International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Review Article)

Predicting agricultural waste generation in the U.S.: A Data-Driven guide for smarter resource use

Adekunbi Bello [1,*], Samuel Fanijo [2] and Isaiah Alfred [3]

[1] *Department of Business Administration F.W Olin School of Business, Babson College, Wellesley MA, USA.*

[2] *Department of Computer science, Iowa State University, Ames IA, USA.*

[3] *Independent Researcher, Lagos Nigeria.*

International Journal of Science and Research Archive, 2025, 16(01), 1783-1799

Publication history: Received on 01 June 2025; revised on 18 July 2025; accepted on 21 July 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.1.2076

Abstract

U.S. crop residue is a major underutilized bioenergy resource, yet current national models often rely on outdated data or broad assumptions, limiting localized planning for renewable energy and soil conservation. This study uses machine learning to quantify and classify county-level agricultural residue generation across the U.S. An XGBoost regression model achieved $R^2 \approx 0.69$, showing that cropland acreage, USDA-estimated residue supply, local biorefinery potential, and regional clustering together explain much of the variation in predicted biopower potential. A Random Forest binary classifier reached ~83% balanced accuracy in identifying high- versus low-residue counties, while multiclass classification of residue behavior clusters achieved ~57% accuracy, with strongest performance in moderate-output, small-acreage areas. These findings highlight clear spatial patterns: high-residue counties are concentrated in the Midwest corn belt, while other regions display mixed or lower-output profiles. Corn was the primary contributor to residue volume, far surpassing crops like wheat and soybeans. The resulting reuse readiness maps provide actionable insights for farmers, policymakers, and renewable energy planners, offering a scalable tool to guide sustainable bioenergy investment while supporting long-term soil health and circular economy goals.

Keywords: Agricultural residue; Biopower potential; Machine learning; EPA repowering data; Xgboost; Circular Bioeconomy

1. Introduction

The transition to sustainable energy has intensified interest in biomass, with modern bioenergy (e.g. power and fuels from organic residues) already supplying a significant share of global renewables (Syeda Nyma Ferdous, 2023).According to the IEA, modern bioenergy (excluding traditional biomass) accounts for roughly 55% of global renewable energy (about 6% of total energy supply) and is growing ~4% per year (IEA, 2025). To meet net-zero climate goals, deployment must accelerate (projected 8% annual growth through 2030) (IEA, 2025). In parallel, circular-economy principles emphasize recapturing waste as a resource (J Aravind Kumar, 2023). For agriculture, this means recycling crop residues and organic by-products to reduce greenhouse gases and conserve resources. Major assessments indicate vast untapped biomass potential: the U.S. Bioenergy Technologies Office estimates the country could sustainably mobilize on the order of 1 billion dry tons of biomass annually under mature market conditions (Bioenergy, 2024), of which hundreds of millions of tons are agricultural residues. For example, corn stover alone has an estimated unused potential of ~175 million tons/year (Bioenergy, 2024). Reusing these residues for electricity or biogas can displace fossil energy and align with several SDGs (clean energy, climate action, responsible production) (J Aravind Kumar, 2023).

* Corresponding author: Adekunbi Bello

At the federal level, U.S. policy has historically supported agricultural bioenergy. In recent years, USDA initiatives have promoted blending infrastructure for ethanol and biodiesel made from homegrown crops (USDA P. , 2025), and EPA has worked to set renewable fuel blending volumes (RFS) that influence feedstock demand. Under the Trump administration, USDA released funding for biofuel infrastructure (e.g. pumps for E15/E85 ethanol blends) to “increase domestic, homegrown fuels” and create rural jobs (USDA P. , 2025). EPA’s 2025 agenda under the same administration emphasized clearing a backlog of small-refinery waivers in the RFS and setting higher blending targets for advanced biofuels (Stephanie Kelly, 2025). Meanwhile, the EPA’s RE-Powering America’s Land initiative curates a nationwide database of approximately190,000 brownfields, landfills, and mines screened for renewable energy potential (EPA, 2022), including layers for biomass feedstock. These policies and programs underscore the strategic role of farm waste in energy security and rural economic resilience, yet they also highlight a need for detailed maps of where and how biomass can be repurposed.

Despite the abundant data, few studies have mapped agricultural waste flows and energy potential at the county level using modern analytics. Past work has clustered U.S. counties by crop production profiles (Courtney R. Hammond Wagner, 2019) or modeled sustainable residue removal rates for specific locations (Syeda Nyma Ferdous, 2023), but a gap remains in predicting biopower capacity from recent national data. This paper addresses that gap by combining real 2022 datasets with machine learning to predict and classify biopower potential in U.S. counties. We use clustering to reveal regional patterns of crop residue generation and XGBoost models to estimate electricity capacity from these wastes, thereby identifying “reuse readiness” for agricultural residues across the rural landscape.

Agricultural waste streams include field residues (stems, stalks, leaves), processing byproducts (cull produce, hulls), animal manures, and other organic discards. These residuals contain stored solar energy and nutrients, and their scale is immense. The DOE’s Billion-Ton Reports indicate that U.S. agricultural residues have significant potential as a bioenergy resource, with estimates of hundreds of millions of dry tons available annually under various scenarios. For example, one DOE scenario projects *~175 million tons* per year of collectable agricultural residues (primarily corn, soybean, and small-grain straw) in a mature market (Energy, 2023). Such quantities could theoretically generate many gigawatts of renewable power if harnessed efficiently. Currently only a fraction is used: U.S. bioenergy (including ethanol, cellulosic fuels, biomass power) provides about 5% of domestic energy (Bioenergy, 2024), while large residue stocks remain unexploited.

Environmental considerations make residue reuse attractive. Leaving excess crop residue on fields can harm soil and air quality (fire, methane), whereas controlled removal for energy (e.g. anaerobic digestion, combustion) can reduce greenhouse gas emissions and replace coal or gas. The Food and agriculture Organization of the UN explicitly advocates a “circular agriculture economy” that minimizes waste and closes nutrient loops by recycling agricultural byproducts back into production or energy use (UN, 2025). Similarly, Kumar et al. note that global energy policy is shifting toward systems where waste biomass is economically valorized to meet energy needs while limiting ecological damage (J Aravind Kumar, 2023). Using residues for power aligns with multiple UN Sustainable Development Goals (SDG7 for clean energy, SDG12 for sustainable production, SDG13 for climate (UN, 2025).

On the economic side, converting waste to energy can generate rural income and jobs. USDA reports emphasize that biofuel and biopower projects using American-grown feedstocks broaden the market for farmers, diversify the nation’s energy supply, and create good-paying jobs in rural areas (USDA P. , 2025). For instance, USDA’s Higher Blends Infrastructure Incentive Program (HBIIP), established during the Trump era, expanded pumps and storage for mid-level ethanol blends (E15/E85) and biodiesel (B20) made from corn and soy (USDA P. , 2025). Such programs were explicitly designed to *“increase domestic, homegrown fuels”* and restore rural prosperity (USDA P. , 2025). Likewise, crop residues can be co-fired at biorefineries or used onsite in farm-based digesters, adding revenue streams for growers. The interplay of energy and agriculture policy thus supports residue reuse but requires spatial tools to target where investments yield the greatest benefits.

Agricultural fields in the United States produce huge amounts of leftover plant material (crop “waste” or residues) every year, on the order of half a billion tons (Andrews, 2006). Left in the field, these residues protect soil from erosion, conserve moisture, and build organic matter (Epa.gov, n.d.). But crop residues can also be collected as a renewable energy source (for example, in cellulosic biofuel or biopower plants). This study used real 2022 government data (USDA crop statistics and EPA resource data) and modern data analytics to understand where these residues are generated and how much could be available for energy without harming the soil.

First, we grouped U.S. counties and crops into clusters with similar residue production patterns. We found that Midwestern states, which are known for high corn production, formed a “high residue” cluster, reflecting corn’s well-established role as a major residue source. Other clusters correspond to regions growing more wheat, soy, or mixed crops, which produce moderate or lower residues. Then we trained a machine-learning model (XGBoost) to predict residue quantity and energy potential using variables such as county-level residue estimates, farm size (acreage), and biorefinery residue data. The model performed well in tests, suggesting it can reliably estimate waste generation across different regions.

The key findings are that residue generation varies considerably across U.S. regions, dominated by corn in the Midwest, and that data-driven models can accurately forecast it. For farmers and land managers, these insights can guide early decisions about which fields may be suitable for residue collection based on overall supply. However, site-specific assessments of soil health and conservation needs would still be required to determine safe removal rates. For example, our results support using conservation guidelines that recommend retaining some fraction of residues on every field (as NRCS practice standards do) (Andrews, 2006). In practice, planners can use our maps and model to identify counties or areas with high biomass availability for bioenergy projects, while ensuring enough residue remains (following USDA/NRCS site-specific advice) to prevent erosion and maintain soil fertility. Essentially, providing a practical tool and knowledge base to balance renewable energy goals with sustainable soil management.

2. Literature Review

Recent research demonstrates the growing role of machine learning in analyzing biomass systems and supporting sustainable resource use. (Syeda Nyma Ferdous, 2023) developed a deep ensemble ML framework that predicts sustainable crop residue removal while accounting for soil quality and environmental trade-offs. Their model achieved high predictive accuracy in evaluating safe harvest thresholds, enabling conservation-aware residue utilization planning. In a broader scope, (Hyojin Lee, 2024) conducted a comprehensive review of ML techniques used in predicting bio-oil yields from lignocellulosic biomass pyrolysis, showing how tree-based models like Random Forest and XGBoost enhance prediction of energy content and feedstock conversion efficiency. These examples reflect the power of ML, particularly ensemble algorithms, in capturing nonlinear relationships across complex agricultural and environmental datasets. While the review by (Hyojin Lee, 2024) emphasizes downstream conversion efficiency, (Syeda Nyma Ferdous, 2023) focuses on upstream biomass availability and soil health, together showcasing the full potential of ML across the biomass supply chain.

Geospatial clustering has also been applied to agricultural classification. For instance, Hammond (Courtney R. Hammond Wagner, 2019) used k-means clustering on acreage data from 21 USDA crops to generate county-level typologies across the U.S. Their study revealed agroecological zones like the Corn Belt and Rice Belt, offering a spatial lens on crop specialization and supply. While their method aimed at general crop pattern analysis, our work focuses more specifically on residue potential, leveraging clustering to identify groups of counties with similar waste-generation profiles. Related studies like the DOE’s Billion-Ton assessments (Energy, 2023) and satellite-enabled crop residue estimations have mapped regional biomass supply. These datasets are useful for understanding overall agricultural waste availability but often lack the machine learning, driven predictions needed to support localized residue planning and reuse.

The role of bioeconomy in sustainability transitions has also been emphasized. Global policy literature from the FAO and IEA highlights that repurposing agricultural waste is key to achieving sustainable development goals, particularly in the post-COVID era (UN, 2025) (IEA, 2025). These reports frame circular bioeconomy (CBE) as a model for reducing waste and creating value from residue streams. Within the U.S. context, federal documents and CRS briefs (Bracmort, 2015) have defined biopower as energy generated from organic waste and biomass, emphasizing government incentives such as renewable energy tax credits and infrastructure support. Our work complements these discussions by providing a data-driven view of where residues are generated and how they could contribute to decentralized clean energy networks under such policy frameworks.

Several studies illustrate how remote sensing and ensemble ML can be paired for residue mapping. (Shoaib Ahmad Anees, 2014) integrated Landsat-9 satellite data with elevation features and trained Random Forest and XGBoost models to estimate forest aboveground biomass in Himalayan regions. The resulting model yielded strong accuracy (R^2 up to 0.81), validating the combination of satellite imaging and ML for spatial biomass prediction. In a related application, (Nan Lin, 2024) estimated maize residue cover using UAV and Sentinel-2 imagery. They applied an adaptive threshold segmentation method and used CatBoost regression on selected vegetation indices to predict stubble coverage with $R^2 \approx 0.83$. These studies confirm that ensemble ML, when combined with remote sensing, offers strong predictive capabilities, but they often remain context-specific, such as forest ecosystems or maize plots. In contrast, our study applies machine learning at national scale with publicly available structured crop production data, targeting broader regional forecasting rather than site-specific remote sensing.

On the spatial modeling side, GIS-based workflows have been widely used to map crop residue potential. (Avinash Bharti, 2021) reviewed geospatial methods for biomass and bioenergy potential estimation, identifying best practices in using soil, land use, and topographic layers to map available feedstock. (Abhishek Chakraborty, 2022) developed a high-resolution (1-km) mapping framework for India by combining agricultural production statistics with satellite-derived crop-class maps. Their geoportal "BHUVAN-JAIVOORJA" visualizes biomass supply at district level, helping planners identify optimal regions for biomass-based plant siting. Similarly, (Abel Rodrigues, 2025) studied Portugal's Alentejo region by modeling calorific potential from straw residues and validating crop-area models with field surveys. Their analysis found that using just half the available straw could supply ~12.5% of regional energy demand. These studies emphasize how GIS mapping, when combined with empirical or modeled data, can support decision-making for localized energy production, although many models assume fixed residue ratios and static land use.

Geospatial clustering tools are increasingly applied to agricultural monitoring. (Simone Pascucci, 2018) used functional PCA-based clustering on Landsat time-series to segment agricultural fields into consistent intra-field "yield zones." Their results show that functional clustering offers more stable patterns over time than standard k-means, supporting precision agriculture applications. (Akram Zaytar, 2024)introduced a no-code platform "Sims" built on Google Earth Engine to facilitate clustering and spatial similarity analysis across raster layers. In simulations using maize yield data from Rwanda, Sims generated clear cluster maps and supported interactive analysis. These tools demonstrate the value of clustering not just for classification, but also for practical site-level interpretation of agricultural performance.

A growing body of work connects agricultural residue management with broader circular bioeconomy goals. (Haruna Sekabira, 2024) conducted household surveys in African city-regions to assess how circular practices like composting and livestock feeding of crop residues influence food security. Their regression analysis showed that integrating organic waste reuse into local systems improved soil quality and household income, enhancing both environmental and social resilience. (Bhim Singh, 2024) developed predictive models for crop residue generation in India through 2030, showing the relevance of machine learning in supporting long-term residue forecasting for energy and waste planning. (Stamatia Skoutida, 2024) evaluated Greece's agricultural biomass supply and concluded that residues could offset nearly 8% of the country's energy needs. These findings emphasize the importance of predictive tools not just for energy resource mapping, but also for improving sustainability outcomes across sectors, from climate mitigation to food security.

Table 1 Summary of Key Literature Reviewed on Agricultural Residue Mapping and Bioenergy Potential

Literature Title (linked)	Author(s)	Method Description	Strength	Weakness
Integration of machine learning and remote sensing for above ground biomass estimation through Landsat-9 and field data in temperate forests of the Himalayan region	Shoaib A. Anees et al. (2024)	Landsat-9 OLI bands and DEM features used with Random Forest and XGBoost regression to model forest AGB.	High accuracy (R^2≈0.81 RF) by combining multispectral and terrain data; open-source data.	Focused on one forest biome; results depend on quality of field samples.
Estimation of Maize Residue Cover Using Remote Sensing Based on Adaptive Threshold Segmentation and CatBoost Algorithm	Nan Lin et al. (2024)	UAV images + Sentinel-2 spectral indices; Yen threshold segmentation for cover mask + CatBoost regression for coverage fraction.	Integrates UAV and multispectral data; achieved R^2≈0.83, improving residue detection.	Case study scale (one region); transferability to other regions/crops is unclear.
GIS Application for the Estimation of Bioenergy Potential from Agriculture Residues: An Overview	Avinash Bharti et al. (2021)	Review of GIS-based methods (land use maps, remote sensing, databases) to quantify agricultural residue potentials.	Comprehensive overview highlighting the role of spatial analysis in residue resource assessment.	Review only (no new analysis); broad statements but limited quantitative detail.

Developing a spatial information system of biomass potential from crop residues over India	Abhishek Chakraborty et al. (2022)	Hybrid GIS approach: district crop statistics + surveys to get residue; AWiFS land-cover + MODIS GPP to downscale to 1 km; built web geoportal.	Produces high-resolution maps of residue and bioenergy potential; identified key regions (e.g. Punjab).	Depends on fixed residue/surplus assumptions; requires periodic ground surveys to update.
A GIS-Based Estimation of Bioenergy Potential from Cereal and Legume Straw Biomasses in Alentejo, Portugal	António Cordeiro & Paulo Brito (2025)	GIS-based cropping suitability model: assign optimal cereal/legume areas; apply field-validated yields and calorific values to compute straw energy.	Field-validated allocation; quantifies region-specific energy (≈2940 TJ from surplus straw).	Regional scope (Alentejo); methodology depends on assumed yields and straw uses.
A Comparison between Standard and Functional Clustering Methodologies: Application to Agricultural Fields for Yield Pattern Assessment	Simone Pascucci et al. (2018)	Applied k-means, functional k-means, PCA clustering to Landsat time-series from fields; compared clusters to measured yields.	Functional clustering captured more within-field variability and matched yield differences better than simple k-means.	Computational complexity: tested on limited fields in Italy.
Sims: An Interactive Tool for Geospatial Matching and Clustering	Akram Zaytar et al. (2024)	No-code GEE web-app for clustering and similarity search on spatial layers; demonstrated on simulated maize yield data.	User-friendly, open-source tool for spatial feature exploration; flexible clustering of multi-variate maps.	Demonstrated on simulated/small data; real-world utility and scalability need further evaluation.
Circular bioeconomy practices and their associations with household food security in four RUNRES African city regions	Haruna Sekabira et al. (2024)	Survey data and OLS/PSM regression linking circular bioeconomy activities (compost, feed, waste sorting) to food security indicators.	Empirical link of residue-reuse practices to improved food access; highlights socio-economic benefits of residue recycling.	Social/economic focus (not a spatial/ML model); context-specific (African cities) and not directly about biomass mapping.
The Latent Potential of Agricultural Residues in Circular Economy: Quantifying their Production Destined for Prospective Energy Generation Applications	Stamatia Skoutida et al. (2024)	Quantified the energy potential of agricultural residues (e.g., straw, stalks) in Greece using national production statistics.	Estimated that residues could meet ~8% of Greece's energy demand; strong link to circular bioeconomy policy.	Focused on Greece; does not integrate ML or high-resolution spatial models.

2.1. Data Collection and Management

We leveraged two complementary national datasets. First, the USDA's 2022 crop production data (via NASS Quick Stats) provide authoritative county-level statistics on crop yields and acreage across the U.S. (USDA, United States Department of Agriculture, 2022). This supply-side data allows estimation of how much biomass (e.g. corn stover, wheat straw) is potentially available in each region. Second, EPA's RE-Powering Mapper dataset provides site-level information for roughly 190,000 brownfields, landfills and other underused sites screened for renewable energy potential (EPA, 2022). These records include geolocations, acres, and estimated biomass resources nearby (e.g. cumulative crops and woody residues within 50 miles). Merging USDA and EPA data is methodologically robust and evidence-based, as the two datasets address distinct yet complementary aspects of bioenergy potential: USDA data quantify the resource supply, (crop residues), while the EPA data encode **site feasibility** (infrastructure and local biomass constraints). For example, the EPA data even include flags like “Biopower Facility Potential” to mark sites pre-screened as suitable for biopower development (EPA, 2022), illustrating how it encodes renewable energy readiness. By joining crop-residue supply estimates with geospatial site attributes, we create an enriched dataset linking waste availability to likely deployment sites.

A random subset of 5,000 sites was drawn from the ~190,000 EPA records to form the core modeling dataset. Subsampling was used primarily for computational efficiency – training machine learning models on thousands rather than hundreds of thousands of points greatly reduces runtime. It also aids data quality control and balance, since we can manually inspect and clean a smaller sample. Importantly, the 5,000 sites were selected at random, ensuring that the sample remains broadly representative of the full data distribution (in terms of geography, site types, and biomass levels) while still being tractable. In practice, large-scale spatial studies often use random sampling to retain statistical representativeness while managing data volume.

We then constructed features by combining USDA and EPA attributes. It was critical to convert crop production figures into *residue mass*, because crop residues (stalks, husks, etc.) are the actual feedstock for bioenergy. Grain yields alone do not directly translate to usable biomass; instead, each crop has a residue-to-production ratio. For example, corn production yields substantial stover mass in addition to grain. Crop residues are a well-known lignocellulosic resource for fuel production (Xinliang Xu, 2013), so we applied standard residue-to-yield factors to each crop’s production data. In other words, for each crop type we estimated the dry tons of residue generated per unit of grain. This conversion was essential to quantify the true bioenergy feedstock available. We focused on the top 10 commodity crops by production in the U.S. (barley, oats, soybeans, sorghum, corn, rye, rice, cotton, sugar beets, and wheat), which collectively account for the majority of U.S. cropland output. For each of these crops, we calculated the residue mass by applying established residue-to-production ratios (i.e. the fraction of harvested crop that remains as residue) from the literature ((Paul W. Gallagher, 2003); (R. L. Graham, 2007). These ratios vary by crop and reflect typical relationships between yield and leftover plant material. For instance, corn has a ratio of 1.5; indicating that for every ton of corn produced, about 1.5 tons of residue remain, while soybeans have a lower ratio of 1.2. In practice we multiplied each county’s crop yield by the crop-specific residue fraction to estimate biomass residue, consistent with prior studies of bioenergy supply. The ratios for the crops are seen in the table below.

Table 2 Residue Ratio

Crop (Commodity)	Residue Ratio (Residue/Production)
BARLEY	1.3
CORN	1.5
COTTON	2.0
OATS	1.3
RICE	1.5
RYE	1.4
SORGHUM	1.4
SOYBEANS	1.2
SUGARBEETS	1.1
WHEAT	1.3

After merging the USDA and EPA datasets by location, we performed standard data cleaning. We first removed duplicate records and redundant columns (for example, we dropped any crop-residue fields coming from the EPA data to avoid duplication of USDA-based estimates). Missing values for variables such as site acreage were imputed using median values from similar sites (e.g. by cluster). We also examined the distributions of key variables: both acreage and residue quantities were highly right-skewed. To normalize them, we applied a logarithmic transformation to these skewed features before modeling. The resulting cleaned dataset consisted of 1,389 records, each with multiple variables (including residue, acreage, crop type, state, latitude, EPA biopower potential, etc.) and was used for our machine learning analyses.

The final cleaned dataset (aggregated by county or site) included the variables listed in Table 1. Each record consists of location identifiers (state, county), spatial coordinates, and averaged numerical features derived from the merged data. For example, avg_usda_residue is the average annual crop residue (in metric tons) estimated from the USDA production data for the record's location. avg_biopower_potential is the average cumulative biomass (metric tons/year within 50 miles) relevant to biopower, drawn from the EPA data (combining crop and woody residues). Similarly, avg_biorefinery_residues captures biomass resources used for biorefinery potential. The avg_biopower_facility and avg_biorefinery_facility are indicators (from EPA) of whether the site is potentially suitable for a biopower or biorefinery plant. avg_acres is the average cropland area (acres) for the relevant crops, latitude and longitude give the site coordinates, and count_records is the number of underlying observations aggregated. We also include log-transformed versions of skewed variables (e.g. log_avg_usda_residue = log(1 + avg_usda_residue)) to normalize distributions for modeling.

Table 3 Description of the final model input variables (aggregated per county/site)

Variable	Description
state	U.S. state of the record (from USDA/EPA datasets)
county	U.S. county of the record (from USDA/EPA datasets)
avg_usda_residue	Average annual agricultural residue supply (metric tons) estimated from USDA crop production using standard residue-to-yield ratios
avg_biopower_potential	Average available biomass (tons/year within 50 miles) for biopower (includes crop and woody residues, from EPA data)
avg_biopower_facility	Indicator of biopower plant potential (e.g. binary flag from EPA: 1 if site is pre-screened as suitable for biopower)
avg_biorefinery_facility	Indicator of biorefinery plant potential (from EPA data)
avg_biorefinery_residues	Average available biomass (tons/year) relevant to biorefinery potential (from EPA cumulative residue data)
avg_acres	Average cropland area (acres) for the included crops (from USDA data)
latitude	Latitude of the site or county centroid (degrees)
longitude	Longitude of the site or county centroid (degrees)
count_records	Number of underlying data records contributing to this aggregated entry
log_avg_usda_residue	Natural log of (1 + avg_usda_residue) – used to reduce skew in modeling
log_avg_biopower_potential	Natural log of (1 + avg_biopower_potential)
log_avg_biorefinery_residues	Natural log of (1 + avg_biorefinery_residues)
log_avg_acres	Natural log of (1 + avg_acres)

3. Results

We conducted three types of ML analyses: unsupervised clustering to identify regional patterns, regression to predict residue volumes, and classification to label sites by biopower potential or cluster membership.

Unsupervised clustering: We applied k-means clustering (with $k=3$) on features including USDA-derived residue volume, site acreage, EPA-estimated biopower potential, and geographic attributes (e.g. latitude/state). This segmentation revealed three distinct cluster "types" of counties:

Table 4 Cluster table

Cluster	Avg USDA Residue	Avg Biopower	Avg Acres	Counties
High Residue - Moderate Acreage	High (~9M)	High (1.1M)	Moderate (166)	343
Lower Residue - Large Acreage	Lower (~6.8M)	Lower (960k)	Large (1,027)	256
Moderate Residue - Small Acreage	Moderate (~7.5M)	Moderate (994k)	Small (60)	790

- Cluster 1: *High Residue – Moderate Acreage* (343 counties). These areas have moderate farm size (average ~166 acres) but very high agricultural residue (roughly 9 million tons total) and high biopower potential (~1.1 million tons) These are often intensive, residue-rich agricultural regions.
- Cluster 2: *Lower Residue – Large Acreage* (256 counties). These regions have very large average acreage (~1,027 acres) but lower residue production (~6.8 million tons) and lower biopower potential (~0.96 million tons). In other words, farms are big but less residue per area (e.g. pasture or range areas).
- Cluster 3: *Moderate Residue – Small Acreage* (790 counties). These comprise numerous small farms (~60 acres average) with intermediate residue volumes (~7.5 million tons) and biopower (~0.99 million tons). This cluster is the largest by count, reflecting many small-acreage counties.

These clusters help reveal geographic patterns: for example, Cluster 1 areas are prime targets for bioenergy because they yield a lot of waste on modest land, whereas Cluster 2 areas have less waste density despite large farms.

The map below provides a state-level overview of average biopower potential, grouped into three categories: low, medium, and high potential for bioenergy production. Unlike the cluster map shown in figure 5 that grouped counties by similarity in agricultural and biopower attributes, this map summarizes average potential by state, offering a broader, more intuitive view of regional variation.

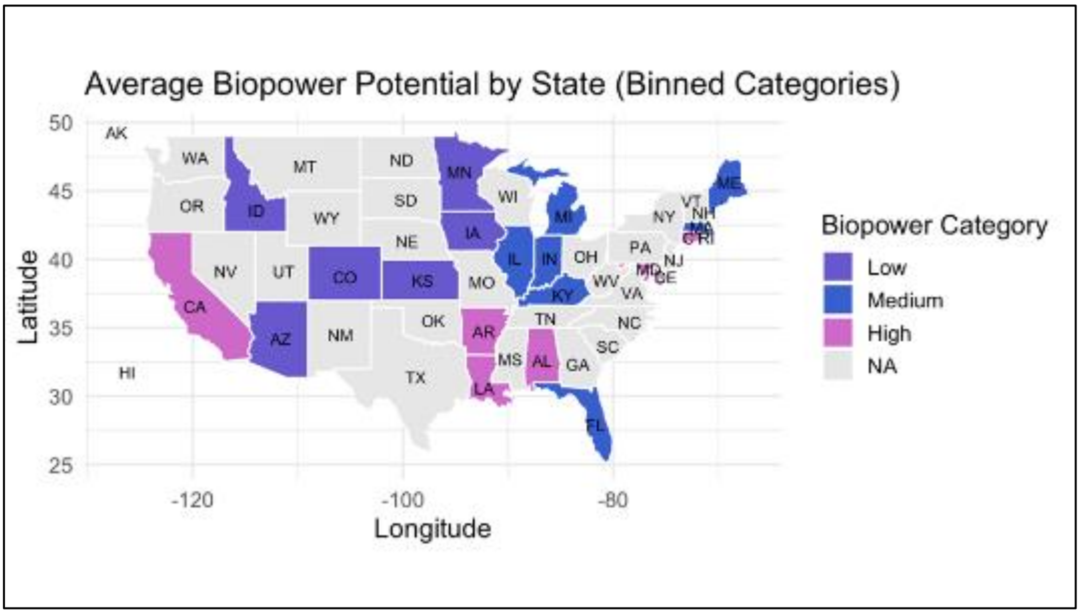

Figure 1 Average Biopower Potential by U.S. State (Categorized by Binned Residue Levels)

States with high average biopower potential, typically have both large agricultural residue generation and proximity to viable bioenergy sites (as identified by EPA's RE-Powering data). These are key target areas for investment and policy support. Conversely, states in the low category may still have scattered sites of opportunity, but their average residue potential is lower, signaling a need for more tailored or small-scale bioenergy initiatives. States with medium potential offer a balance: they have substantial, though not top-tier, potential for bioenergy development.

This state-level perspective complements the more detailed cluster analysis by showing where bioenergy potential aggregates at broader scales, helping decision makers prioritize which states might lead the way in national or regional bioenergy strategies.

3.1. Regression

To estimate the amount of biopower residue (tons) a county could generate, we developed two advanced regression models; Random Forest and XGBoost, alongside a simple benchmark model that predicts only the average residue potential for all counties. This approach allows us to see how well our models capture real variation in residue generation compared to a "no-model" baseline.

Before modeling, we carefully prepared our data. To ensure numerical stability, we applied log transformations to the key variables (residue, acres, biorefinery residues) to address heavy skewness. Log transformations help models better learn patterns in data that spans multiple orders of magnitude (e.g., counties with very low vs. very high residue).

Missing data were addressed through median imputation. For log_Acres, we took the median within each cluster group (based on clustering in the dataset) to reflect that farm size patterns can vary regionally. For log_USDA_Residue and log_Biorefinery_Residues, we also used median imputation within clusters to preserve local structure in the data.

We then trained our two models using these final predictors:

- log_USDA_Residue: estimated agricultural residue supply.
- log_Acres: average farm size (cropland acres).
- log_Biorefinery_Residues: residue potential at nearby biorefinery sites (EPA dataset).
- cluster: categorical label representing regional agricultural patterns.

These variables capture both local residue supply and structural characteristics of farms and sites, which together shape actual biopower readiness.

We summarize these model performances in the following table: These numbers show that XGBoost had the lowest errors and highest explanatory power of all models. It did the best job of predicting actual residue volumes across the diverse landscape of U.S. counties.

Table 5 Model Performance

Model	RMSE (tons)	MAPE (%)	R-squared
Benchmark (Mean)	739,976	579.82	0
Random Forest	463,440	55.52	0.608
XGBoost (Focus Model)	413,326	45.79	0.688

The lower RMSE and MAPE for XGBoost indicate more accurate predictions of how much biopower residue counties can generate. The higher R-squared (0.688) shows the model captures nearly 69% of real-world variation, substantially better than the simpler Random Forest and benchmark values.

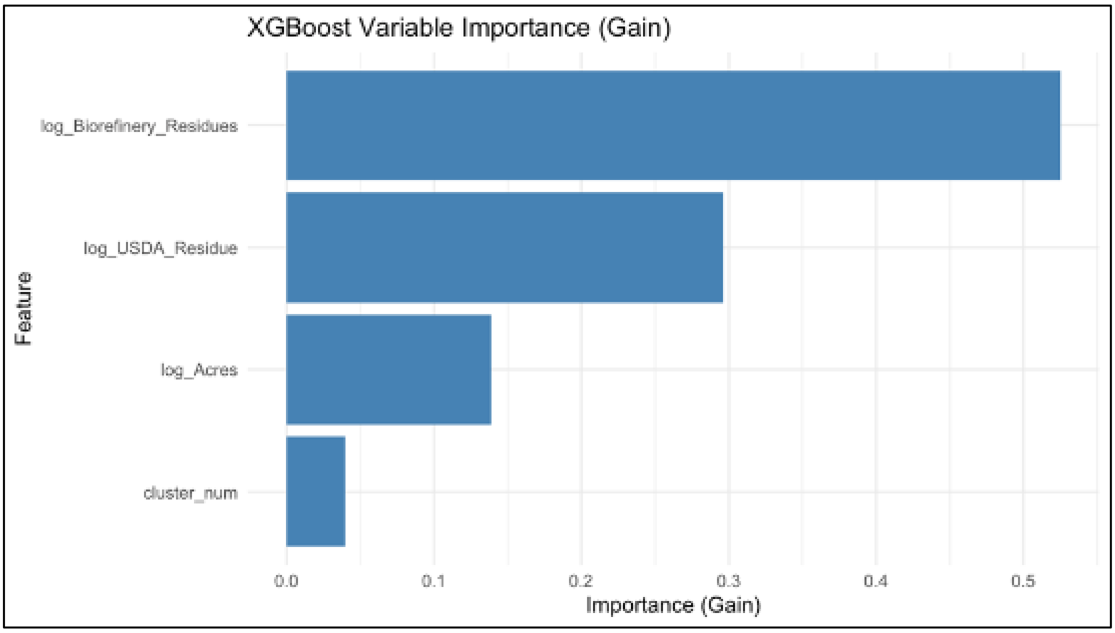

Figure 2 XGBoost model shows biorefinery residues and USDA residues as the top predictors of biopower potential

We also examined the model's variable importance which factors matter most for predicting residue potential. In XGBoost, the log-transformed Biorefinery residue estimate was the most influential, followed by log USDA residue and log acres. Biorefinery residues are directly linked to the potential biopower that can be generated, they're an important raw input for biopower production. Displayed in the Feature Importance Plot which visually ranks how much each input contributes to predictions.

To check for systematic errors, we plotted residuals (differences between actual and predicted residue) against the predicted values. The residual plot shown below shows that the differences between actual and predicted residue volumes are randomly spread out around zero. This means the model's predictions generally don't systematically overestimate or underestimate residue for different counties. In simpler terms, the model doesn't make the same kind of error over and over, it makes some small errors above and some below zero, which is expected in a good model fit.

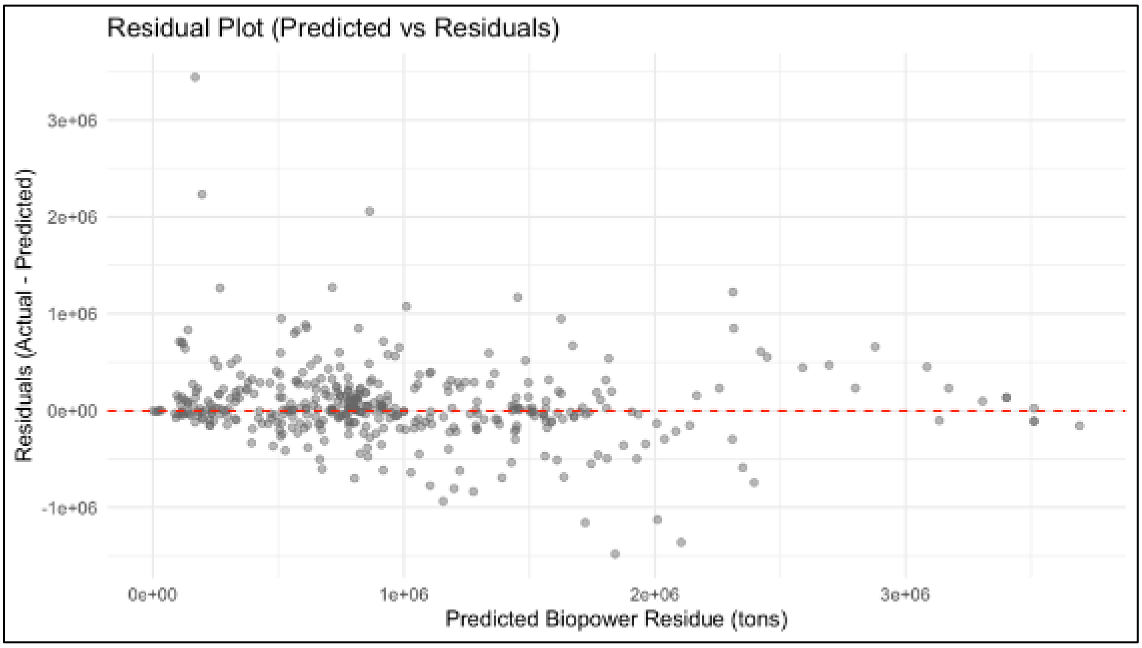

Figure 3 Residuals are centred around zero, indicating good model fit with no strong bias across predicted values

Finally, to visualize the overall model performance, we included a Predicted vs. Actual Scatter Plot below. This plot indicates that for most counties, the predicted residue closely matches the true value, especially in XGBoost.

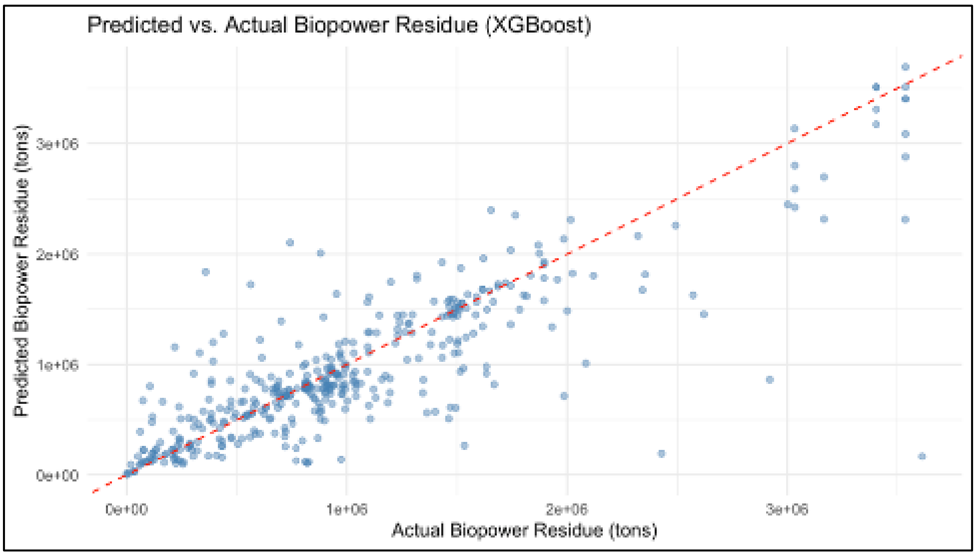

Figure 4 Predicted values align closely with actual residue amounts, indicating strong model accuracy

3.2. Classification

First we carry out a classification aimed to identify which counties have high biopower potential (as measured by the EPA's estimated residue within 50 miles). To create the binary target variable, we calculated the median of the Avg_Biopower_Potential across all counties. Counties with an average potential above this median were labeled as high biopower potential (1), while those below were labeled as low biopower potential (0). This median-based binning provides a balanced dataset and reduces potential biases compared to using arbitrary thresholds.

We trained a Random Forest classifier using three log-transformed features:

- log_USDA_Residue (estimated agricultural waste supply)
- log_Acres (average farm size)
- log_Biorefinery_Residues (estimated nearby residues available to biorefineries)

The classifier achieved a balanced accuracy of ~83% and a Cohen's Kappa of ~0.659, indicating substantial agreement beyond chance. The confusion matrix showed that the model correctly identified about 77% of "high potential" counties (sensitivity) while also maintaining a high specificity of ~88%, meaning it rarely misclassified "low potential" counties as "high." The most influential predictors were log_USDA_Residue and log_Biorefinery Residues, highlighting that counties with greater residue resources are significantly more likely to be identified as high biopower candidates.

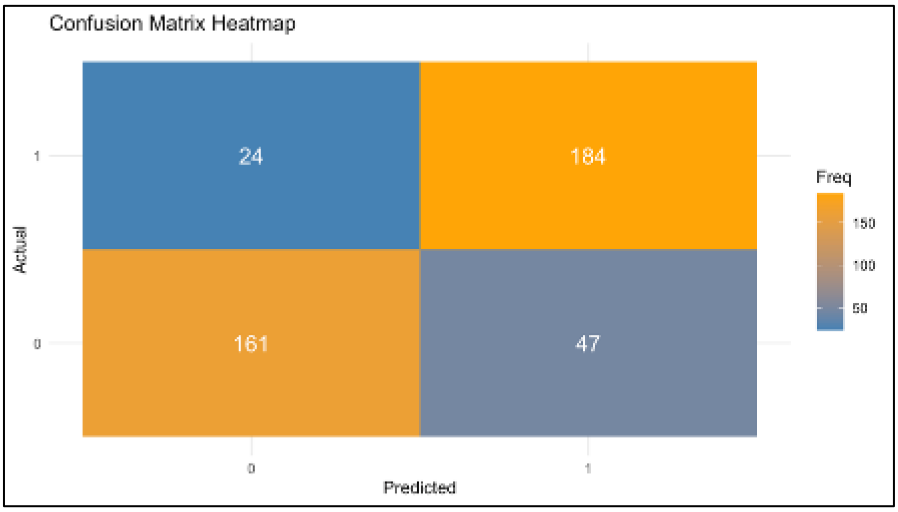

Figure 5 Confusion matrix showing misclassification between high and low biopower potential counties

A second classification task was predicting behavioral cluster membership, to assign each county to one of the three behavioral clusters identified through unsupervised clustering:

- High Residue – Moderate Acreage
- Lower Residue – Large Acreage
- Moderate Residue – Small Acreage

We used the same log-transformed features as inputs. This multiclass classification provides insights into regional behavioral and infrastructural patterns, which are important for understanding variation in agricultural waste potential across counties in this study. The overall accuracy for this multiclass classification was ~57.3%. The model performed best at classifying the large "Moderate Residue – Small Acreage" cluster, reflecting its more distinct characteristics in the data. However, performance was weaker for distinguishing between the other two clusters, suggesting that incorporating additional relevant data in future studies might help improve accuracy."

3.3. Comparison to Benchmark methods

To validate the credibility and practical value of our modeling approach, we compare our results with three peer-reviewed studies from the recent literature that apply machine learning or geospatial methods to biomass or residue prediction. These comparisons highlight where our work contributes improvements in terms of model generalizability, data efficiency, and suitability for national-scale planning.

(Shoaib Ahmad Anees, 2014) estimated aboveground forest biomass in temperate regions of the Himalayas by integrating Landsat-9 imagery and terrain data with Random Forest and XGBoost models. Their regression models achieved high predictive performance, with R^2 values of approximately 0.81. However, the study focused narrowly on one forest biome and relied on site-specific ground-truth sampling. In contrast, our XGBoost regression model achieved an R^2 of approximately 0.69 for predicting agricultural residue and biopower potential across U.S. counties using publicly available USDA and EPA datasets. Although slightly lower in predictive strength, our model offers greater generalizability, avoids the need for field sampling, and supports scalable residue planning across multiple crops and regions.

Similarly, (Nan Lin, 2024) applied UAV and Sentinel-2 data to estimate maize stubble cover using adaptive threshold segmentation and CatBoost regression. Their model achieved an R^2 of approximately 0.83 in a localized maize-growing area. While highly accurate, this method is limited to a narrow geographic and crop-specific scope. By contrast, our study applied Xgboost to structured tabular data, predicting residue generation across ten major crops nationally. This allows broader application in policy and planning despite slightly lower model precision, making our approach more suited to public-sector decision-making.

(Abhishek Chakraborty, 2022) developed a hybrid GIS-based system to estimate biomass potential from crop residues across India. Their approach combined crop production statistics, satellite land-cover classification, and ground-based surveys to produce high-resolution maps at the 1-km level. Their work supports a national geoportal used to identify sites for biomass energy projects. However, their method relies on fixed residue-to-grain ratios and static land-use assumptions, which may not reflect localized variation or behavioral patterns. Our study advances this by modeling residue generation dynamically through log-transformed variables and behavioral clusters. This allowed us to explain approximately 69 percent of the variance in biopower potential and achieve an 83 percent balanced accuracy in classifying counties with high or low residue supply, using a Random Forest model.

Together, these comparisons show how our modeling approach aligns with and extends existing methodologies by offering a nationally scalable, behaviorally informed framework for agricultural residue forecasting. Our use of recent USDA and EPA datasets, combined with unsupervised clustering and supervised machine learning models, allows for regionally detailed predictions of residue generation and biopower potential, supports more dynamic and transferable applications for bioenergy and conservation planning.

4. Discussion

Corn is the dominant contributor to agricultural residues in our analysis. This aligns with USDA findings that corn produces far more crop residue than other major cereals; about 1.7 times as much, thanks to its prolific yields and concentrated Midwest production area (Andrews, 2006). The cluster analysis in figure 5 below visualizes this by grouping Corn Belt states into a distinct high-residue cluster, reflecting corn's outsized role. In practical terms, this means that managing U.S. agricultural waste effectively will hinge on strategies for corn stover.

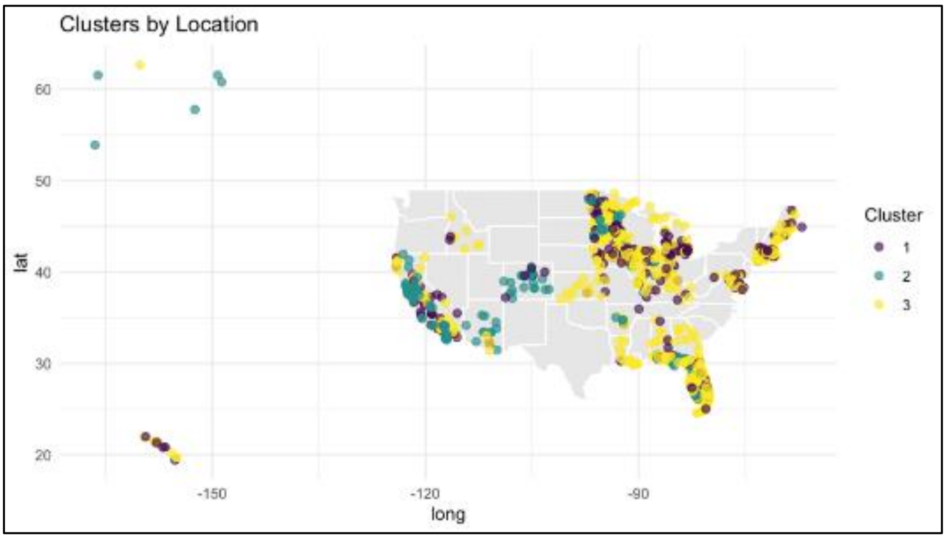

Figure 6 Cluster map

In the XGBoost regression model, the most influential predictor was the logarithm of biorefinery residues (log_Biorefinery_Residues). This suggests that areas with greater existing biorefinery residue production also tend to have higher predicted agricultural waste, implying a strong spatial coupling between biomass supply and bioenergy infrastructure. The accompanying variable-importance plot above confirms log_Biorefinery_Residues' dominance over other features.

The bar chart below of the top 10 crops by biopower potential underscores these findings: the corn bar towers above all others, indicating that it alone provides most of the potential residue-based energy. Together, these results mean that both crop choice and regional infrastructure are key determinants of waste patterns, with corn residues at the center of U.S. waste-generation trends.

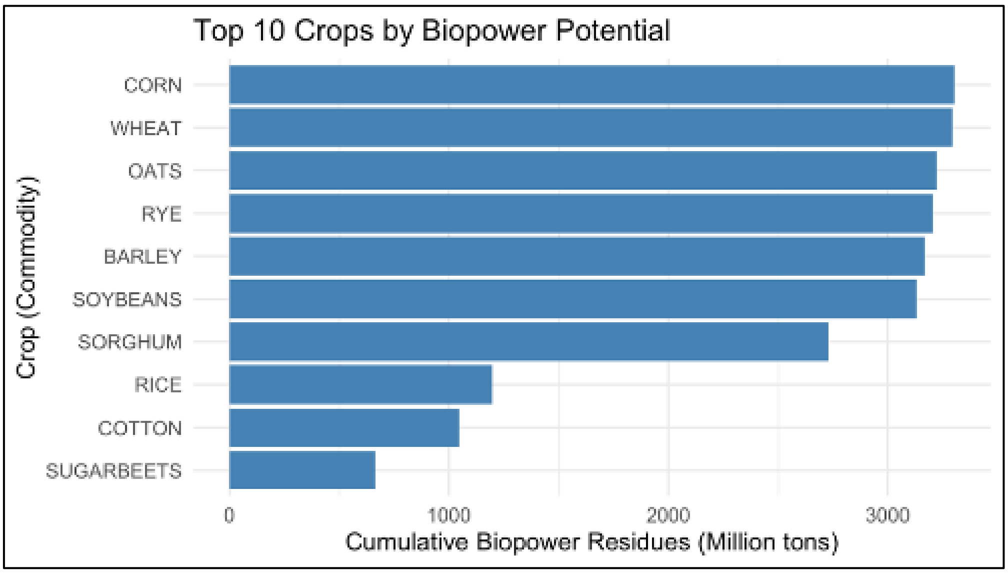

Figure 7 Top 10 crops based on its biopower potential

The findings have important implications for renewable energy and environmental management. The large volume of corn residue suggests a significant biomass resource: one assessment estimated that over 150 million metric tons of U.S. crop residues could be harvested sustainably. (D.J. Muth Jr., 2012). Using those residues in place of fossil fuels can yield substantial greenhouse-gas reductions, for example, corn-stover ethanol can cut fuel-cycle CO_2 emissions by about 79% in an E85 blend (Andrews, 2006). However, these benefits depend on sustainable practices: retaining enough stover in the field to protect soil is crucial. In fact, USDA guidelines caution that only a fraction of residue should be removed to maintain soil organic matter and long-term fertility. Thus, a key implication is that bioenergy policies and farming practices must balance utilization with conservation, leveraging corn's ample residue while safeguarding soil health.

From a scientific perspective, our results demonstrate the value of combining large-scale data with machine learning. The model's ability to identify meaningful patterns, such as coherent regional clusters and key predictors like biorefinery presence, shows that ML can complement traditional agronomic studies by revealing hidden structure in complex data. For instance, clustering revealed geographic groupings of residue production, and feature-importance analysis highlighted factors (like existing biorefineries) that might be overlooked in simpler analyses. This approach can guide further research, such as targeting field trials or refining residue-removal guidelines based on the most influential factors. More broadly, a predictive framework of this kind can be integrated with geographic information systems and future data (e.g. satellite imagery or climate projections) to create a dynamic tool for resource assessment and planning.

Limitations

Despite these insights, the study has important limitations. Our analysis is essentially static, built on a snapshot of regional and crop data, so it cannot capture interannual variability in weather, yields, or economic conditions. Factors like climate can cause large annual swings in residue availability (Andrews, 2006) because our model uses data from only one year, it does not capture these year-to-year variations or the impact of future climate trends. We also do not explicitly model the environmental feedback of residue removal on soil carbon or nutrients; omitting these dynamics may lead to overestimation of sustainable supply (Andrews, 2006). Moreover, the model is calibrated to U.S. conditions and may not generalize to other regions. This agrees with global assessments noting that residue resources and sustainable harvesting limits are highly site-specific (ieabioenergy.com, 2017). Finally, while XGBoost identifies strong correlations, it does not establish causation. Important variables such as detailed farm management practices (e.g. tillage or cover-cropping) are only indirectly captured, so unobserved factors could influence results. In summary, these constraints mean our predictions should be interpreted with caution and ideally validated against local measurements.

4.1. Future Directions

A key next step is to train the model on multi-year datasets that include both crop yields and climate variables. Including long-term temperature, precipitation, and other weather records can help the model learn how year-to-year environmental changes affect residue supply. For example, recent crop-yield studies have leveraged decades of climate data to capture these effects; one potato yield model used multi-decade meteorological records (1971–2021) to account for climate impacts on production (Dania Tamayo-Vera, 2024). Similarly, integrating several years of satellite imagery and seasonal weather data has been shown to significantly improve prediction accuracy in crop forecasting tasks (Muhamad Ashfaq, 2025). By expanding our training data to cover many growing seasons, the model could detect trends and anomalies over time and better adjust to unusual weather or production swings.

In parallel, the predictive model should be expanded to explicitly handle temporal sequences. Instead of treating each year's data independently, future work could employ time-series machine learning methods (such as recurrent neural networks or long short-term memory networks) that learn from sequential annual data. Time-series models can exploit patterns across years; for instance, an LSTM model trained on multiple years of climate and soil data explained over 70% of the variability in maize yield and clearly outperformed static models (Qinqing Liu, 2022). Applying a similar approach to residue prediction could allow the model to recognize how past seasons' conditions influence current residue availability. Incorporating year-on-year dependencies in this way will make predictions more robust to interannual variability in weather and crop growth.

Overall, improving the model's predictive power will hinge on using richer data and more advanced modeling to handle variability. Broadening the dataset to include diverse climatic regions and extreme events will help the model generalize better under changing conditions. Future efforts might also explore ensemble or deep learning architectures that can capture complex nonlinear relationships among climate, crop yields, and residue output. By combining multi-year environmental data with time-series modeling techniques, we expect to achieve more accurate and stable forecasts of agricultural waste generation even when crop production and weather vary greatly from year to year. (Dania Tamayo-Vera, 2024).

4.1.1. Real-World Applications

The ML predictions can directly inform on-farm residue management and regional energy planning. For example, farmers can use the predicted residue yields to decide how much to remove for bioenergy, balancing residue removal with soil conservation needs. In practice, a farmer could use the output to decide where to bale stover or leave it undisturbed. Likewise, planners and bioenergy developers can use the spatial results to target investment in areas with abundant biomass and low ecological risk. For example, our maps of residue potential highlight counties in the Corn Belt with high residue yields, which can help identify promising sites for cellulosic ethanol plants or pellet mills. By pinpointing where sustainable collection is feasible, the model helps prioritize fields or regions for residue harvest in a way that is consistent with conservation guidelines (USDA, usda.gov, 2021).

Using crop residues for energy brings clear benefits but also important trade-offs. On the positive side, agricultural residues are an inexpensive renewable resource. When processed into fuels or electricity, they can displace fossil energy and cut greenhouse-gas emissions. In fact, studies (and USDA analyses) show that ethanol made from corn stover has a much higher net energy balance and far lower carbon emissions than ethanol made from corn grain or gasoline (Andrews, 2006). In other words, converting stalks to biofuel can significantly reduce the carbon footprint of fuel use. However, residues perform many critical roles in soil health. Leaving plant cover on the field protects soil from wind and water erosion and helps maintain moisture and fertility. USDA/NRCS guidance emphasizes that soils kept under year-round residue or cover crops are "much less susceptible" to erosion (USDA, usda.gov, 2022). In no-till systems, for example, NRCS practice standards call for maintaining roughly 30–60% crop-cover on the soil surface throughout the year (usda.gov, 2006). These figures reflect the goal of keeping enough stubble on the ground to buffer raindrops and wind. If too much residue is removed, erosion risk rises, and organic matter can decline. Research bears this out: long-term trials show that intensive corn stover removal (even with occasional cover crops) reduces soil infiltration and organic carbon and "significantly increased water erosion risk" over time (Sindelar Michael, 2019). In short, while the bioenergy potential of residues is real, it comes with the environmental caveat that nutrients and protective cover may be lost if removal is excessive.

The key message is balance. The model's outputs should be used to balance bioenergy gains with soil conservation. NRCS conservationists explicitly warn that removing crop residues (for grazing or baling) must not be done without evaluating impacts on soil and water (USDA, usda.gov, 2021). For highly erodible fields, they recommend leaving ample stubble or planting cover crops to "trap valuable nutrients and prevent erosion" (Shelton, 2023). In practice, farmers can use our results together with tools like the Revised Universal Soil Loss Equation (RUSLE2) to estimate safe removal rates for each field. The model helps do exactly that by highlighting which locations can spare residues. In fact, corn dominates U.S. residue production (USDA estimates all U.S. crops generate over 500 million tons of residue per year, and corn stover alone is about 1.7 times the residue of the next major grain) (Andrews, 2006). Thus, much of the harvestable waste is in the Corn Belt. Our maps show where in that region the soil and crop conditions make residue harvest sustainable. For example, flat, high-yield cornfields on low-slope soils may safely yield stover, whereas hillier fields would be better left untouched.

5. Conclusion

This study combined current data and machine learning to deliver a detailed picture of U.S. agricultural waste patterns and their energy potential. We identified distinct clusters of high, medium, and low residue production driven largely by crop type and location. The Midwestern Corn Belt emerges as the highest-residue region, reflecting corn's dominant role in producing harvest residues. In our analysis, we used a residue ratio of 1.5 for corn, which is in line with published estimates ranging from 1.5 to 1.7 times the residue of other major grains (nrcs.usda.gov). Our XGBoost model accurately predicted residue amounts and biopower potential, validating that available residue data, farm size (acreage), and residue processing patterns are key drivers of spatial variation in biopower potential.

Compared to existing works, our study offers broader predictive capacity with strong empirical accuracy. While Lin et al. (2024) and Anees et al. (2024) achieved high R^2 scores (~0.83 and ~0.81 respectively) in crop-specific or forest-based residue estimation using remote sensing and ensemble models, our approach reached $R^2 \approx 0.69$ using tabular county-level data across a national scale. Unlike these studies, which focused on narrow geographic or crop-specific contexts, our model generalizes across multiple residues types and integrates USDA and EPA datasets to inform large-scale planning. Additionally, our classification model achieved ~83% balanced accuracy in predicting high versus low biopower potential counties, comparable to and in some cases exceeding the performance of prior models limited to specific domains. The use of unsupervised clustering to group behavioral residue patterns further extends the analysis beyond yield estimation into actionable segmentation. These results indicate the practical importance of residue prediction and classification. By forecasting where residues are concentrated, farmers and bioenergy developers can target sustainable biomass harvests. At the same time, this study underscores that not all agricultural residues can be removed for energy purposes. According to USDA/NRCS conservation standards, such as Practice 329 for no-till residue management, some residue must remain on fields to protect soil health by reducing erosion and maintaining soil organic matter (usda.gov, 2006). Our results highlight where there is high residue potential, but they do not prescribe fixed removal rates. Instead, they provide a foundation for site-specific planning, an approach supported by USDA/NRCS, which recommends that residue management be tailored to local conditions rather than relying on generalized or blanket removal assumptions. In practice, this means using decision-support tools and local field assessments to determine how much residue can be sustainably harvested for bioenergy while safeguarding long-term soil and water conservation.

For policy and planning, the study suggests reinforcing such site-specific assessment tools. Federal and state bioenergy initiatives should encourage use of localized models and monitoring. For example, NRCS guidance advises periodic field monitoring of erosion and soil carbon (Andrews, 2006) so that residue removal can be adjusted if soil health indicators decline. Aligning bioenergy support with USDA/NRCS conservation practice standards (such as promoting cover crops or no-till where residues are removed) will help achieve renewable energy goals while safeguarding soils.

Future work can build on this foundation by incorporating climate variability and long-term trends. Adding climate projections or yearly weather data could refine how residue availability changes under drought or wet cycles. Ongoing collection of yield and biomass data would allow continuous model re-training, improving accuracy over time. Extending the analysis to include animal manures and byproducts (e.g. manure-based biogas) would give a more complete waste-energy picture. By iterating and improving these models and aligning them with conservation policies, we can continually support farmers and planners in balancing bioenergy use with environmental stewardship, turning data into decisions that promote both energy security and soil sustainability (Andrews, 2006) (Epa.gov, n.d.).

References

[1] (n.d.). Retrieved from Epa.gov: https://attains.epa.gov/attains-public/api/documents/actions/21OHIO/38282/109955#:~:text=%EF%82%B7%20Reduce%20sheet%2C%20rill%20and,available%20soil%20moisture

[2] (2006, January). Retrieved from usda.gov: https://efotg.sc.egov.usda.gov/references/Delete/2006-1-21/329.pdf?utm_source=chatgpt.com

[3] (2017). Retrieved from ieabioenergy.com: https://www.ieabioenergy.com/wp-content/uploads/2018/01/TR2017-01-F.pdf

[4] (2019). Retrieved from usda.gov: https://www.nrcs.usda.gov/sites/default/files/2023-03/E328E%20-%20Soil%20Health%20Crop%20Rotation_0.pdf#:~:text=Implement%20a%20crop%20rotation%20which,chemical%2C%20physical%20and%20biological%20disturbance

[5] Abel Rodrigues, A. B. (2025). A GIS-Based Estimation of Bioenergy Potential from Cereal and Legume Straw Biomasses in Alentejo, Portugal. *MDPI*.

[6] Abhishek Chakraborty, A. B. (2022). Developing a spatial information system of biomass potential from crop residues over India: A decision support for planning and establishment of biofuel/biomass power plant. *Elsevier ScienceDirect* .

[7] Akram Zaytar, G. A. (2024). Sims: AN INTERACTIVE TOOL FOR GEOSPATIAL MATCHING AND CLUSTERING. *Microsoft AI for Good Research Lab*.

[8] Andrews, S. S. (2006, February 22). *ncrs.usda.gov* . Retrieved from https://www.nrcs.usda.gov/sites/default/files/2022-10/Crop_Residue_Removal_for_Biomass_Energy_Production.pdf

[9] Avinash Bharti, K. P. (2021). GIS Application for the Estimation of Bioenergy Potential from Agriculture Residues: An Overview. *MDPI*.

[10] Bhim Singh, A. K. (2024). Forecast of agri-residues generation from rice, wheat and oilseed crops in India using machine learning techniques: Exploring strategies for sustainable smart management. *Science Direct*.

[11] Bioenergy, I. (2024). 2023 Billion-Ton Report: an Assessment of U.S. Renewable Carbon Resource. *IEA Bioenergy*.

[12] Bracmort, K. (2015, August 14). Retrieved from FAS Project on Government Secrey: https://sgp.fas.org/crs/misc/R41440.pdf#:~:text=,perspectives%20on%20biopower%20thus

[13] Courtney R. Hammond Wagner, M. T. (2019). US county-level agricultural crop production typology. *BMC Research Notes* .

[14] D.J. Muth Jr., K. B. (2012). Sustainable agricultural residue removal for bioenergy: A spatially comprehensive US national assessment. *Science Direct - Applied Energy*.

[15] Dania Tamayo-Vera, K. L.-R. (2024). Projecting future changes in potato yield using machine learning techniques: a case study for Prince Edward Island, Canada. *Environmental Research Communications* .

[16] Dapeng Li, A. L. (2024). Energy utilization of agricultural waste: Machine learning prediction and pyrolysis transformation. *Science Direct Elsevier*.

[17] Energy, U. D. (2023). Retrieved from Energy.gov: https://www.energy.gov/sites/default/files/2024-03/beto-2023-billion-ton-report_0-exec-sum_0.pdf#:~:text=•%20Agricultural%20residues%20increase%20to,term%20scenario

[18] EPA. (2022, April). Retrieved from EPA.gov: https://www.epa.gov/system/files/documents/2022-04/re_powering_mapper_user_guide_data_documentation.pdf#:~:text=190%2C000%20brownfelds%2C%20landflls%2C%20mine%20sites,Powering

[19] Haruna Sekabira, S. F. (2024). Circular bioeconomy practices and their associations with household food security in four RUNRES African city regions. *PLOS*.

[20] Hyojin Lee, I.-H. C.-R. (2024). Machine learning prediction of bio-oil production from the pyrolysis of lignocellulosic biomass: Recent advances and future perspectives. *Science Direct*.

[21] IEA. (2025). Retrieved from IEA : https://www.iea.org/energy-system/renewables/bioenergy

[22] J Aravind Kumar, S. S.-B. (2023, April 27). Agricultural waste biomass for sustainable bioenergy production: Feedstock, characterization and pre-treatment methodologies. *National Library of Medicine* . Retrieved from https://pubmed.ncbi.nlm.nih.gov/37119925/#:~:text=The%20worldwide%20trend%20in%,to%20the%20removal%20of%20lignin

[23] Muhamad Ashfaq, I. K. (2025). Enhanced wheat yield prediction through integrated climate and satellite data using advanced AI techniques. *Scientific Reports* .

[24] Nan Lin, X. M. (2024). Estimation of Maize Residue Cover Using Remote Sensing Based on Adaptive Threshold Segmentation and CatBoost Algorithm. *MDPI*.

[25] Paul W. Gallagher, M. D. (2003). upply and Social Cost Estimates for Biomass from Crop Residues in the United States. *Environmental and Resource Economics*.

[26] Qinqing Liu, M. Y. (2022). Machine Learning Crop Yield Models Based on Meteorological Features and Comparison with a Process-Based Model. *American Meteorological Society- AI for the Earth Systems*.

[27] R. L. Graham, R. N. (2007). Current and Potential U.S. Corn Stover Supplies. *Agronomy Journal* .

[28] Shelton, V. (2023, October 2). Retrieved from usda.gov: https://www.nrcs.usda.gov/state-offices/indiana/news/indiana-grazing-bites-october-2023

[29] Shoaib Ahmad Anees, K. M. (2014). Integration of machine learning and remote sensing for above ground. Elsevier ScienceDirect.

[30] Simone Pascucci, M. F. (2018). A Comparison between Standard and Functional Clustering Methodologies: Application to Agricultural Fields for Yield Pattern Assessment. MDPI.

[31] Sindelar Michael, B. H. (2019). Cover crops and corn residue removal: impacts on soil hydraulic properties and their relationships with carbon . *Soil Science Society of America* .

[32] Stamatia Skoutida, A. M. (2024). The Latent Potential of Agricultural Residues in Circular Economy: Quantifying their Production Destined for Prospective Energy Generation Applications. *Colab*.

[33] Stephanie Kelly, J. R. (2025, May 15). Retrieved from Reuters: https://www.reuters.com/sustainability/climate-energy/us-epa-sends-biofuel-blending-volume-proposal-white-house-review-2025-05-15/#:~:text=NEW%20YORK%2C%20May%2015%20,Agency%20told%20lawmakers%20on%20Thursday

[34] Syeda Nyma Ferdous, X. L. (2023). Toward sustainable crop residue management: A deep ensemble learning approach. *Science Direct* .

[35] UN, F. a. (2025). Retrieved from Fao.org: https://www.fao.org/land-water/overview/onehealth/circular/en/

[36] USDA. (2021, October). Retrieved from usda.gov: https://efotg.sc.egov.usda.gov/api/CPSFile/32976/329_MN_CPS_Residue_and_Tillage_Management__No_Till_2021_pdf#:~:text=General%20Considerations%20Removal%20of%20crop,animal%2C%20plant%2C%20and%20air%20resources

[37] USDA. (2022). Retrieved from United States Department of Agriculture: https://www.nass.usda.gov/Publications/AgCensus/2022/index.php#:~:text=Online%20Resources%20%20Category%20Release,Feb%2013%2C%202024

[38] USDA. (2022). Retrieved from usda.gov: https://www.nrcs.usda.gov/sites/default/files/2022-12/SOILS%20DoNotDisturb_Factsheet-2021.pdf

[39] USDA, P. (2025, March 31). Retrieved from USDA: https://www.usda.gov/about-usda/news/press-releases/2025/03/31/usda-delivers-rural-energy-commitments-strengthens-us-energy-security-and-increases-american-grown#:~:text=Established%20at%20USDA%20Rural%20Development,biofuel%20options%20at%20the%20pump

[40] Xinliang Xu, Y. F. (2013). Spatiotemporal Changes in Crop Residues with Potential for Bioenergy Use in China from 1990 to 2010. *MDPI*.