16 \frac{c^2}{h^2} \delta.$$

From the definitions of $D_h(\delta)$ and the \flat -transform, it follows directly that $D_h^\flat(\delta) \leq 16 \frac{c^2}{h^2} \delta$. \square

Given the bound on $D_h^\flat(\delta)$ from Proposition 1, we return to defining the second part of σ_n^t . Let s_n^t be the unique solution in δ to

$$\sqrt{16 \frac{c^2}{h^2}} \sqrt{\frac{t}{n\delta}} + \frac{t}{n\delta} = 1/8,$$

given by

$$\frac{t}{n} \left(\sqrt{\left(\frac{2c}{h} \right)^2 + \frac{1}{8}} - \frac{2c}{h} \right)^{-2}. \quad (15)$$

Since both terms in the right-hand side expression of Equation (11) are decreasing in δ , it straightforwardly holds that $\sigma_n^t \leq \phi_{h,n}^\sharp(1/8) + s_n^t$. Plugging this into Equation (13), we obtain

$$\bar{P}_h^*(\mathcal{E}(\hat{f})) \geq \phi_{h,n}^\sharp(1/8) + s_n^t \leq e^{1-t}. \quad (16)$$

In the following steps we aim to express $\phi_{h,n}^\sharp$ in terms of the empirical distribution using the geometry of the function class \mathcal{F}_h .

B.3 Symmetrization

The object $\phi_{h,n}^\sharp$ still depends on the unknown distribution \bar{P}_h^* via Equation (10). Our goal in this step is to express the difference between empirical and theoretical distribution solely in terms of the empirical distribution using Rademacher sums. These can then be upper bounded in terms of the complexity of the class \mathcal{F}_h . For any $\delta > 0$, we introduce the set of differences

$$\mathcal{G}(\delta) := \{f - f_{\mathcal{F}_h}^* : f \in \mathcal{F}_h(\delta)\}.$$

Let $\epsilon_{11}, \dots, \epsilon_{nM}$ be independent Rademacher random variables drawn independently of $\bar{P}_{h,n}$. For any $g \in \mathcal{G}(\delta)$, we define the Rademacher sum

$$R_{h,n}g := \frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M \epsilon_{im} g(Y'_{im} | X_i),$$

and denote its supremum analogously to Equation (9) as

$$\|R_{h,n}\|_{\mathcal{G}(\delta)} = \sup_{g \in \mathcal{G}(\delta)} |R_{h,n}g|.$$

We now relate first $\phi_{h,n}(\delta)$ and then $\phi_{h,n}^\sharp(\delta)$, whose definitions involve $\mathcal{F}_h(\delta)$, to $\|R_{h,n}\|_{\mathcal{G}(\delta)}$, whose definition involves $\mathcal{G}(\delta)$. Let henceforth $\mathbb{E}_{\bar{P}_h^*, \epsilon}$ be the expectation operator with respect to the product of $(\bar{P}_h^*)^{\otimes n}$ and $\text{Rad}(1/2)^{\otimes nM}$. By the standard symmetrization argument (see e.g. the rightmost inequality of the particular case of Koltchinskii 2011, Theorem 2.1), we have that

$$\phi_{h,n}(\delta) \leq 2\mathbb{E}_{\bar{P}_h^*, \epsilon} \left[\sup_{f_1, f_2 \in \mathcal{F}_h(\delta)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_{i1} (\ell[f_1] - \ell[f_2])(X_i, \bar{Y}_i', \bar{Z}_i) \right| \right]. \quad (17)$$

The absolute value can be dropped. Then, using $\ell[f_{\mathcal{F}_h}^*]$ as a pivot, and the fact that Rademacher law is symmetric around zero, Equation (17) is equivalent to

$$\phi_{h,n}(\delta) \leq 4\mathbb{E}_{\bar{P}_h^*, \epsilon} \left[\sup_{f \in \mathcal{F}_h(\delta)} \left(\frac{1}{n} \sum_{i=1}^n \epsilon_{i1} (\ell[f] - \ell[f_{\mathcal{F}_h}^*])(X_i, \bar{Y}_i', \bar{Z}_i) \right) \right]. \quad (18)$$

For any $f \in \mathcal{F}_h$, $\ell[f](X_i, \bar{Y}_i', \bar{Z}_i)$ can be seen as the value of the averaging function $u \mapsto h(u) = M^{-1} \sum_{m=1}^M u_m$ at $\bar{\ell}[f](X_i, \bar{Y}_i', \bar{Z}_i)$ whose m -th component is $(f(Y'_{im}|X_i) - Z_{im})^2$. Corollary 1 in Maurer (2016) provides a version of the contraction inequality (see e.g. Koltchinskii 2011, Theorem 2.2) for this case. The function h is $M^{-1/2}$ -Lipschitz, that is, $|h(u) - h(u')| \leq M^{-1/2} d_2(u, u')$ for all $u, u' \in \mathbb{R}$ and with d_2 the Euclidean distance on \mathbb{R}^M . Thus, by the corollary, Equation (18) yields

$$\begin{aligned} \phi_{h,n}(\delta) &\leq 4\mathbb{E}_{\bar{P}_h^*, \epsilon} \left[\sup_{f \in \mathcal{F}_h(\delta)} \left(\frac{1}{n} \sum_{i=1}^n \epsilon_{i1} h((\bar{\ell}[f] - \bar{\ell}[f_{\mathcal{F}_h}^*])(X_i, \bar{Y}_i', \bar{Z}_i)) \right) \right] \\ &\leq \frac{4\sqrt{2}}{\sqrt{M}} \mathbb{E}_{\bar{P}_h^*, \epsilon} \left[\sup_{f \in \mathcal{F}_h(\delta)} \left(\frac{1}{n} \sum_{i=1}^n \sum_{m=1}^M \epsilon_{im} \left[(f(Y'_{im}|X_i) - Z_{im})^2 - (f_{\mathcal{F}_h}^*(Y'_{im}|X_i) - Z_{im})^2 \right] \right) \right]. \end{aligned} \quad (19)$$

Note that the $M^{-1/2}$ term here is why our final result does not capture an influence of M . The expression in square brackets in Equation (19) is of the form $(a-x)^2 - (b-x)^2$ and can be expressed in terms of the function $\phi_{b,x}: t \mapsto (t+b-x)^2 - (b-x)^2$ over $[-c/h, c/h]$ if one chooses $t = a - b$. For any $b, x \in [0, c/h]$ the function $\phi_{b,x}$ satisfies $\phi_{b,x}(0) = 0$ and

$$|\phi_{b,x}(u) - \phi_{b,x}(v)| = |(u-v)[(u-(x-b)) + (v-(x-b))]| \leq L|u-v|$$

with $L = 4c/h$, since $u, v \in [-c/h, c/h]$. It follows that the function $u \mapsto \phi_{b,x}(u)h/(4c)$ is a contraction. Thus, by the aforementioned contraction inequality, Equation (19) is itself smaller than

$$\frac{32\sqrt{2}}{\sqrt{M}} \frac{c}{h} \mathbb{E}_{\bar{P}_h^*, \epsilon} \left[\sup_{f \in \mathcal{F}_h(\delta)} \left(\frac{1}{n} \sum_{i=1}^n \sum_{m=1}^M \epsilon_{im} [f(Y'_{im}|X_i) - f_{\mathcal{F}_h}^*(Y'_{im}|X_i)] \right) \right] = \frac{32c\sqrt{2M}}{h} \mathbb{E}_{\bar{P}_h^*, \epsilon} \|R_{h,n}\|_{\mathcal{G}(\delta)}.$$

It follows that

$$\phi_{h,n}^\flat(\delta) \leq \frac{32c\sqrt{2M}}{h} \sup_{\sigma \geq \delta} \frac{\mathbb{E}_{\bar{P}_h^*, \epsilon} \|R_{h,n}\|_{\mathcal{G}(\sigma)}}{\sigma}. \quad (20)$$

This puts us in a position where we can upper bound $\phi_{h,n}^\sharp$ by upper bounding the Rademacher sum for a given complexity of \mathcal{F}_h .

B.4 Bounding the excess risk with respect to the complexity of \mathcal{F}_h .

On the basis of the covering number assumption from Equation (4), we adapt Koltchinskii (2011, Theorem 3.12) to our case in order to obtain a simpler bound on the excess risk.

In the previous Appendix B.3 we rely on the laws \bar{P}_h^* and $\bar{P}_{h,n}$ on $\mathcal{X} \times [0, 1]^M \times \mathbb{R}_+^M$. The functions in $\mathcal{G}(\delta)$, like those in $\mathcal{F}_h(\delta)$ are defined on $\mathcal{X} \times [0, 1]$. We therefore introduce Q_{Mn} , the empirical law that puts mass $1/(nM)$ on each (X_i, Y'_{im}) . The law Q_{Mn} is a marginal law over $\mathcal{X} \times [0, 1]$ derived from $\bar{P}_{h,n}$, just like how Q^* is derived from \bar{P}_h^* . In view of Equation (4) we then capture the richness (see Koltchinskii 2011, Section 3.4) of the class $\mathcal{G}(\delta)$ as

$$\sigma_n^2 := \sup_{g \in \mathcal{G}(\delta)} Q_{Mn} g^2. \quad (21)$$

The covering number condition in Equation (4) also holds for $\mathcal{G}(\delta)$, which is also uniformly absolutely bounded by c/h for any δ . We then have by the version in Koltchinskii (2011, Theorem 3.11) of Dudley's entropy integral (see Dudley 2014) that

$$\begin{aligned} \mathbb{E}_{\bar{P}_h^*, \epsilon} \|R_{h,n}\|_{\mathcal{G}(\delta)} &\lesssim \frac{1}{\sqrt{nM}} \mathbb{E}_{\bar{P}_h^*} \int_0^{2\sigma_n} \sqrt{\log N(\epsilon, \mathcal{F}_h, L^2(\bar{P}_{h,n}))} d\epsilon \\ &= \frac{1}{\sqrt{nM}} \mathbb{E}_{\bar{P}_h^*} \int_0^{2\sigma_n} \sqrt{\log N(\epsilon, \mathcal{F}_h, L^2(Q_{Mn}))} d\epsilon \\ &\lesssim \frac{1}{\sqrt{nM}} \mathbb{E}_{\bar{P}_h^*} \int_0^{2\sigma_n} \left(\frac{\|F_h\|_{L^2(Q_{Mn})}}{\epsilon} \right)^\rho d\epsilon, \end{aligned} \quad (22)$$

where the last step uses Equation (4). We aim to upper bound the integral in Equation (22). By change of variable with $u := \epsilon / \|F_h\|_{L^2(Q_{Mn})}$,

$$\int_0^{2\sigma_n} \left(\frac{\|F_h\|_{L^2(Q_{Mn})}}{\epsilon} \right)^\rho d\epsilon = \|F_h\|_{L^2(Q_{Mn})} \int_0^{2\sigma_n / \|F_h\|_{L^2(Q_{Mn})}} u^{-\rho} du.$$

Importantly we have that for all $f \in \mathcal{F}_h$ and any law Q on $\mathcal{X} \times [0, 1]$, it holds that $Qf^2 \leq QF^2$, hence we can upper bound the previous equality by

$$\|F_h\|_{L^2(Q_{Mn})} \int_0^2 u^{-\rho} du = \|F_h\|_{L^2(Q_{Mn})} \frac{2^{1-\rho}}{1-\rho}. \quad (23)$$

Without loss of generality, we can assume that $0 \leq F_h \leq c/h$. Therefore,

$$\mathbb{E}_{\bar{P}_h^*, \epsilon} \|R_{h,n}\|_{\mathcal{G}(\delta)} \lesssim \frac{1}{\sqrt{nM}} \frac{c}{h} \frac{2^{1-\rho}}{1-\rho}.$$

B.5 Combining the Parts

Making use of Equation (23) we can upper bound Equation (20) and write

$$\phi_{h,n}^b(\delta) \lesssim \frac{c\sqrt{M}}{h\delta} \frac{1}{\sqrt{nM}} \frac{c}{h} \frac{2^{1-\rho}}{1-\rho} = \frac{c^2}{h^2\delta\sqrt{n}} \frac{2^{1-\rho}}{1-\rho}.$$

Note that here one can see how the effect of M cancels out. This bound is strictly decreasing in δ . Thus, using the definition of $\phi_{h,n}^\sharp(1/8)$ in Equation (14), it suffices to find δ such that

$$\frac{c^2}{h^2\delta\sqrt{n}} \frac{2^{1-\rho}}{1-\rho} = \frac{1}{8}$$

to reveal that

$$\phi_{h,n}^\sharp \lesssim \frac{c^2 2^{1-\rho}}{h^2 \sqrt{n} (1-\rho)}. \quad (24)$$

Using Equation (15) we get our result in Equation (5).

C Approximating the Identity

For a formal definition of the approximate identity introduced in Section 3, first recall that the convolution of two Lebesgue integrable functions g and h in $L^1(\mathbb{R})$ is defined by

$$(g * h)(y) := \int g(y - y')h(y')dy'.$$

Now recall that a collection $\{K_h : h > 0\} \subset L^1(\mathbb{R})$ of real-valued Lebesgue-integrable functions satisfying

- (i) $\forall h > 0, \int K_h(t)dt = 1,$
- (ii) $\sup_{h>0} \int |K_h(t)|dt < \infty,$
- (iii) $\forall \varepsilon > 0, \limsup_{h \rightarrow 0} \int K_h(t) \mathbb{1}_{\{|t| > \varepsilon\}} dt = 0,$

is called an approximate identity. For instance, if K_1 is the standard Gaussian density and $K_h(\cdot) := h^{-1}K_1(\cdot/h)$, then (i) and (ii) are obviously met, as well as (iii) by the dominated convergence theorem. The name ‘approximate identity’ derives from the following property.

Proposition 2. *Let $\{K_h : h > 0\}$ be an approximate identity. If $\varphi \in L^p(\mathbb{R})$ for some finite $p \geq 1$, then $\limsup_{h \rightarrow 0} \int |(K_h * \varphi) - \varphi|^p dt = 0$. Moreover, if φ is bounded and uniformly continuous, then $\limsup_{h \rightarrow 0} \|(K_h * \varphi) - \varphi\|_\infty = 0$.*

Rather than recalling the full proof of Proposition 2, let us give the simpler proof of the following result adapted to our case.

Lemma 2. *If $f^*(\cdot|x) \in L^2([0,1])$ for (almost) every $x \in \mathcal{X}$ and if $\sup_{x \in \mathcal{X}} \|f(\cdot|x)\|_2 < \infty$, then $\limsup_{h \rightarrow 0} \|f_h^*(\cdot) - f^*(\cdot)\|_2 = 0$.*

Proof. Set $h > 0$ and observe that, for (almost all) $(x, y) \in \mathcal{X} \times [0, 1]$,

$$\begin{aligned} f_h^*(y|x) &= \mathbb{E}_{\tilde{P}_h^*}(K_h(Y - Y')|Y' = y, X = x) \\ &= \int_{[0,1]} K_h(\gamma - y) f^*(\gamma|x) d\gamma \\ &= \int K_h(t) f^*(t + y|x) dt, \end{aligned}$$

with the convention that $f^*(\gamma|x) = 0$ for all $\gamma \notin [0, 1]$. Therefore, using $\int K_h(t)dt = 1$ (second equality), the Cauchy-Schwarz inequality (the inequality), and Fubini’s theorem (third equality) yields

$$\begin{aligned} \|f_h^*(\cdot) - f^*(\cdot)\|_2^2 &= \int_{\mathcal{X}} \int_{[0,1]} \left(\int K_h(t) f^*(t + y|x) dt - f^*(y|x) \right)^2 f^*(x) dx dy \\ &= \int_{\mathcal{X}} \int_{[0,1]} \left(\int K_h(t) [f^*(t + y|x) - f^*(y|x)] dt \right)^2 f^*(x) dx dy \\ &\leq \int |K_h(t)| dt \times \int_{\mathcal{X}} \int_{[0,1]} \left(\int |K_h(t)| [f^*(t + y|x) - f^*(y|x)]^2 dt \right) f^*(x) dx dy \\ &= \int |K_h(t)| dt \times \int_{\mathcal{X}} \left(\int |K_h(t)| \times \|\tau_t f^*(\cdot|x) - f^*(\cdot|x)\|_2^2 dt \right) f^*(x) dx, \end{aligned} \quad (25)$$

where $\tau_t f^* : y \mapsto f^*(t + y)$ and $\|\cdot\|_2$ denotes the $L^2(\mathbb{R})$ -norm.

Set arbitrarily $\varepsilon > 0$ and $x \in \mathcal{X}$. Since $f^*(\cdot|x) \in L^2([0, 1])$, the triangle inequality yields

$$\|\tau_t f^*(\cdot|x) - f^*(\cdot|x)\|_2 \leq 2\|f^*(\cdot|x)\|_2 \leq 2 \sup_{x \in \mathcal{X}} \|f^*(\cdot|x)\|_2 < \infty$$

for all $t \in \mathbb{R}$. Moreover, by Tsybakov (2009, Lemma A.2), there exists $t_x > 0$ such that $\|\tau_t f^*(\cdot|x) - f^*(\cdot|x)\|_2^2 \leq \varepsilon$ for all $|t| < t_x$. Then, by the definition of an approximate identity, there exists $h_x > 0$ such that $0 < h \leq h_x$ implies $\int |K_h(t)| \mathbb{1}\{|t| \geq t_x\} dt \leq \varepsilon$. Consequently if $0 < h \leq h_x$, then

$$\begin{aligned} \int |K_h(t)| \times \|\tau_t f^*(\cdot|x) - f^*(\cdot|x)\|_2^2 dt &= \int |K_h(t)| (\mathbb{1}\{|t| \geq t_x\} + \mathbb{1}\{|t| < t_x\}) \times \|\tau_t f^*(\cdot|x) - f^*(\cdot|x)\|_2^2 dt \\ &\leq \left(4\|f^*(\cdot|x)\|_2^2 + \sup_{h>0} \int |K_h(t)| dt \right) \varepsilon. \end{aligned}$$

Therefore, $x \mapsto \left(\int |K_h(t)| \times \|\tau_t f^*(\cdot|x) - f^*(\cdot|x)\|_2^2 dt \right) f^*(x)$ converges pointwise to zero as h goes to zero. Because it is also upper-bounded by an integrable function independent of h (a constant times $x \mapsto f^*(x)$), the dominated convergence theorem guarantees that the RHS integral in Equation (25) goes to zero as h goes to zero, hence the result. \square

D Hyperparameters

This section provides an overview over the implementation and hyperparameters of the *condensité* methods as well as the other methods from the literature. We specify any values set explicitly by us, for the remaining hyperparameters we keep the default values of the various implementations. In the synthetic data (Section 5) and CPS (Section 6.1) experiments, empirical ISE results are computed using a grid of 500 points, for the satellite image data (Section 6.2) we use only 30 points to reduce the memory footprint. Note that any further data splitting within the methods described below are secondary splits of the 80% training data resulting from the primary splits as described in the main text.

D.1 *condensité*

***condensité (NN)*.** For *condensité (NN)* we use a fully connected neural network with 5 hidden linear layers, batch normalization (Ioffe and Szegedy 2015), and sigmoid-weighted linear unit activations (known as *SiLU* or *swish*, they were originally proposed in Hendrycks and Gimpel 2016; and found to perform well empirically in Elfwing et al. 2018; Ramachandran et al. 2017). Other common activation functions such as *ReLU* (Glorot et al. 2011) also work, but we think it is instructive to show how the predictor architecture can be used to shape the inductive bias of our method, in this case in favor of smoothness. We use a batch size of 1024 and 20 neurons per hidden layer. We train using the *Adam* (Kingma and Ba 2015) optimizer with a learning rate of 10^{-3} and weight decay (see Loshchilov and Hutter 2019) of 10^{-4} . We use 80% of the samples for training, and the remaining 20% for cross-validation. For training, our objective is the mean squared error stated in Equation (1). For cross-validation, we evaluate using the ISE as stated in Equation (6). We train for at most 20 epochs and stop the training if the best previous best cross-validation performance has not been exceeded for 5 rounds.

condensité (tree) For *condensité (tree)* we use a **LightGBM**⁶ gradient boosted tree. As with *condensité (NN)*, we choose $h = 0.01$ and $M = 100$ and perform a 80% to 20% train and cross-validation split. We set the minimal number of data points per leaf to 50, the number of bins to 40 (the same as for *condensier* and *LinCDE*), and the number of leaves to 40.

***condensité (CNN)*.** This *condensité* version consists of a CNN encoder and fully connected head, connected by skip connections. The CNN contains three convolutional layers with kernel size 3, batch normalization, and max pooling with kernel size 2. The inputs are images with three channels, where the third channel consists in the value of the Y'_{im} of the data point in question. We then perform global average pooling on the activations of each convolutional layer, concatenate them with one another and again with the Y'_{im} feature, and feed the result to the head. The head itself consists of three hidden layers, the first with 128 neurons, the subsequent ones with 64. For training, we use the same hyperparameters as for *condensité (NN)*, but reduce the batch size to 128 for memory reasons.

⁶<https://pypi.org/project/lightgbm/4.6.0/>

D.2 Other Methods

LinCDE. We mostly retain the default hyperparameters of the *LinCDE* implementation⁷, but change the number of trees from the default of 100 to 200, and the depth parameter for the individual trees from 1 to 3 so the expressivity of the estimator is not limited too severely.

DRF. We use the Python version of *DRF* with the default parameters of the implementation⁸ except for increasing the minimal node size parameter to 50 to prevent overfitting, since we work with large datasets (the default value is 15). *DRF* returns an empirical probability mass function, so we apply Gaussian smoothing to obtain a density. We choose the bandwidth using Silverman’s rule of thumb (see e.g. Silverman 2018).

FlexCode (tree). This version of *FlexCode* uses an XGBoost⁹ regressor (Chen and Guestrin 2016) and a cosine basis system with 31 bases.

FlexCode (CNN). This version of *FlexCode* uses the same combination of CNN and fully connected head as *condensité (CNN)* as regressor for the coefficients of a cosine basis system with 31 bases, with the only difference being that here there are no Y'_{im} features, and hence there is no third input channel. We use the same optimizer and training parameters as for *condensité (CNN)*.

condensier. We choose the equal length binning method referred to as default in the documentation of the implementation¹⁰. We increase the number of bins to 40 (the default is 20) to account for the complex nature of our datasets, since equal width bins put a hard limit on the expressivity of the method (40 bins is also the default value used by *LinCDE*).

Feature extractor. In the image application presented in Section 6.2, we use a feature extractor to transfer the two-channel 100×100 image patches into 176 features for *condensité (tree)*, *LinCDE*, *DRF*, and *condensier*. The basis of the feature extractor is the same neural network also used for *FlexCode (CNN)*. Within each of the aforementioned methods, it is trained to predict AGB labels on the fraction of the training data used for fitting the model. The features extracted for a given input image are the global average pooled activations of the convolutional layers which serve as input to the fully connected head.

E Proof of Concept and Hyperparameter Analysis (Additional Figures)

This section provides additional figures and details complementing Section 5 of the main text.

E.1 Synthetic Data – Landmark Analysis

This landmark analysis augments the ISE comparison presented in Section 5.1. We inspect the conditional densities fitted by the different methods at three different landmarks for each mechanism. For the single relevant covariate setting we vary only the relevant covariate. For the data on manifold setting we vary the angle between the first two covariates. For the non-sparse data setting, we set all covariates to a specific value. The landmarks are stated in Table 7. All unspecified covariates are fixed at zero.

⁷<https://github.com/ZijunGao/LinCDE>

⁸<https://github.com/lorismichel/drf>

⁹<https://pypi.org/project/xgboost/3.0.2/>

¹⁰<https://github.com/osofr/condensier>

Setting	Landmark 1	Landmark 2	Landmark 3
Single relevant covariate	$x^{(1)} = 0$	$x^{(1)} = 0.5$	$x^{(1)} = 1$
Data on manifold	$x^{(1)} = \cos \frac{2\pi}{6}$	$x^{(1)} = \cos \pi$	$x^{(1)} = \cos \frac{10\pi}{6}$
	$x^{(2)} = \sin \frac{2\pi}{6}$	$x^{(2)} = \sin \pi$	$x^{(2)} = \sin \frac{10\pi}{6}$
Non-sparse data	$x = (-1, \dots, -1)$	$x = (0, \dots, 0)$	$x = (1, \dots, 1)$

Table 7: Landmark values for each data-generating process to visualize the fitted estimators.

Figure 7 shows the conditional densities estimates for each method and landmark, and a histogram of 1000 samples from the true density for comparison. Overall, we observe good performance for both *condensité* methods. The data on manifold and non-sparse data settings prove somewhat challenging for *condensité (tree)*, whereas *condensité (NN)* gives a close fit everywhere. *LinCDE* performs well in most settings but struggles somewhat with the magnitude of the shifts along the horizontal axis for the non-sparse data. In this respect it is similar to *condensité (tree)*, although it provides a smoother fit. Similarly to the *condensité* methods, *DRF* provides an excellent visual fit on all single relevant covariate landmarks, but underestimates the magnitude of the shifts in some of the other settings. *FlexCode* fails to adapt to the heteroskedasticity of the single relevant covariate landmarks, and struggles in the data on manifold setting. Its estimates also exhibit noticeable spurious peaks in multiple instances. A larger number of bases might give it greater flexibility to adapt to concentrated densities, but would likely also incur even more severe spurious peaks. Lastly, *condensier* consistently estimates the location of the mode well, although it under-smooths rather severely in the non-sparse data setting. In summary, the examples confirm that both versions of *condensité*, as well as the other methods from the literature, capture the essential tendencies of the conditional distributions in our synthetic data examples correctly.

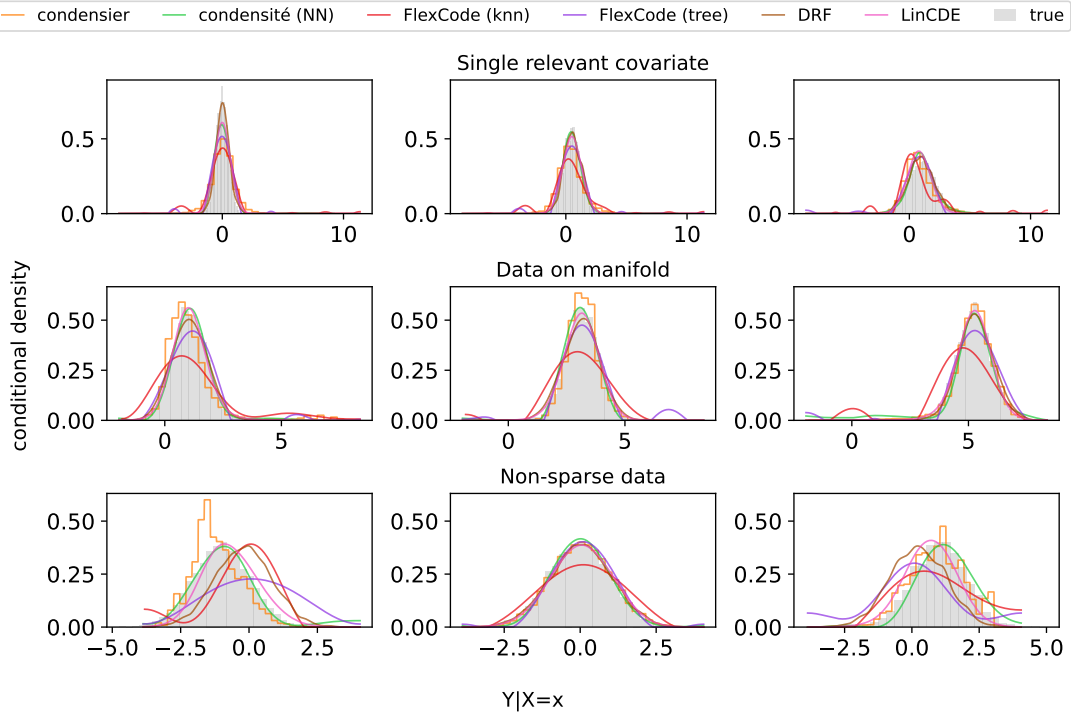


Figure 7: Comparison of the estimates to a sample from the true conditional density on a grid of points.

E.2 The Effect of M and Its Interaction With h – *condensité (tree)*

Figure 8 below shows the experiment exploring the interplay of M and h for *condensité (tree)*. It complements Figure 3 in Section 5.2 of the main text.

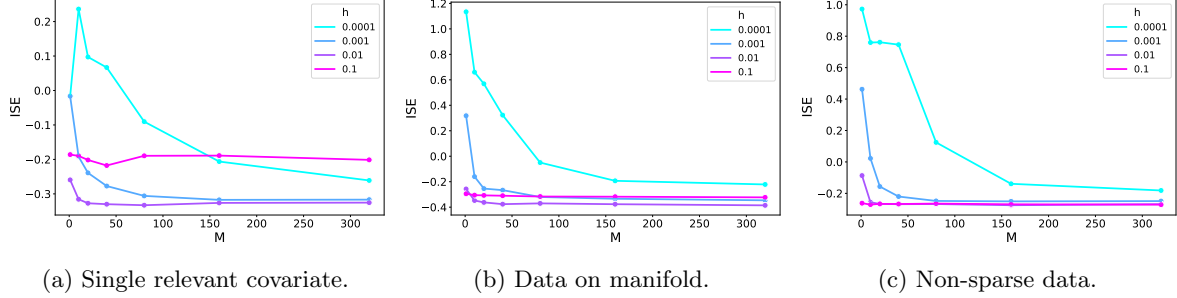


Figure 8: *condensité (tree)*: ISE for different M and h on the synthetic data-generating mechanisms.

F The IPUMS-CPS Dataset (Preprocessing and Additional Figures)

This section contains details on our compilation of the CPS dataset, as well as a landmark analysis for the other methods from the literature.

F.1 Dataset and Preprocessing

For the evaluation of our methods on real-world data in Section 6 of the main text, we use CPS records from 2024 for which the Annual Social and Economic Supplements (ASEC) are available, meaning from March 2024. We create a dataset based on the variable selection outlined in Table 8.

CBSASZ	(population size of household location area)
AGE	(age at last birthday)
SEX	(sex)
NCHILD	(number of own children in household)
YRIMMIG	(year of immigration into the US)
NATIVITY	(foreign birthplace or parentage)
LABFORCE	(labor force status)
UHRSWORKT	(hours usually worked per week at all jobs)
RACE	(race)
EMPSTAT	(employment status)
CLASSWKR	(class of worker)
EDUC	(educational attainment)
INCTOT	(total personal income)

Table 8: IPUMS-CPS variable selection.

We perform the following preprocessing steps:

1. keep only observations that are part of the ASEC,
2. drop all observations with
 - INCTOT negative, unknown, or in excess of 300000\$,
 - unknown or below secondary school education,

- unknown LABFORCE or NATIVITY,
3. re-code SEX and LABFORCE to $\{0, 1\}$ dummy variables,
 4. one-hot encode RACE and EMPSTAT by the first digit,
 5. one-hot encode CLASSWKR,
 6. map EDUC to years of education,
 7. create separate dummy for UHRSWORKT code 997 (varying hours), and set values 997 and 999 (not in universe) to 0.

This leaves us with a total of 113104 observations of 26 covariates (not including INCTOT). The variables AGE, NCHILD, YRIMMIG, NATIVITY, UHRSWORKT, and EDUC_YEARS (derived from EDUC) are multi-valued, with the remaining ones being binary. All variables are encoded numerically. Importantly, whenever we draw a sample for learning or evaluation, we resample with replacement weighted by ASECWT (the ASEC weighting factor) to obtain a representative sample.

F.2 Landmark Analysis for Realistic Use-Cases (Other Methods)

Figure 9 and Figure 10 below show the landmark analyses for the other methods from the literature corresponding respectively to Figure 4 and Figure 5 in the main text. *DRF* and *condensier* fit areas of high density well, but seemingly at the cost of overfitting other areas, resulting in high-variance estimates. *LinCDE* provides a smooth fit but does not manage to fit highly concentrated densities. Despite performing worse in ISE than *condensité (CNN)*, *condensité (tree)*, and *FlexCode (tree)*, the other methods from the literature exhibit the same general trends in line with the manually constructed local empirical density.

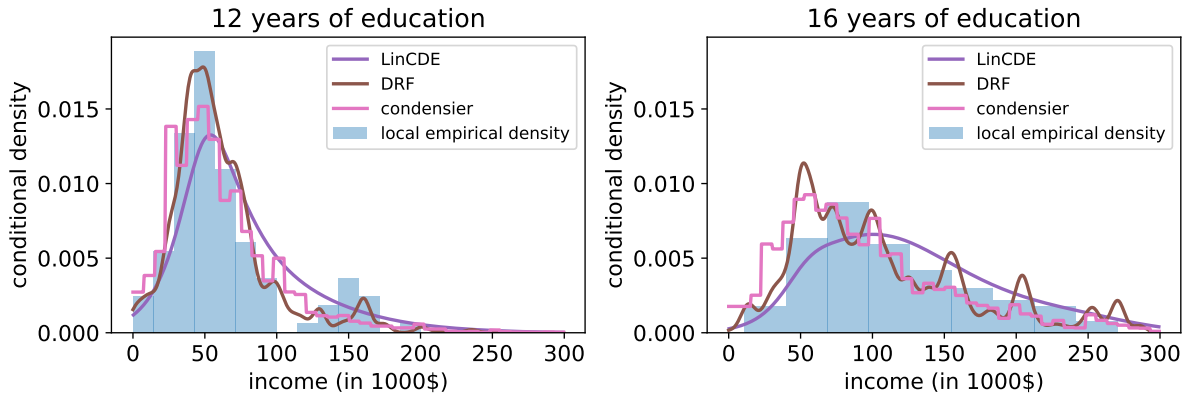


Figure 9: Conditional income density for 12 and 16 years of education with otherwise identical covariates.

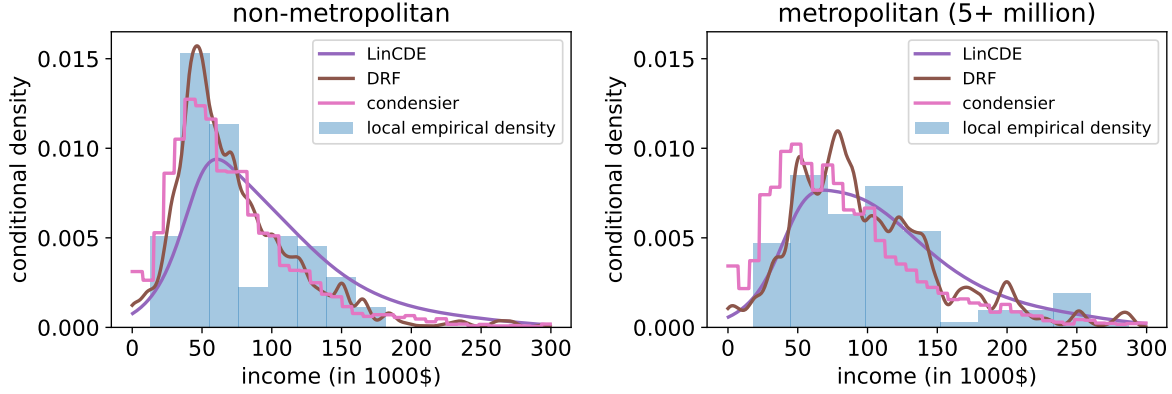


Figure 10: Conditional income density for metropolitan and non-metropolitan inhabitants with otherwise identical covariates.

G ESA ICC Satellite Imaging Data (Preprocessing and Additional Figures)

G.1 Dataset and Preprocessing

The following are details on the dataset and additional results for the AGB satellite image application presented in Section 6.2.

Label data. AGB is a measure of biomass density, defined as dry weight of live wood per unit area. We derive our labels from the biomass dataset provided by the ESA Biomass Climate Change Initiative (Santoro and Cartus 2025). This dataset provides labels computed from satellite images of different wavelengths by a theoretically motivated and empirically calibrated algorithm. Each label corresponds to a specific geographic area of 100×100 meters. The AGB measurements provided are given in megagrams (tons) per hectare and constitute yearly averages. Our experiment sets out to predict biomass measurements from satellite images of the type used for the creation of the Santoro and Cartus (2025) dataset, essentially expanding upon the role of the theoretical model used in the original study, which provides mean estimates, not conditional densities. To this end we choose the two rectangular coordinate regions given in Table 9, one for training, and the other for testing out-of-sample.

	north-west	north-east	south-east	south-west
training region	$[132.5, -14.65]$	$[133.75, -14.65]$	$[133.75, -15.65]$	$[132.5, -15.65]$
test region	$[132.75, -13.25]$	$[134.0, -13.25]$	$[134.0, -14.25]$	$[132.75, -14.25]$

Table 9: Coordinate rectangle.

Both of these are in the Australian Northern Territory. The training patch covers an area to the south-east of Kakadu national park, and the test patch an area to the south-east of the town Katherine. The location is chosen for its low biomass, which can be captured by C-band imaging, and little seasonal change in AGB.

Feature data. To obtain features for each label patch, we retrieve a ground range detected Sentinel-1A satellite image taken in interferometric wide swath mode from the GEODES portal (CNES 2025). We choose the images in Table 10, which contain the corresponding patches described in Table 9. They have been taken on August 15th 2020, that is during the dry season where foliage is low and C-band

measurements are most informative. We ensure cloud cover is less than 5%. We work with an orthorectified version of the images as can be obtained through the GEODES portal, and use the *VV* and *VH* bands. The resolution of the images is 10×10 meters, their IDs are given in Table 10.

training	S1A_IW_GRDH_1SDV_20200815T204041_20200815T204106_033922_03EF62_FF6E
test	S1A_IW_GRDH_1SDV_20200815T204106_20200815T204131_033922_03EF62_BB0B

Table 10: IDs of the Sentinel-1A satellite images used to construct our features.

The images are shown in Figure 11 and Figure 12 for illustration; note that our preprocessing is minimal and not aimed at visualization.

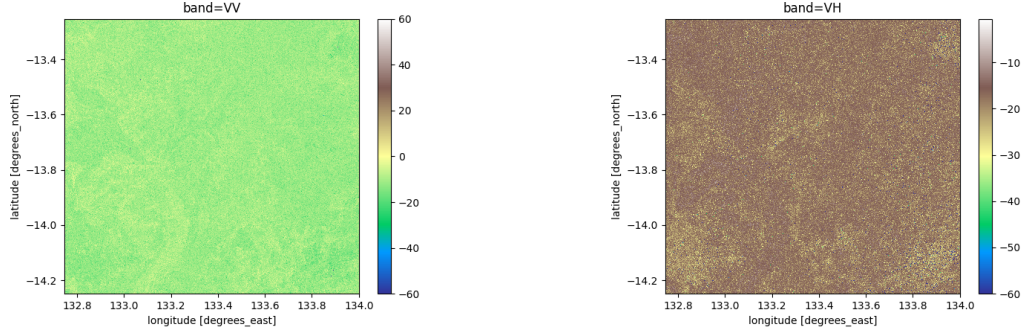


Figure 11: **Training region.** Images (in decibel) used as the basis for our feature patches.

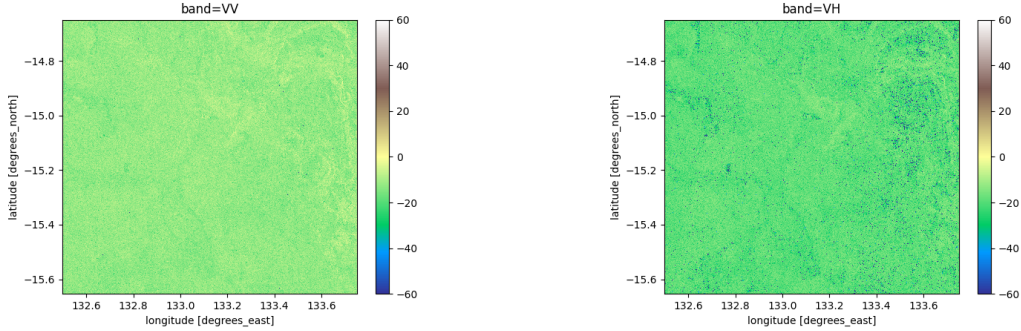


Figure 12: **Test region.** Images (in decibel) used as the basis for feature patches.

Compiling our dataset. To compile our dataset, we perform the following steps:

1. fuse AGB labels into square kilometer patches by taking the average,
2. log-transform target AGB values by $x \mapsto \log(1 + x)$ to avoid dominance of near-zero values,
3. log-transform Sentinel-1A image pixels by $x \mapsto 10 \log_{10}(x)$ to obtain values in decibel,
4. identify the corresponding 100×100 pixel satellite image patch for each fused AGB label by reprojecting into EPSG:3577 (drop all patches with missing data).

After the first step, each AGB label corresponds to a one square kilometer patch, which makes for a 100×100 pixel patch in the satellite images. This results in 14653 observations for the training region, and 14683 observations for the test region. Each observation is a 100×100 image with two channels, *VV* and *VH*.

G.2 Visualization and Qualitative Assessment (Training Region)

Figure 13 below contains the training region counterparts to the test region plots in Figure 6.

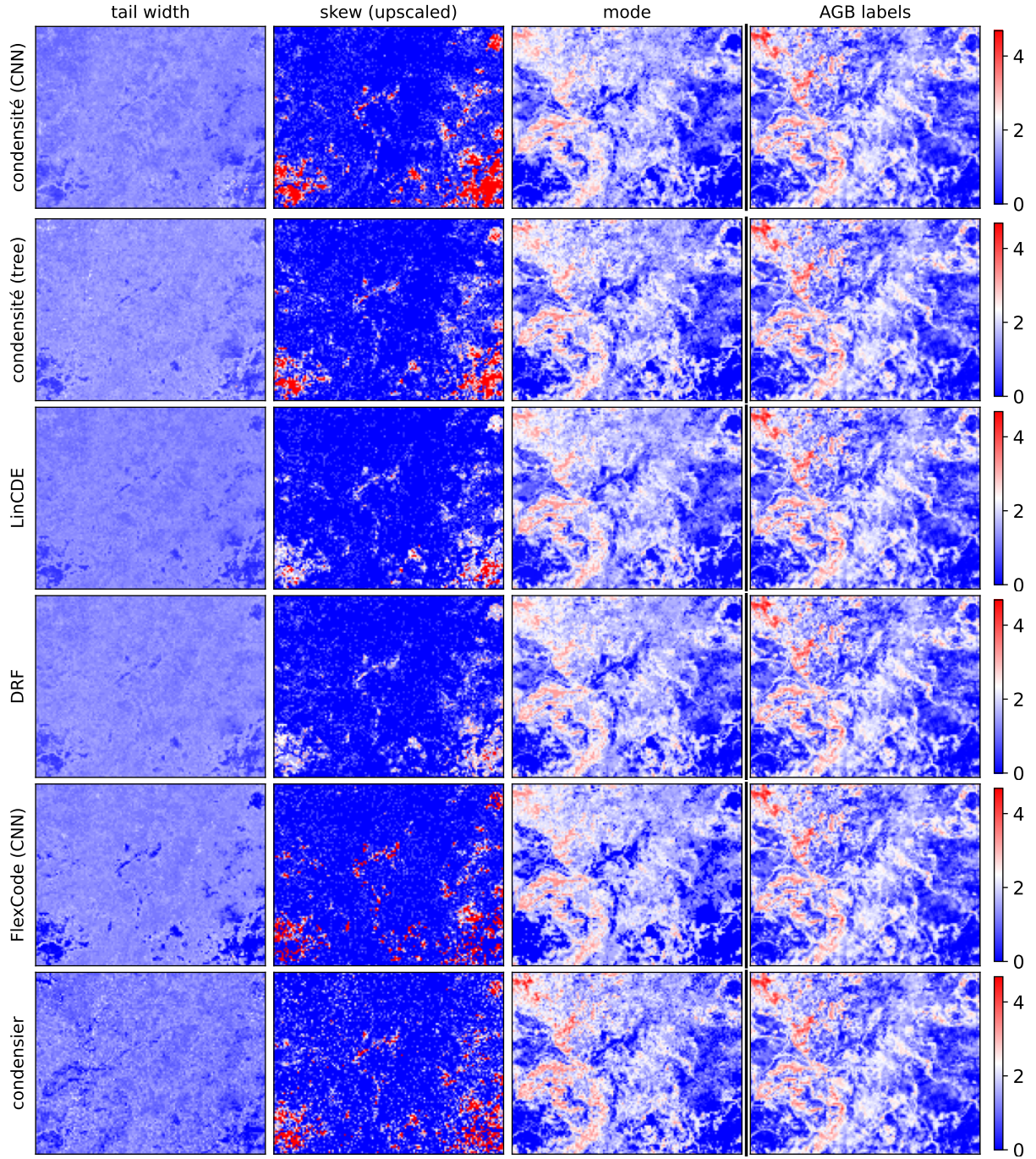


Figure 13: **Training region.** Visualization of method estimate summary statistics and labels.