# npj Science of Food

## Article in Press

# Mapping gas permeability of sustainable packaging materials to link food barrier needs by clustering algorithms

Ting Yu Yeh & Deniz Turan

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

# Mapping Gas Permeability of Sustainable Packaging Materials to Link Food Barrier Needs by Clustering Algorithms

**Ting Yu Yeh**[1], **Deniz Turan** [1*]

[1] *Food Quality and Design, Wageningen University & Research, P.O. Box 17, 6700 AA Wageningen, the Netherlands*
*\* Corresponding author, deniz.turan-kunter@wur.nl*

**Abstract**

Food packaging is critical for ensuring food safety, quality, and shelf life. However, growing environmental concerns with conventional plastics drive the search for sustainable alternatives. A major challenge that many biobased and biodegradable materials show poor barrier properties, limiting their use for foods. This study provides a proof-of-concept for classifying sustainable packaging materials by clustering oxygen transmission rate (OTR) and water vapor transmission rate (WVTR) data. A dataset from 49 studies (2000–2016) was analyzed using K-Means, Gaussian Mixture Model (GMM), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN emerged as best performing algorithm, achieving the highest Silhouette Score (0.910) and lowest Davies-Bouldin Index (0.374). Results validated that while many sustainable films exhibit high permeability, nanocomposites achieved improved barrier performance. This data-driven framework demonstrates clustering as a tool for systematic grouping of packaging materials, with future work requiring broader datasets, industrial benchmarks, and standardized reporting for practical application.

**Keywords**: oxygen transmission rate (OTR), water vapor transmission rate (WVTR), DBSCAN, food preservation, barrier properties, unsupervised learning, machine learning in packaging

## 1.Introduction

Food packaging plays a crucial role in ensuring the safety, quality, and shelf life of food products. By acting as a protective barrier against external factors, such as moisture, oxygen, and contaminants, it reduces food waste by preserving the freshness and nutritional value of the food products[1]. To maintain the shelf life of the products, packaging materials must exhibit appropriate gas-barrier properties. The protection effectiveness of these materials depends on their permeability to gases and vapors, which is defined as the transmission of substances through a material and involves four key processes: sorption on the surface, dissolution within the material, diffusion across the material, and desorption from the opposite surface[2].

In the field of food packaging, the most critical permeability metrics are the water vapor transmission rate (WVTR) and oxygen transmission rate (OTR), as they significantly affect the shelf life of food products[3]. The process of moisture sorption has the potential to induce chemical or physical instability. In addition, the loss of moisture from products through permeation during storage can also lead to changes in their quality[4]. For example, lactose crystallization can occur in milk powder without proper packaging due to moisture sorption[5]. Oxygen ingress contributes to food spoilage through lipid oxidation in fatty foods, further reducing shelf life, so high-fat foods, such as cheese, require high oxygen barrier packaging. Therefore, it is essential to match the barrier properties of packaging materials to the requirements of food groups based on the food characteristics to minimize the food waste.

Plastics dominate food packaging due to their versatility and most importantly tailored to provide varying degrees of barrier protection, depending on the polymer composition[1]. 40% of food in Europe packaged in plastic[6]. However, 95% of plastic packaging, mainly LDPE, HDPE, and PET, is not recycled, costing the global economy $80–120 billion annually. Only 2% is recycled, while most end up in landfills or incineration[7]. This has driven a new regulation toward sustainable alternatives, reusable, recyclable, and biodegradable materials[8]. Publications on sustainable packaging have surged from 70 to 1,300 in a decade (number of publications under keyword "sustainable packaging material" in Web of Science between 2015-2025), highlighting academia's growing research focus on eco-friendly solutions like lightweight biodegradable packaging to reduce waste, material use, and transportation costs[9].

A study done by Guillard et al. (2018) highlighted the two key challenges in food packaging and sustainability that limit the large-scale market adoption. First, while extensive research has focused on the development of bio-based packaging materials derived from renewable and biodegradable sources, these materials often exhibit inferior barrier and mechanical properties compared to conventional plastics[10]. Since many biodegradable polymers are hydrophilic, their permeability to gases tends to increase under high humidity due to water absorption[11]. These properties indicate that such materials may not provide the same level of protection against moisture and gas permeation as conventional materials. Consequently, maintaining the shelf life and quality of food products becomes challenging, potentially leading to increased food waste, which is contrary to the primary purpose of food packaging. Second, there is a lack of tools to help stakeholders match packaging properties with specific food requirements[12]. While decision support tools, such as those from the Netherlands Institute for Sustainable Packaging (KIDV), are available to predict the environmental impact of packaging based on factors such as circularity, material recyclability, and life cycle assessment (LCA)[13], there is still a need for tools that match gas transfer properties with requirements for different food categories. For example, instant coffee requires very low WVTR and OTR to prevent moisture-induced clogging and oxidation-related flavor loss[14]. In contrast, fresh produce, such as fruits and vegetables, require higher OTR and WVTR to allow for post-harvest respiration. High-fat foods such as cheese require low oxygen permeability but can tolerate higher water vapor permeability[15]. The absence of such specialized decision tools hinders stakeholders from effectively matching sustainable packaging materials to the unique barrier needs of different food products.

While existing studies have identified suitable biodegradable materials for food groups such as biodegradable chitosan-cellulose and polycaprolactone film effective for fresh products like shredded lettuce[16] and polyethylene with corn starch for preservation of lean beef[17], these insights are often generalized and lack information on conditions and permeability information. These studies focus primarily on the development of novel materials or the testing of individual materials in isolation, often overlooking a holistic, comparative approach. As a result, many food products remain unsuitable for currently available biodegradable packaging. There is no comprehensive study that systematically matches sustainable packaging materials to food groups based on gas barrier needs. Without a clear framework to identify which sustainable materials offer barrier performance comparable to conventional plastics for specific food types, inappropriate packaging

choices may compromise food safety, quality, and shelf life. As a result, there is an urgent need for a decision tool to match material gas permeability to specific food requirements.

To address this gap, this work aims to demonstrate a proof-of-concept for how clustering approaches can classify sustainable packaging materials based on gas permeability, specifically oxygen and water vapor permeability, thereby supporting the identification of alternatives to conventional plastics. Unlike previous studies that analyze materials in isolation, this data-driven approach provides a structured framework for material selection. By leveraging insights from 49 scientific studies (2000–2016)[11], the analysis illustrates how such clusters can be positioned relative to the barrier ranges reported for different food categories (fresh meat, fruits and vegetables, seafood, bakery products, cheese, peanuts and snacks, coffee, and baby food), without implying prescriptive matching. This framework is intended to establish a methodological foundation for future studies based on more comprehensive, industrially relevant datasets, ultimately supporting more systematic evaluation of sustainable packaging options.

## 2. Results and Discussion

All clustering analyses were performed on logarithmically transformed values of the water vapor transmission rate (WVTR) and oxygen transmission rate (OTR) to group sustainable packaging materials by their permeability characteristics, as defined in the Data preprocessing section 3.2. The distribution of OTR and WVTR was investigated after confirming measurement conditions and was given in **Figure 1.** OTR exhibited a lower average compared to WVTR but had a wider range, with several extreme outliers. The OTR distribution was strongly right skewed, while WVTR showed a slight left skew. These characteristics shows why we applied standardization before modeling. Given the non-centroid distribution of the gas permeability data, we hypothesized that DBSCAN will outperform traditional centroid-based methods (such as K-Means) in accurately classifying sustainable packaging materials. Additionally, the K-means clustering algorithm is used as a baseline for performance comparison. Since the optimal number of clusters is unknown, density-based clustering models such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is employed to address this uncertainty[18].

## 2.1 The chosen model to classify sustainable packaging materials based on their gas permeability

In this section, clustering algorithm performance comparisons (K-Means, DBSCAN, and GMM) were discussed by using the Silhouette Score and the Davies-Bouldin Index. **Table 1** shows Silhouette Scores and Davies–Bouldin Index values before and after data standardization by measurement conditions, referring to screening OTR values to 23 °C and 50% RH and WVTR values to 25 °C and 50% RH, ensuring comparability across studies. This step was performed after confirming measurement conditions (see Methods 3.2) and reduced variability from inconsistent reporting.

After standardization by measurement conditions, DBSCAN showed substantial improvement in both metrics (Silhouette Score 0.558 → 0.910; Davies–Bouldin Index 0.520 → 0.374). K-Means improved strongly in the Silhouette Score (0.583 → 0.905) but showed moderate improvement in the Davies–Bouldin Index (0.673 → 0.482). GMM improved moderately in both metrics (Silhouette Score 0.539 → 0.899; Davies–Bouldin Index 0.772 → 0.445), although the improvement in Silhouette Score was smaller than that observed for DBSCAN and K-Means. These results suggest that DBSCAN provided the most consistent clustering performance with the highest Silhouette Score and the lowest Davies-Bouldin Index after standardization. This outcome reflects the suitability of density-based clustering for handling the non-centroid distribution of gas permeability data, highlighting the role of preprocessing and parameter selection in improving clustering accuracy.

**Table 1.** Comparison of clustering algorithms (K-Means, DBSCAN, and GMM) based on Silhouette Score and Davies-Bouldin Index before and after standardization by measurement conditions for OTR and WVTR to 23 °C/50% RH and 25 °C/50% RH, respectively.

| Metric | Algorithm | Before standardization | After standardization |
|---|---|---|---|
| | K-Means | 0.583 | 0.905 |
| Silhouette Score | **DBSCAN** | **0.558** | **0.910** |
| | GMM | 0.539 | 0.899 |

| | K-Means | 0.673 | 0.482 |
|---|---|---|---|
| Davies- Bouldin Index | **DBSCAN** | **0.520** | **0.374** |
| | GMM | 0.772 | 0.445 |

To date, no studies have specifically applied clustering algorithms to the gas permeability properties of packaging materials. Existing clustering approaches have predominantly focused on classifying materials based on mechanical or physical properties. For instance, K-Means has been used to categorize plastic polymers based on mechanical properties[19], while Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) has been utilized to classify mixed-mail items based on attributes such as flexibility and friction, addressing logistical challenges posed by packaging heterogeneity in parcel shipments[20]. In the study[20], the adjusted mutual information (AMI) score was used to evaluate the clustering model because the benchmark study was conducted to act as ground truth label. This performance index compares the clustering results with the ground truth labels. In our study, although a similar density-based algorithm was used, the clusters were not determined before the algorithm. Therefore, the silhouette score was chosen as the most appropriate evaluation metric.

In contrast, research on gas permeability has largely focused on predicting the permeability of polymers using machine learning techniques. For example, a study[21] used a neural network to correlate the infrared spectrum of a polymer with its permeability for 33 polymers. More recently, another study[22] used a graph neural network trained on both experimental and simulated gas permeability data from 820 polymers to improve prediction accuracy. The limited application of clustering to gas permeability data may be due to the inherent complexity of these datasets, as gas permeability of polymers is influenced by multiple interdependent factors, including polymer structure, free volume, gas-polymer interactions, and processing conditions[22]. Therefore, the high dimensionality and nonlinear relationships among these parameters pose challenges to effectively structuring the data for clustering analysis.

Despite providing promising insights, the study is constrained by a relatively small dataset, that covers publications from 2000 to 2016. During this period, research on sustainable packaging materials was less extensive, resulting in a narrower range of material properties. Data quality was

also a significant challenge, with inconsistencies and missing information for key parameters. The barrier performance of the materials can be influenced by various external and internal factors, such as the sample preparation methods, test conditions as well as micromorphology[23]. Among these factors, test conditions such as temperature, relative humidity, and material composition significantly affects the reliability and reproducibility of permeability measurements. Additionally, mislabeled data, such as oxygen reported mistakenly for carbon dioxide, highlighted the need for standardized reporting protocols. Given the proliferation of sustainable packaging studies over the past decade, future research can benefit from a broader and more current dataset, provided that issue of data accuracy and consistency issues are addressed.

Since the research used only a small amount of data on gas permeability, expanding the dataset either on gas permeability or beyond gas permeability such as mechanical strength would help to linking packaging properties directly to food product shelf-life. Because the measurement condition was reported ambiguously, standardized and transparent reporting of measurement conditions, material compositions, and processing parameters is essential to ensure reliable, comparable results. Researchers and industry professionals will be better equipped to identify the correct barrier properties of sustainable packaging materials. While this study demonstrates the viability of clustering algorithms in clustering sustainable materials according to gas permeability, future work with more comprehensive or higher quality data is crucial for closing the gap between biodegradable and conventional packaging solutions.

## 2.2 Matching gas barrier properties of sustainable packaging materials with barrier needs of foods

The barrier ranges for different food categories, namely, fresh meat, fruits& vegetables, seafood, bakery products, cheese, peanuts & snacks, coffee and baby food were adapted[15]. According to the evaluation metrics, the DBSCAN algorithm produced the best clustering outcome with a higher Silhouette Score (0.910) and a lower Davies-Bouldin Index (0.374) for clustering sustainable packaging materials' gas permeability after standardization. Therefore, the following matching discussion is based on the DBSCAN clustering model results, which identified three distinct groups (purple, orange and green), while the remaining points were classified as noise. The clustering results of K-Means and GMM can be referred to **Supplementary Figure 1 and 2.**

As shown in **Figure 2** and summarized in **Table 2**, Cluster 1 (purple, n=57) consists primarily of fish-gelatin nanocomposites with montmorillonite (see **Supplementary Table 1**). These materials exhibit very low OTR (mean $\approx$ 1.9 cm$^3$/m$^2$day) and moderate WVTR ($\approx 10^3$ g/m$^2$day). In Figure 2, these materials appear positioned near the lower-OTR range reported for cheese packaging; however, this proximity should be interpreted descriptively rather than as an indicator of suitability. No quantitative distance or overlap validation was performed in this study, the apparent alignment reflects only the relative placement of clusters within the log-transformed OTR/WVTR feature space. The dataset is limited and does not capture the influence of processing conditions or scalability. For instance, the article in the dataset discussed ultrasonic dispersion proving critical for achieving high-barrier performance[24], which shows the sensitivity of nanocomposite performance to processing conditions. This highlights the importance of interpreting the clusters as emergent patterns in the literature rather than as ready-to-use design guidelines.

Recent research has increasingly shifted from passive barrier enhancement toward active PLA-based nanocomposites. In these systems, nanofillers alone are often insufficient, however multi-week shelf-life extension for cheese has been achieved through the incorporation of active agents such as essential oils[25], Ag–graphene–TiO$_2$[26], or zero-valent iron (ZVI) nanoparticles[27]. This transition reflects a broader move toward multifunctional packaging systems. However, these advances come with trade-offs. For example, active PLA nanocomposites frequently show reduced transparency, changes in color, and altered mechanical strength, all of which must be addressed for industrial adoption. In addition, potential toxicological and migration concerns associated with nanomaterials require careful assessment to ensure compliance with EU food contact regulations[28].

A further limitation arises from the dataset construction itself. Some well-known studies on nanocomposites, for example, the incorporation of organically modified montmorillonite (OMMT, Cloisite) into thermoplastic starch (TPS), which enhanced tensile strength and reduced WVTR[29], were not captured in TRANSMAT due to the strict ontological criteria, manual annotation, and copyright restrictions applied during curation. While this focused methodology to build the current dataset ensures high-quality annotation of gas permeability and packaging composition, it inevitably excludes other relevant publications, reducing representativeness. This highlights the need for broader and more automated data curation in future efforts to strengthen the robustness of clustering outcomes.

**Table 2**. Summary of DBSCAN cluster outcomes, showing the average (± standard deviation) OTR (cm$^3$/m$^2$day) at 23 °C and 50% RH, and WVTR (g/m$^2$day) at 25 °C and 50% RH for assumed thickness 100 μm in each cluster, along with their dominant base material, type, and secondary material

| Cluster | 1 (n = 57) | 2 (n = 12) | 3 (n = 180) |
|---|---|---|---|
| OTR (cm$^3$/m$^2$day) | 1.95 ± 0.48 | 6,409.6 ± 7309.3 | 10.3±5.4 |
| WVTR (g/m$^2$day) | 1,125.2 ± 282.2 | 2,158.8 ± 1,632.4 | 264,385.3±106,946.5 |
| Dominant Base Material | fish gelatin | polylactic acid | carrot puree |
| Dominant Type | nanocomposite | individual | blend |
| Dominant Secondary Material | montmorillonite | None | CMC |

Cluster 2 (orange, n=12) includes both conventional polyolefins (LDPE, HDPE) and biobased, industrially biodegradable polymer PLA (see **Supplementary Table 2**). Polyolefins are represented by triangles appearing at the margins of Cluster 2 in **Figure 2**. This cluster is characterized by very high OTR (6,409.6 ± 7,309.3 cm$^3$/m$^2$day) and intermediate WVTR (2,158.8 ± 1,632.4 g/m$^2$day) (**Table 2**). Polyolefins appear toward the lower-WVTR side of the cluster, while PLA films span a broader range, reflecting known variability linked to processing and crystallinity. DBSCAN identified Cluster 2 as a small, coherent group seperating from gelatin nanocomposites (Cluster 1) and polysaccharide blends (Cluster 3). However, the small sample size (n = 12) and high standard deviation mean that the grouping should be interpreted cautiously. Although PLA appears positioned near permeability ranges reported for fruits, vegetables, and bakery products as sustainable alternative to polyolefins based on clustering results, this outcome reflects the relative distribution of PLA values within the limited dataset rather than a true functional equivalence to polyolefins. The relatively small dataset used in this study limited the granularity of clustering. Expanding the dataset to include a larger and more diverse set of permeability values would prevent such forced groupings, improve representativeness, and allow clustering to reveal more fine-grained distinctions relevant to food packaging applications.

There have been several studies on improving the barrier properties of neat PLA in the dataset, however, those could not be incorporated here for clustering due to dataset constraints. For example, a study[30] showed that annealing PLA to increase crystallinity from 0% to 46% significantly enhanced oxygen barrier performance, and that organically modified montmorillonite (OMMT) nanofillers further improved gas barrier properties. However, relative humidity data were not reported in the study, hence limiting comparability. Thus, values of this study were not included in our model. Similarly, another study[31] found that surfactant-modified cellulose nanocrystals with silver nanoparticles increased barrier effects up to 60%, but OP data lacked humidity conditions. Moreover, Rhim et al. (2009) demonstrated that organomodified nanoclays decreased WVP in PLA films, but oxygen permeability data were missing, this study was not included in the clustering model as well[32].

Similar to nanocomposite studies, coating strategies could also offer potential for enhancing barrier properties. For instance, PLA/ aluminum oxide (AlOx) paperboard[33] or soy protein isolate (SPI) films coated with PLA[34] showed barrier improvements but they were not captured in original TRANSMAT dataset for similar reasons mentioned before. These cases illustrate how methodological reporting gaps and the dataset's strict inclusion criteria can limit coverage, emphasizing the need for broader curation pipelines in future clustering studies while expanding the dataset.

Moreover, recent applications indirectly support the feasibility of this placement. For instance, for fruits, PLA composites incorporating tung oil derivatives significantly reduced WVTR and extended strawberry shelf life from 3–4 to 5 days[35], while PLA/essential-oil nanofiber systems suppressed fungal growth in table grapes for >10 days[36]. For vegetables, ultra-thin poly(L-lactic acid) (PLLA) films combined with poly(L-lactic acid-co-butyrate itaconate) (P(LA-BI)) films reduced OTR and improved cherry tomato preservation up to 28 days[36], and PLA and polybutylene adipate-co-terephthalate (Ecoflex®) films reinforced with lignocellulose nanofibers showed lower WVTR, suitable for fresh-cut lettuce[37]. For bakery products, bilayer PLA/konjac glucomannan/wheat gluten films with cinnamaldehyde reduced mold counts by up to 5.8 log CFU/g, extending bread shelf life from 3–4 to 10 days[38]. While many recent PLA-based packaging studies report limited OTR/WVTR data, they provide application trials showing extended shelf life. These studies demonstrate that PLA films can, in practice, serve in fruits, vegetables, salads

and bakery applications when combined with fillers, multilayers, or active agents. Thus, they help to validate that the current cluster captures a realistic functional overview for PLA between high-barrier gelatin nanocomposites and moisture-sensitive polysaccharide blends in the dataset. Within the Cluster 3 (green, n = 180), edible films formulated from carrot puree, carboxymethyl cellulose (CMC), corn starch and gelatin were studied (see **Supplementary Table 3**). Statistically, it is characterized by relatively low OTR values ($10.3 \pm 5.4$ cm³/m²·day) but extremely high WVTR values ($264,385.3 \pm 106,946.5$ g/m²·day). The clustering result therefore clearly separates these films from conventional polymers and nanocomposites (Clusters 1 and 2), reflecting their fundamentally different composition and water-rich matrix. DBSCAN clustering approach highlighted that the grouping is primarily driven by WVTR rather than OTR. At the same time, these results must be interpreted with caution. The dataset contains carrot blends materials in Cluster 3 with more than 100 data rows compared to smaller sample sizes in Cluster 1 and 2. This imbalance in this analysis may therefore reflect sample overrepresentation in the dataset rather than industrial relevance. While we recognize that imbalanced datasets can bias clustering outcomes, the goal of this study was not to deliver a fully validated predictive model but to demonstrate a methodological proof-of-concept for classifying sustainable materials based on gas permeability. Given the scarcity and heterogeneity of permeability data in the literature, we consider this an essential intermediate step.

Moreover, some permeability values were converted from permeability coefficients to transmission rates using an assumed film thickness of 100 μm and standardized measurement conditions (23 °C/50% RH for OTR; 25 °C/50% RH for WVTR). These choices are consistent with common practice but inevitably introduce uncertainty in the absolute positioning of samples in feature space. In extreme cases, especially for hygroscopic materials whose WVTR is strongly RH-dependent, cluster membership could shift if the true thickness or relative humidity deviate substantially from the assumed values. A systematic sensitivity analysis that perturbs thickness and experimentally well-characterized temperature/RH conditions, once richer metadata are available, is therefore a first priority for future work to quantitatively assess the robustness of the clustering outcome. Additionally, future studies should compare these blend films with broader categories of edible coatings (e.g., starch, protein, or alginate systems) to evaluate whether carrot purée behaves uniquely or simply represents a broader class of high-WVTR biopolymer coatings.

Beyond visual overlap, future research should also employ simple quantitative metrics to validate the alignment between clusters and food category requirements. For example, distances between cluster centroids and food-specific permeability ranges, or the fraction of samples falling within each food group's thresholds, could provide a more objective measure of correspondence. Such approaches were not feasible here due to dataset size and heterogeneity, but would become feasible once permeability requirements for food groups are available in standardized numerical ranges and the dataset contains sufficient, balanced data across material classes.

Overall, this study revealed that the density-based clustering algorithm, DBSCAN, was the most effective algorithm for grouping sustainable packaging materials based on gas permeability, with the highest Silhouette Score (0.900) and lowest Davies-Bouldin Index (0.388). Additionally, the results indicate that most sustainable packaging materials have high gas permeability, whereas nanocomposites and coated materials demonstrate improved barrier performance as observed within this limited dataset. This study serves as a proof-of-concept to illustrate the potential of clustering approaches for systematically classifying packaging materials. To strengthen practical relevance, future research should focus on careful data curation and integration of larger, more diverse datasets that capture advances beyond 2016. Standardized reporting of permeability conditions, material compositions, and processing parameters is essential to improve comparability. Ultimately, expanding this framework with broader datasets and advanced machine learning methods could bridge the gap between academic data and real-world application. While the present work demonstrates the method as a proof-of-concept research tool, we envision that with expanded datasets and integration of additional parameters, this framework could be further developed into a practical decision-support tool for food companies and packaging designers.

## 2.3 Research Limitations

The dataset provides initial insights for further sustainable materials studies, but it is not fully representative of the wide range of films currently reported in the literature or commercialized by industry. First, the available dataset is limited to 49 scientific articles selected based on the availability of structured OTR and WVTR values, following the TRANSMAT Ontological and Terminological Resource. While this ensured transparency and consistency, it excluded many materials for which data were not reported in a structured or standardized way. As a result, only a

small pool of materials and monomer sources were included, which restricts the generalizability of the findings.

We acknowledge that this limited dataset does not capture the hundreds of sustainable films developed by academia and industry, nor the breadth of processing technologies and testing conditions available. Furthermore, inconsistencies in unit reporting (e.g., cm³·mm/m²·24h·atm vs. g/m²·s·Pa) and measurement conditions (temperature, humidity) complicated direct comparisons. For instance, the dataset contains instances where identical values appear in columns, except for the "Original_Value" field. For example, the same "Doc," "Target," and "Type" entries linked with different permeability values in "Original_Value", such as $1.6$ $cm^3 \cdot mm/m^2 \cdot 24h \cdot atm$ and $0.16 \times 10^{-5}$ $cm^3 mm/m^2 day atm$ , making it challenging to accurately interpret the conditions under which these values were recorded. In such cases, further manual investigation into each original article is applied to resolve discrepancies. Although dataset creators questioned controlled parameter values for permeability measurements, some permeability values were missing or reported under different temperature and humidity conditions, which can further complicate clustering analysis and comparison. As a result, manual intervention or supplementary data processing required to address these inconsistencies.

Therefore, the present study should be interpreted as a methodological demonstration of clustering rather than a comprehensive survey of sustainable packaging materials. To improve data reliability and applicability, future studies must expand the dataset to include a wider range of materials, ensure standardized reporting of measurement conditions and units, and integrate industrially relevant data such as PET and PP alongside conventional sustainable options.

In addition, certain assumptions were necessary to harmonize the dataset where original publications lacked complete metadata. For example, if film thickness was not specified, a default of 100 μm was applied, and when test conditions were missing, standard values of 23 °C / 50% RH (for OTR) and 25 °C / 50% RH (for WVTR) were assumed. While these assumptions allowed comparability across studies, they introduce potential sources of bias, particularly for humidity-sensitive biopolymers or films. Since DBSCAN clustering emphasizes the relative distribution of data points rather than absolute values, we expect the overall cluster structure to remain robust. Nevertheless, cluster membership could shift under alternative assumptions for film thickness or

measurement conditions. Thus, a formal sensitivity analysis (e.g. varying the assumed thickness) should therefore be a priority for future work once more complete datasets with reported measurement conditions are available. This will help verify whether the cluster boundaries remain stable when key parameters are varied. Acknowledging these limitations is crucial to ensure transparent interpretation of the present findings as a methodological demonstration rather than a definitive classification of all sustainable films.

## 2.4 Research Outlook and Recommendation

The study shows that clustering algorithms can effectively classify sustainable materials. Future research could expand dataset and improve data quality to better connect material characteristics with food preservation needs. Incorporating a broader range of sustainable materials, including newly developed biopolymers and advanced coatings, and benchmark materials such as PET and PP alongside conventional sustainable options will enable more meaningful comparisons with industry-relevant packaging. Additionally, extending the dataset beyond 2016 to include recent advancements in material science will improve the relevance of the findings. Including more gas permeability data, as well as additional material properties such as mechanical strength, biodegradability and recyclability will enable a more balanced assessment of sustainable packaging performance. To further strengthen the robustness of clustering results, future datasets should minimize reliance on assumed values such as default thicknesses or standard testing conditions. Instead, efforts should be made to encourage standardized metadata reporting (e.g., film thickness, temperature, relative humidity) and harmonized units across publications. Such improvements will reduce uncertainty, eliminate the need for assumptions, and enable more rigorous sensitivity analyses. Ultimately, this will allow clustering outcomes to reflect true material performance under realistic conditions, thereby improving the applicability of the methodology for industrial packaging design.

Studies showed that nanocomposites have improved barrier properties and can be designed with sustainable materials to enhance performance. The use of active packaging technologies, such as antimicrobial coatings, may enhance sustainable packaging's functionality. Combining different sustainable materials or developing multilayered structures could improve barrier performance while maintaining environmental benefits. This study should be regarded as a proof-of-concept

demonstration of clustering methodology rather than a comprehensive survey of all available permeability data. The dataset used was limited in scope, but it provided a structured foundation to test whether clustering can serve as a useful classification tool. Future work should focus on curating and publishing a larger dataset that includes a wider variety of sustainable films, industrial benchmarks (e.g., PET, PP), and more recent data generated with modern characterization methods. Once richer datasets are available, future studies can apply this framework to optimize hybrid packaging approaches for food preservation.

Beyond permeability, real-world packaging choices are also driven by properties such as mechanical strength, optical clarity, printability, sealing performance, cost, and regulatory compliance. While these parameters were outside the scope of this proof-of-concept, the clustering framework we present can be extended to incorporate such multidimensional datasets. By integrating standardized metrics (e.g., tensile strength, recyclability indicators, cost indices, and compliance status with EFSA/FDA regulations), future iterations of the model could generate clusters that not only reflect barrier performance but also practical trade-offs relevant to industrial adoption. This would transform the methodology from an exploratory classification tool into a more holistic decision-support approach for sustainable packaging development. Using more advanced machine learning techniques, such as deep learning or reinforcement learning, could improve material classification and prediction accuracy. Additionally, integrating real-world performance data from packaging trials could enhance the model's robustness.

## 3. Methods

### 3.1. Data Description

This study utilizes a text-mined dataset compiled from 49 scientific articles[11] published between 2000 and 2016. The raw data, stored in four comma-separated value files, includes information on 7 variables (**Table 3**). Python 3.8 was used for all subsequent data cleaning, feature engineering, and analysis.

After merging and de-duplicating the four datasets, the final dataset contained 1,591 entries, including 1,154 symbolic, 403 quantitative, and 394 addimensional entries. Of these, 22 articles initially contributed 197 extractable gas permeability data points (oxygen, carbon dioxide, and water vapor), while the remaining articles provided primarily symbolic data, such as material types

and quantities of impact components in the dataset. To expand the dataset, we manually extracted additional values ('oxygen transmission rate' (OTR), 'water vapor transmission rate' (WVTR), 'temperature,' and 'relative humidity') from the remaining 27 articles, resulting in a total of 295 permeability data points. These data points correspond to a limited set of material categories: (i) gelatin-based films (including gelatin + nanoclays), (ii) conventional polymers (LDPE, HDPE, PLA), and (iii) carrot puree blends with starch, gelatin, or CMC (see Supplementary Tables 1–3). Carbon dioxide permeability was excluded due to the limited number of values reported. This clarification ensures that the subsequent clustering analysis is interpreted within the scope of a limited but structured dataset.

**Table 3.** Variables of raw data, including gas permeability characteristics. The article titled "barrier and surface properties of chitosan-coated greaseproof paper" reports a carbon dioxide permeability of 3400 (cm³ mm)/m²·atm·day as a quantity/count characteristic.

| Variable | Explanation | Values |
|---|---|---|
| Doc | Article title | Barrier and surface properties of chitosan-coated greaseproof paper |
| DOI | Digital Object Identifier | https://doi.org/10.1016/j.carbpol.2006.02.005 |
| Target | generic concept represented | Permeability |
| Type | Ontology concept category, symbolic, quantitative or addimensional | QUANTITY |
| Original_Value | a list of annotated tokens for symbolic data, two lists of annotated tokens for quantitative data | (['3400'], ['cm','^','3','mm','/','(','m','^',2', 'atm','day',')')']) |
| Attached_Value | a list of annotated tokens to disambiguate a measure unit when necessary for quantitative data. None for symbolic data. | ['carbon', 'dioxide'] |
| Annotator | annotator id | 1 |

The reported units for barrier properties were not standardized, provided as either permeability coefficients or transmission rates. Across all barrier properties, 34 unique units were identified, which were standardized during data preprocessing. For oxygen permeability, the most common coefficient unit was $cm^3 cm/m^2\,Pa\,s$, while for water vapor it was $g.\,mm/m^2\,day\,Pa$. Transmission rates were most often reported as cm³/m²·day (oxygen) and g/m²·day (water vapor).

The relationship between the oxygen permeability (OP) and the oxygen transmission rate (OTR) is expressed as[38]:

$$OP = \frac{OTR \times l}{\Delta P} \qquad (1)$$

where $l$ stands for film thickness and $\Delta P$ is pressure difference across the material.

However, since the barrier properties of packaging materials are often reported as transmission rates, the study used the equation (1) to convert permeability coefficients to transmission rates transmission rates for consistency with the earlier study[15]. Moreover, some inconsistencies in reporting were identified in several studies (e.g., "cm³·cm/m²·s" reported as a transmission rate). In such cases, the reported value was reinterpreted as a permeability data. All final data were standardized to transmission rates with day as the time unit, enabling consistent comparison across materials and studies. This standardization supported clustering analysis of sustainable packaging materials based on their barrier performance for different food applications.

Measurement conditions have a significant effect on gas permeability performance. To assess this impact, the distributions of temperature and relative humidity (RH) were analyzed (**Table 4**). For oxygen transmission rate (OTR), the most commonly used test temperature was 23°C, while for water vapor transmission rate (WVTR), it was 25°C. Additionally, 50% RH was the most applied condition in both permeabilities, aligning with standard measurement practices for permeability (Sánchez-Tamayo et al., 2020). These conditions are consistent with established test methods, such as ASTM D3985 for OTR and ASTM E96 for WVTR, so that all subsequent analyses were performed under these standardized conditions: 23 °C and 50% RH for OTR, and 25 °C and 50% RH for WVTR. When measurement conditions were missing, these standard values were assumed.

This filtering ensured the reliability and comparability of the results, as measurement variations can significantly affect the gas barrier properties of packaging materials[38].

**Table 4.** Most commonly reported measurement conditions for oxygen transmission rate (OTR) and water vapor transmission rate (WVTR). The most frequent conditions were 23 °C and 50% RH for OTR, and 25 °C and 50% RH for WVTR. *A significant portion of data lacked measurement condition details. To ensure comparability, the dataset was screened and standardized to these conditions; when measurement conditions were missing, these values were assumed for subsequent analyses.*

| OTR | | | | WVTR | | | |
|---|---|---|---|---|---|---|---|
| Temp (°C) | Quantity | RH (%) | Quantity | Temp (°C) | Quantity | RH (%) | Quantity |
| 23 | 59 | 50 | 46 | 25 | 87 | 50 | 48 |
| 25 | 26 | 0 | 26 | 20 | 15 | 90 | 14 |
| 20 | 20 | 57 | 4 | 23 | 9 | 75 | 11 |
| missing | 8 | missing | 39 | missing | 22 | missing | 37 |

### 3.2. Data preprocessing

Although the raw dataset was relatively clean, some duplicate data was identified, and permeability units were not standardized yet. To address this, duplicates were removed, and transmission rate values for oxygen and water vapor were standardized before entering training step. If permeability was reported as a permeability value, it was converted into a transmission rate by multiplying by the material thickness. If thickness was not specified, a default value of 100 μm was assumed. Finally, oxygen transmission rate (OTR) was standardized to $cm^3/m^2day$ and water vapor transmission rate (WVTR) to $g/m^2day$. Data processing flowchart is given in **Figure 3**

Feature engineering consists of text preprocessing, natural language process and lastly feature selection. The purpose of feature engineering is to extract the materials and type of the packaging materials from the article title in variable 'Doc'. Before performing any analysis, the text data in the 'Doc' column underwent cleaning to remove noise such as punctuation, special characters, and irrelevant symbols. This step ensured that the text is cleaner and easier to process. The cleaned text is then converted to lowercase for consistency.

To process the unstructured text data in the "Doc" column, several Natural Language Processing (NLP) techniques are applied to identify key components such as sustainable packaging materials, base materials, packaging forms, and modified materials. As the majority of "Doc" entries describe the barrier properties of specific sustainable materials, the analysis begins by refining the text. The following steps are implemented: (1) Preprocessing Text for Parsing: When the word "of" appears in an article title, the text following its first occurrence is extracted. This extracted segment is then analyzed using the SpaCy 3.7.8. library to determine its grammatical structure and syntactic dependencies. This process aids in identifying key components such as base materials and modification type and (2) Dependency Parsing (DP): Dependency Parsing analyzes the grammatical structure of a sentence to determine the relationships between words and their roles in the text[39]. This method was effective for extracting key components in scientific article titles, which typically follow structured grammatical patterns. After parsing, several rules were set to extract key components.

Rules for extracting key components was dependent on base materials: (1) Keyword matching: if a noun or proper noun matches predefined keywords such as "starch," "chitosan," or "zein," it is identified as the base material. (2) Film/Films: if the word "film" or "films" is present and no base material has been identified, the noun immediately preceding "film" or the direct object (dobj) of "film" is set as the base material. If no preceding noun is available, the function looks for modifiers (e.g., adjectives) to refine the base material, such as in "chitosan films." (3) "Based" Clause: if the word "based" appears in the text, its associated modifier is treated as the base material (e.g., "starch-based").

If terms like "nanocomposite" or "coated/coatings" are detected, they are classified as the material type. Keywords such as "montmorillonite," "clay," and "silica" are used to identify secondary materials. If any of these keywords are found, the associated term is extracted as the secondary material.

After text preprocessing and feature extraction, three key features are extracted from the 'Doc' column, namely based material, modification type, secondary material. For entries where these features could not be automatically extracted using the predefined extraction rules, manual verification was conducted to ensure the correct identification of the relevant features. In order to

make comparisons with the gas permeability requirements for various food groups identified in previous studies[15], the logarithms of the OTR and WVTR were selected as the primary features for clustering. Although all permeability data were first standardized into transmission rates, the values still spanned several orders of magnitude across different materials. The logarithmic transformation reduced this skewness, normalized the scale, and prevented clustering algorithms from being dominated by extreme values, thereby improving the consistency of the clustering analysis. Further standardization by measurement conditions was applied to the dataset before clustering. Specifically, OTR values were screened to 23 °C and 50% RH and WVTR values to 25 °C and 50% RH, reflecting the most commonly reported test conditions (see **Table 4**). This standardization step was essential to reduce variability across studies and ensure comparability between materials prior to clustering. From the initial 295 data points extracted from the literature, 249 were retained after preprocessing for clustering, while the remaining data points were considered noise or excluded due to incomplete measurement conditions or values.

## 3.3. Research Models

K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Model (GMM) were applied to identify distinct groups of packaging materials based on their gas permeability features. Clustering is an unsupervised learning method that groups objects according to similarities in their attributes. Clustering techniques can be categorized into hierarchical, partition-based, density-based, grid-based and model-based algorithms[40]. Since the present dataset is neither high-dimensional nor hierarchical, grid-based and hierarchical algorithms were not considered. The goal of clustering is to maximize intracluster similarity and minimize intercluster similarity.

K-Means is a partition-based clustering algorithm that groups data points into k-clusters using Euclidean distance between points, with k specified as a prerequisite. The optimal number of clusters can be determined using the elbow method or silhouette score[41]. The algorithm minimizes the sum of squared errors for each cluster[42] as represented

$$\min(X) = \sum_{k=1}^{K} \sum_{x_i \in c_k} ||x_i - \mu_k||^2 \qquad (2)$$

where $x_i$ is data point in cluster $c_k$, $\mu_k$ is the mean of the cluster and K is the number of clusters.

While computationally efficient, K-Means assumes clusters are spherical and equidistant, which may not fit all datasets, including those with non-centroid distributions.

Gaussian Mixture Model (GMM) is a model-based clustering method that estimates parameters of multiple Gaussian distribution in the dataset[40]. For each of the k components, the parameters include: (i) Means ($\mu_k$) the center of the cluster, (ii) Covariance ($\Sigma_k$) shape and orientation of the cluster, (iii) Mixing Coefficient ($\pi_k$) weight of each Gaussian component (sum = 1)[43]. The parameter of GMM were estimated by using Expectation-Maximization (EM) algorithm, which iteratively refines the model parameters through: (i) Expectation (E-step) assigns probabilities for each data point belonging to a cluster, (ii) Maximization (M-step) updates the parameters to maximize the likelihood[44]. A pre-defined cluster number k is necessary for K-Means and GMM, and elbow method is widely used to identify cluster number k of K-Means Clustering[45]. However, the elbow point was ambiguous in the result, so Silhouette Score was applied to determine the optimal clustering number[46] in both K-Means and GMM as shown in **Supplementary Figure 3**. Unlike K-Means, GMM does not require spherical clusters and can model clusters of varying shapes, sizes, and densities, making it more flexible for complex data distributions[47].

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm that does not require a predefined number of clusters as K-Means and Gaussian Mixture Model. It identifies clusters of arbitrary shapes and can detect outliers or noise points[40]. Key parameters of DBSCAN include (i) Epsilon ($\epsilon$) maximum radius of a neighborhood and (ii) MinPts minimum number of points required to form a cluster. This method is particularly suited for datasets where clusters are not centroid based. As for DBSCAN, according to rule of thumb, the MinPts should larger than dimension 1 and it's better to be twice as the dimension[48]. Since the dimension of the data is 2, with only OTR and WVTR, the minimum points was 4. According to k-distance result, epsilon was set as elbow point, which was 1.17 (**Supplementary Figure 4**).

**3.4. Model Evaluation**

Silhouette and Davies-Bouldin index and visual domain evaluation are used to determine the performance of the algorithm. Higher Silhouette Score and a lower Davies-Bouldin Index indicate better clustering performance. The Silhouette Score is calculated[49] as:

$$s = \frac{b - a}{\max(a, \ b)} \qquad (3)$$

where a represents the average distance between a sample and all other points in the same cluster, while b is the average distance between a sample and all other points in the next nearest cluster. The Davies-Bouldin index is defined as:

$$db = \frac{1}{k} \sum_{i=1}^{k} max \left( \frac{\Delta X_i + \Delta X_j}{\delta(X_i, X_j)} \right) \qquad (4)$$

where $\Delta X_i$ is the distance within the cluster and $\delta(X_i, X_j)$ is the distance between the cluster $i$ and cluster $j$.

Silhouette Score can measure how well each material fits within its assigned cluster compared to other clusters while Davies-Bouldin Index can assess the similarity between clusters and the separation between them[50].

After clustering, the results were visualized using scatter plots to map oxygen and water vapor transmission rates. These visualizations facilitate direct comparisons between the gas transmission rate characteristics of sustainable packaging materials and the specific barrier requirements of eight food groups as identified in previous studies[15].

**Acknowledgement**

**Data availability**

Raw data source from Lentschat study can access via CIRAD open access portal https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/U7HK8J

Processed data can be accessed in supplementary Table 4.

**Code availability**

Processed dataset package is available and code can be accessed via GitHub repository

https://github.com/whps0620/Food-Pack-Mapper/tree/main and in supplementary material.

**Competing interests**

The authors declare no competing interests.

**AUTHOR CONTRIBUTIONS**

D.T. conceptualized the study and did the project administration; D.T. performed the investigation, T.Y.Y. developed the methodology, and executed the model; D.T developed the paper concept and T.Y.Y., D.T. wrote the original draft of the paper; T.Y.Y. did the visualization; D.T. performed critical review and editing.

**Declaration of generative AI and AI-assisted technologies in the writing process**

During the preparation of this work, the authors used Grammarly (full paper) in order to improve the spelling, grammar, and style of the text. No additional original content was generated using these AI-assisted technologies. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

# References

1 Marsh, K. & Bugusu, B. Food packaging—roles, materials, and environmental issues. *Journal of food science* **72**, R39-R55 (2007).

2 Choudalakis, G. & Gotsis, A. Permeability of polymer/clay nanocomposites: A review. *European polymer journal* **45**, 967-984 (2009).

3 Tyagi, P., Salem, K. S., Hubbe, M. A. & Pal, L. Advances in barrier coatings and film technologies for achieving sustainable packaging of food products–a review. *Trends in Food Science & Technology* **115**, 461-485 (2021).

4 Chen, Y. & Li, Y. Determination of water vapor transmission rate (WVTR) of HDPE bottles for pharmaceutical products. *International journal of pharmaceutics* **358**, 137-143 (2008).

5 Jouppila, K. & Roos, Y. Water sorption and time-dependent phenomena of milk powders. *Journal of dairy science* **77**, 1798-1808 (1994).

6 Geijer, T. Plastic packaging in the food sector. *Six ways to tackle the plastic puzzle. ING THINK Economic and Financial Analysis* (2019).

7 Defruyt, S. Towards a new plastics economy. *Field actions Science reports. The journal of field actions*, 78-81 (2019).

8 Parliament, E. Regulation (EU) 2025/40 of the European Parliament and of the Council of 19 December 2024 on Packaging and Packaging Waste, Amending Regulation (EU) 2019/1020 and Directive (EU) 2019/904, and Repealing Directive 94/62/EC. *Official Journal of the European Union* (2024).

9 Ncube, L. K., Ude, A. U., Ogunmuyiwa, E. N., Zulkifli, R. & Beas, I. N. Environmental impact of food packaging materials: A review of contemporary development from conventional plastics to polylactic acid based materials. *Materials* **13**, 4994 (2020).

10 Sangroniz, A. *et al.* Packaging materials with desired mechanical and barrier properties and full chemical recyclability. *Nature Communications* **10**, 3559 (2019).

11 Lentschat, M., Buche, P., Dibie-Barthelemy, J., Menut, L. & Roche, M. Food packaging permeability and composition dataset dedicated to text-mining. *Data in Brief* **36**, 107135 (2021).

12 Guillard, V. *et al.* The next generation of sustainable food packaging to preserve our environment in a circular economy context. *Frontiers in nutrition* **5**, 121 (2018).

13 Schaarsberg, N. G. & van de Stadt, K. in *24 Th IAPRI World Packaging Conference.*

14 Rankin, J., Wolff, I., Davis, H. & Rist, C. Permeability of Amylose Film to Moisture Vapor, Selected Organic Vapors, and the Common Gases. *Industrial & Engineering Chemistry Chemical and Engineering Data Series* **3**, 120-123 (1958).

15 Stocchetti, G. Technology that bridges the gap. *Packaging Films* **3**, 16-18 (2012).

16 Makino, Y. & Hirata, T. Modified atmosphere packaging of fresh produce with a biodegradable laminate of chitosan-cellulose and polycaprolactone. *Postharvest Biology and Technology* **10**, 247-254 (1997).

17 Strantz, A. & Zottola, E. Bacterial survival on lean beef and Bologna wrapped with cornstarch-containing polyethylene film. *Journal of food protection* **55**, 782-786 (1992).

18 Ester, M., Kriegel, H. P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).  226-231 (1996).

19    Pezer, D. Application of Cluster Analysis for Polymer Classification According to Mechanical Properties. *Technium* **3**, 67-75 (2021).

20    Stadlthanner, D., Steinkellner, H., Landschützer, C. & Kaever, D. A hierarchical density-based clustering method applied to mixed-mail in Austria. *Logistics Research* **17,** 1-17 (2024).

21    Wessling, M. *et al.* Modelling the permeability of polymers: a neural network approach. *Journal of membrane science* **86**, 193-198 (1994).

22    Phan, B. K. *et al.* Gas permeability, diffusivity, and solubility in polymers: Simulation-experiment data fusion and multi-task machine learning. *npj Computational Materials* **10**, 186 (2024).

23    Wu, F., Misra, M. & Mohanty, A. K. Challenges and new opportunities on barrier performance of biodegradable polymers for sustainable packaging. *Progress in Polymer Science* **117**, 101395 (2021).

24    Bae, H. J. *et al.* Effect of clay content, homogenization RPM, pH, and ultrasonication on mechanical and barrier properties of fish gelatin/montmorillonite nanocomposite films. *LWT-Food Science and Technology* **42**, 1179-1186 (2009).

25    Sharma, S. *et al.* Active film packaging based on bio-nanocomposite TiO2 and cinnamon essential oil for enhanced preservation of cheese quality. *Food Chemistry* **405**, 134798 (2023).

26    Peter, A. *et al.* Chemical and organoleptic changes of curd cheese stored in new and reused active packaging systems made of Ag-graphene-TiO2-PLA. *Food Chemistry* **363**, 130341 (2021).

27    Ligaj, M., Tichoniuk, M., Cierpiszewski, R. & Foltynowicz, Z. Efficiency of novel antimicrobial coating based on iron nanoparticles for dairy products' packaging. *Coatings* **10**, 156 (2020).

28    European Commission.    (2009).

29    Park, H. M. *et al.* Preparation and properties of biodegradable thermoplastic starch/clay hybrids. *Macromolecular Materials and Engineering* **287**, 553-558 (2002).

30    Picard, E., Espuche, E. & Fulchiron, R. Effect of an organo-modified montmorillonite on PLA crystallization and gas barrier properties. *Applied Clay Science* **53**, 58-65 (2011).

31    Fortunati, E., Peltzer, M., Armentano, I., Jiménez, A. & Kenny, J. M. Combined effects of cellulose nanocrystals and silver nanoparticles on the barrier and migration properties of PLA nano-biocomposites. *Journal of Food Engineering* **118**, 117-124 (2013).

32    Rhim, J.-W., Hong, S.-I. & Ha, C.-S. Tensile, water vapor barrier and antimicrobial properties of PLA/nanoclay composite films. *LWT-Food Science and Technology* **42**, 612-617 (2009).

33    Peelman, N. *et al.* Application of bioplastics for food packaging. *Trends in Food Science & Technology* **32**, 128-141 (2013).

34    Rhim, J.-W., Lee, J. H. & Ng, P. K. Mechanical and barrier properties of biodegradable soy protein isolate-based films coated with polylactic acid. *LWT-Food Science and Technology* **40**, 232-238 (2007).

35    Zong, Y. *et al.* Fabrication of antimicrobial and high-toughness poly (lactic acid) composite films using tung oil derivatives. *International Journal of Biological Macromolecules* **254**, 127792 (2024).

36      Brandão, R. M. *et al.* Active packaging of poly (lactic acid) nanofibers and essential oils with antifungal action on table grapes. *FEMS Microbiology Letters* **369**, fnac116 (2022).

37      Bascón-Villegas, I. *et al.* A new eco-friendly packaging system incorporating lignocellulose nanofibres from agri-food residues applied to fresh-cut lettuce. *Journal of Cleaner Production* **372**, 133597 (2022).

38      Wang, L., Zhang, Y., Xing, Q., Xu, J. & Li, L. Quality and microbial diversity of homemade bread packaged in cinnamaldehyde loaded poly (lactic acid)/konjac glucomannan/wheat gluten bilayer film during storage. *Food Chemistry* **402**, 134259 (2023).

39      Nivre, J. Dependency parsing. *Language and Linguistics Compass* **4**, 138-152 (2010).

40      Mehta, V., Bawa, S. & Singh, J. Analytical review of clustering techniques and proximity measures. *Artificial Intelligence Review* **53**, 5995-6023 (2020).

41      Nanjundan, S., Sankaran, S., Arjun, C. & Anand, G. P. Identifying the number of clusters for K-Means: A hypersphere density based approach. *arXiv preprint arXiv:1912.00643* (2019).

42      Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B. & Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences* **622**, 178-210 (2023).

43      Quiñones-Grueiro, M., Prieto-Moreno, A., Verde, C. & Llanes-Santiago, O. Data-driven monitoring of multimode continuous processes: A review. *Chemometrics and Intelligent Laboratory Systems* **189**, 56-71 (2019).

44      Xuan, G., Zhang, W. & Chai, P. in *Proceedings 2001 international conference on image processing (Cat. No. 01CH37205).* 145-148 (IEEE).

45      Cui, M. Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance* **1**, 5-8 (2020).

46      Shahapure, K. R. & Nicholas, C. in *2020 IEEE 7th international conference on data science and advanced analytics (DSAA).* 747-748 (IEEE).

47      Vila, J. P. & Schniter, P. Expectation-maximization Gaussian-mixture approximate message passing. *IEEE Transactions on Signal Processing* **61**, 4658-4672 (2013).

48      Giri, K., Biswas, T. K., & Sarkar, P. ECR-DBSCAN: An improved DBSCAN based on computational geometry. *Machine Learning with Applications*, *6*, 100148 (2021).

49      Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53-65 (1987).

50      Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 224-227 (2009).

**Figure and Table Legends**

**Figure 1.** Distribution of oxygen transmission rate (OTR) and water vapor transmission rate (WVTR), both exhibiting skewness, highlighting the need for standardization before modeling

**Figure 2.** Clusters based on log-transformed OTR (oxygen transmission rate) and WVTR (water vapor transmission rate) values identified by DBSCAN, color-coded as follows: Cluster 1 (Purple), Cluster 2 (Orange), Cluster 3 (Green). Circular markers (●) represent sustainable materials; triangular markers (▲) represent conventional plastics (LDPE, HDPE), whereas grey circular markers (●) indicate noise. Because most food groups demand very low WVTR and OTR, most sustainable materials in this study do not satisfy those requirements, leading to limited clustering overlap. Adapted from[15]

**Figure 3.** Data processing flowchart

**Table 1.** Comparison of clustering algorithms (K-Means, DBSCAN, and GMM) based on Silhouette Score and Davies-Bouldin Index before and after standardization by measurement conditions for OTR and WVTR to 23 °C/50% RH and 25 °C/50% RH, respectively.

**Table 2.** Summary of DBSCAN cluster outcomes, showing the average (± standard deviation) OTR ($cm^3/m^2day$) at 23 °C and 50% RH, and WVTR ($g/m^2day$) at 25 °C and 50% RH for assumed thickness 100 μm in each cluster, along with their dominant base material, type, and secondary material
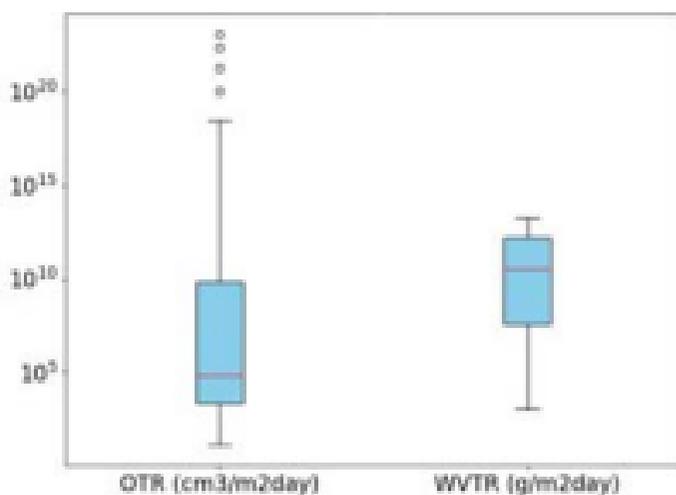
**Table 3.** Variables of raw data, including gas permeability characteristics. The article titled "barrier and surface properties of chitosan-coated greaseproof paper" reports a carbon dioxide permeability of 3400 ($cm^3$ mm)/m²·atm·day as a quantity/count characteristic.
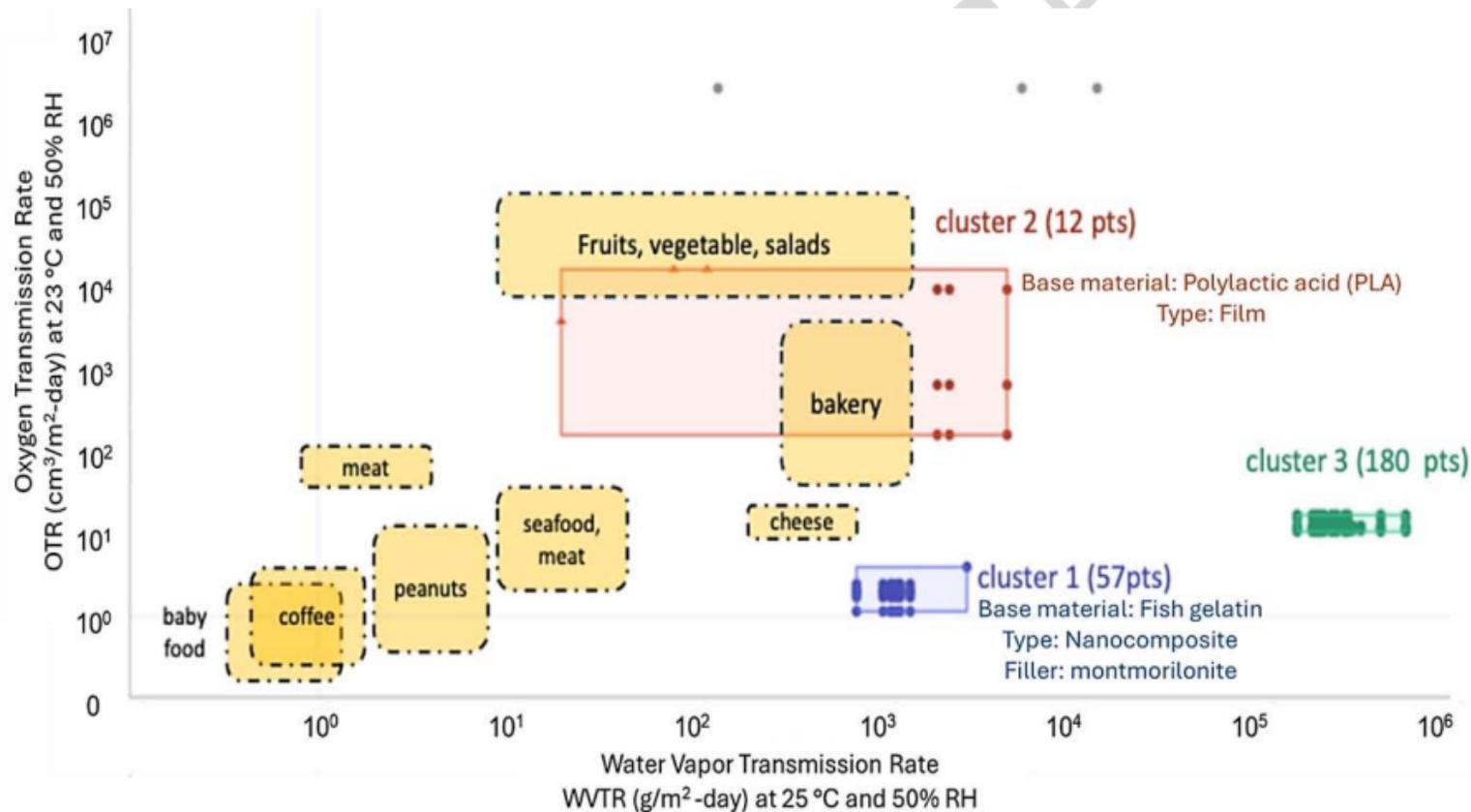
**Table 4.** Most commonly reported measurement conditions for oxygen transmission rate (OTR) and water vapor transmission rate (WVTR). The most frequent conditions were 23 °C and 50% RH for OTR, and 25 °C and 50% RH for WVTR. *A significant portion of data lacked measurement condition details. To ensure comparability, the dataset was screened*

*and standardized to these conditions; when measurement conditions were missing, these values were assumed for subsequent analyses.*

Clustering Algorithm

Data → Preprocessing → Feature Engineering → [ K-Means / DBSCAN / GMM ] → Visualization