International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 5, Sep.-Oct., 2023 pp: 78-88 www.ijemh.com

Limitations of Near-Infrared Spectroscopy Dataset of *Panax Notoginseng*: Private Datasets Challenges and Solutions

Xuefeng Cheng[1,2], Abudhahir Buhari[2], Hongmei Zhu[1], Tadiwa Elisha Nyamasvisva[2], Li Zhenxiang[2], Jin Yuanrong[2], and Juan Liu[3]
[1]*School of Big Data and Information Industry, Chongqing City Management College, Chongqing, China*
[2]*Faculty of Engineering, Science & Technology, Infrastructure University Kuala Lumpur, Malaysia*
[3]*School of Artificial Intelligence, Chongqing University of Education, Chongqing, China*.
Corresponding Author: Juan Liu, liujuan@cque.edu.cn.

Date of Submission: 09-09-2023 Date of Acceptance: 22-09-2023

ABSTRACT: This paper addresses a crucial issue in geographical origin research of *Panax notoginseng*, which is the prevalence of private near-infrared spectroscopy datasets. Currently, researchers often collect data independently, including the collection of *Panax notoginseng* samples and the generation of near-infrared spectroscopy data. However, these private datasets are often not publicly shared, posing challenges to classification studies related to geographical origin of *Panax notoginseng*. This paper begins by introducing the significance of this field and the application of near-infrared spectroscopy technology. It then delves into the problems associated with private datasets, including potential issues during data collection, challenges in data quality, and concerns regarding data privacy and security. To address these issues, principles for constructing datasets are proposed, including guidelines for sample selection, instrument calibration, data preprocessing, and cleaning processes. Through this paper, the aim is to draw researchers' attention to the issue of private datasets, provide solutions and guidance, promote further development in geographical origin research of *Panax notoginseng*, ensure the reliability and comparability of research, and underscore the importance of data sharing.

KEYWORDS: *Panax notoginseng*, geographical origin, private datasets, data mining models, data sharing, data quality.

I. INTRODUCTION

Panax notoginseng, a valuable medicinal plant, is highly esteemed in the field of traditional Chinese medicine for its wide-ranging therapeutic properties[1]. As research on the geographical origin and quality of *Panax notoginseng* deepens, near-infrared spectroscopy technology has emerged as a pivotal tool. Near-infrared spectroscopy allows for non-destructive acquisition of comprehensive information about *Panax notoginseng* samples, including chemical composition and quality characteristics, making it widely applicable in geographical origin research of *Panax notoginseng*[2].

Research on the geographical origin of *Panax notoginseng* holds significant importance for scientists and the traditional Chinese medicine industry. *Panax notoginseng*, recognized as a precious Chinese medicinal herb, has been extensively utilized in traditional Chinese medicine due to its various pharmacologically active components, such as saponins, triterpene saponins, and flavonoids[3]. These components offer notable health benefits, including anti-inflammatory, antioxidant, and anticoagulant effects. The efficacy of *Panax notoginseng* is closely tied to its geographical origin, with variations in chemical composition and efficacy observed among specimens from different regions[4]. Therefore, investigating the geographical origin of *Panax notoginseng* is crucial for guiding the rational collection and quality control of medicinal herbs, enhancing the efficacy and safety of *Panax notoginseng* products.

Simultaneously, with the continual expansion of the global traditional Chinese medicine market, the demand for *Panax notoginseng* is on the rise[5]. However, variations in quality may exist among *Panax notoginseng* from different regions, influenced by environmental factors and cultivation conditions. Hence, studying the geographical origin of *Panax notoginseng* allows for a better understanding of the quality characteristics of specimens from diverse regions, facilitating market regulation and consumer choices.

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 5, Sep.-Oct., 2023 pp: 78-88 www.ijemh.com

Near-infrared spectroscopy technology has found extensive applications in various domains, including food, pharmaceuticals, and agriculture. Its non-destructive and efficient nature makes it an ideal tool for researching *Panax notoginseng*. Through near-infrared spectroscopy, rapid acquisition of chemical information in *Panax notoginseng* samples, encompassing organic components, trace elements, and bioactive substances, becomes feasible. These data can be employed for differentiating *Panax notoginseng* from various geographical origins and assessing its quality characteristics, thereby aiding in geographical origin determination[6].

Furthermore, near-infrared spectroscopy can be utilized for swiftly screening the quality of *Panax notoginseng* medicinal herbs, ensuring their compliance with standards. This holds paramount importance for market regulation and consumer health in the traditional Chinese medicinal herb sector. Therefore, the application of near-infrared spectroscopy technology offers an efficient and accurate analytical tool for geographical origin research of *Panax notoginseng*[7].

Despite the growing significance of geographical origin research of *Panax notoginseng*, it faces a significant challenge—the proliferation of private datasets[8]. Researchers often autonomously collect data, encompassing the collection of *Panax notoginseng* samples and the generation of near-infrared spectroscopy data. However, these datasets are frequently not publicly shared. This trend has adverse repercussions, restricting data accessibility and usability, hindering comparability and replicability of research.

The existence of private datasets implies that it is challenging for other researchers to assess the quality and reliability of this data, as well as to compare differences among various datasets. Additionally, the proliferation of private datasets raises concerns about data privacy and security, necessitating more regulations and solutions to protect sensitive information[9]. Hence, addressing the issue of private datasets is crucial for geographical origin research of *Panax notoginseng* and necessitates action to promote data sharing and collaboration.

The primary objective of this paper is to delve into the issue of private datasets in geographical origin research of *Panax notoginseng* and provide solutions and guidance. To achieve this goal, this paper will be organized as follows: firstly, a comprehensive discussion of the challenges posed by private datasets, including potential issues during data collection and challenges in data quality, will be undertaken. Strategies to overcome these issues and enhance data availability and reliability will be explored. Secondly, principles for constructing datasets, encompassing sample selection, instrument calibration, data preprocessing, and cleaning processes, will be examined. These principles will aid researchers in creating more accurate and consistent datasets. Lastly, the conclusion section will summarize the core findings of this paper and provide recommendations for future geographical origin research of *Panax notoginseng*, with a particular emphasis on the importance of data sharing to advance the research field.

II. CURRENT STATE OF PRIVATE DATASETS

In the field of geographical origin research of *Panax notoginseng*, there is a noticeable trend where researchers are increasingly inclined to autonomously collect data rather than relying on publicly available data resources[10]. This trend is largely influenced by several factors:

Firstly, geographical origin research of *Panax notoginseng* typically demands extensive datasets, comprising a significant number of *Panax notoginseng* samples and their corresponding near-infrared spectroscopy data. Given the wide range of medicinal uses of *Panax notoginseng*, researchers need to encompass data from multiple origins and different batches to ensure comprehensiveness and reliability in their studies. In such cases, autonomous data collection appears to be a reasonable choice as it allows for the collection of samples and data tailored to specific research needs.

Secondly, researchers place a high premium on data quality and control. The quality of near-infrared spectroscopy data directly impacts the reliability of research findings. Researchers are often more inclined to collect data themselves to ensure the accuracy and consistency of the data collection process. Additionally, they can calibrate and fine-tune instruments as needed to enhance data credibility.

However, this trend of autonomous data collection poses several challenges. Firstly, it increases the time and resource costs associated with research work. Data collection and processing are time-consuming and labor-intensive tasks that require a significant investment in manpower and equipment. Secondly, autonomous datasets are often closed, making it difficult for other researchers to access them. This leads to data confinement and opacity, limiting comparability and replicability of research. Lastly, autonomous datasets may be influenced by selection bias, as researchers may be

more inclined to collect readily available samples, overlooking potentially valuable data sources.

Despite the substantial data requirements for geographical origin research of *Panax notoginseng*, there are currently almost no publicly available *Panax notoginseng* near-infrared spectroscopy datasets[11]. This implies that researchers conducting geographical origin classification studies on *Panax notoginseng* have limited access to public data resources for reference and comparison. Several factors contribute to this phenomenon:

Firstly, geographical origin research of *Panax notoginseng* is a relatively new field and has not yet established the tradition of large-scale data sharing. In contrast, some other fields, such as genomics and bioinformatics, have developed extensive public databases to support scientific research. However, the field of geographical origin research of *Panax notoginseng* has not reached this level of data sharing.

Secondly, acquiring and processing data requires significant resources and technical expertise. Generating near-infrared spectroscopy data necessitates specialized instruments and trained, experienced operators, increasing the threshold for data sharing. Furthermore, data cleaning and calibration require specialized knowledge to ensure data quality and accuracy. These factors make data creation and maintenance an expensive undertaking.

Lastly, data privacy and security concerns are also challenges faced in data sharing. Near-infrared spectroscopy data may contain proprietary or sensitive information, making researchers reluctant to make it public. In terms of data sharing, clear privacy protection policies and security measures need to be established to ensure data safety.

The proliferation of private datasets and the lack of publicly available datasets have had a notable impact on geographical origin research of *Panax notoginseng*. Firstly, the presence of private datasets means that it is challenging for other researchers to assess the quality and reliability of this data. Since data generation and processing are typically conducted by individual research teams, external assessors struggle to obtain sufficient information to evaluate data accuracy and consistency[12]. This increases uncertainty in research and diminishes the credibility of research results.

Secondly, the prevalence of private datasets restricts comparisons and amalgamation of different datasets. If different research teams employ varying data collection methods and data processing procedures, integrating their data into a unified analytical framework becomes challenging. This limits the broad application and comprehensive analysis of research, hindering in-depth exploration of *Panax notoginseng* geographical origin issues.

Additionally, the closed nature of private datasets fosters a lack of collaboration and a culture of knowledge sharing within the field of geographical origin research of *Panax notoginseng*. Researchers often focus solely on their own datasets, disregarding other potential sources of valuable data for their research. This restricts collaboration and communication within the field, potentially impeding research progress[13].

The proliferation of private datasets and the dearth of publicly available datasets pose challenges to geographical origin research of *Panax notoginseng*, limiting comparability, replicability, and comprehensiveness of research. Addressing this issue requires measures to encourage data sharing and collaboration, promoting further development in the research field. In the following sections, we will explore how to overcome the challenges posed by private datasets and enhance data availability and credibility.

III. NEAR-INFRARED SPECTROSCOPY DATASET OF *PANAX NOTOGINSENG*

During the process of collecting *Panax notoginseng* near-infrared spectroscopy data, potential issues related to sample selection can arise. Different research teams or individual researchers may opt for various *Panax notoginseng* samples for near-infrared spectroscopy analysis, influenced by factors such as sample availability, cost, and research objectives. This inconsistency in sample selection can result in data imbalance, where certain geographical origins or specific sample types dominate the dataset, while others are overlooked. This bias may introduce prejudice, impacting the reliability and consistency of the research[14].

The dataset size plays a crucial role in model training and generalization as shown in Table 1. A small dataset may lead to overfitting, where the model performs well on training data but fails to generalize to unseen samples. Conversely, a large and diverse dataset enhances the model's ability to capture the underlying patterns and variations related to geographical origin identification.

Sample distribution is another critical factor. Ensuring a representative distribution of samples from different geographical origins helps to

capture the inherent variations in *Panax notoginseng* related to different regions. This consideration is essential to avoid biased or skewed models that may only excel in identifying a particular geographical origin while struggling with others.

TABLE 1: LIMITATIONS OF THE CURRENT DATASET OF *PANAX NOTOGINSENG*

NO	SAMPLES	METHODS	SPONSORS	EVALUATION
1	210 *Panax notoginseng* samples from five cities in Yunnan. Honghe--37 Kunming--52 Wenshan--46 Yuxi--49 Qujing--26	Fourier transform-mid infrared spectra were recorded within the spectra range of 4000-400 cm^{-1} with 4 cm^{-1} resolution, and each sample was scanned 64 times. Near-infrared spectra were recorded as 64 scans in the spectra range of 10000-4000 cm^{-1} with 4 cm^{-1} resolution.	Yunnan Provincial Traditional Chinese Medicine Joint Project	i. The distribution of samples in the dataset was imbalance.
2	93 *Panax Potoginseng* samples from five counties in Wenshan, Yunnan. Yanshan--27 Xichou--18 Maguan--20 Qiubei--28	HPLC. Near-infrared spectra were recorded as 64 scans in the spectra range of 10000-4000 cm^{-1} with 4 cm^{-1} resolution.	Natural Science Foundation of Yunnan Province of China	i. The dataset was few samples. ii. The data acquisition was complex, costly, and time-consuming.
3	359 Panax notoginseng samples were collected from four origins in Yunnan province, Honghe, Kunming, Qujing, and Wenshan.	HPLC. Attenuated total reflectance-Fourier transform infrared spectra were recorded as 16 scans in the spectra range of 4000-400 cm^{-1} with 4 cm^{-1} resolution.	National Natural Science Foundation of China	i. The distribution of samples from different origins is not given.
4	192 *Panax notoginseng* samples from four origins. Wenshan –48 Panxian—48 Yongzhou—48 Guilin--48	Terahertz time-domain spectroscopy, the femtosecond laser has a central wavelength of 780 nm, a pulse width of 100 fs, a repetition frequency of 80 MHz, and an average output power of 140 mW.	Natural Science Foundation of Guangxi, National Natural Science Foundation of China, Guangxi Key Laboratory of Automatic Detecting Technology and Instruments	i. The dataset was few samples. ii. The feature of *Panax notoginseng* is concentrated in the near-infrared and MIR spectra, not terahertz.

5	89 samples from five origins in Yunnan province, Honghe, Kunming, Puer, Qujing, and Wenshan.	HLPC.	Major Science and Technology Project of Yunnan and Kunming	i. The dataset was few samples. ii. Data acquisition equipment is expensive, and it is a non-mainstream data acquisition method.
6	182 samples from four origins in Yunnan province. Honghe--62 Kunming--32 Qujing--26 Wenshan--62	HPLC. Attenuated total reflectance-Fourier transform infrared spectra were provided with a spectral resolution of 4 cm^{-1} from 10000-4000 cm^{-1}.	Natural Science Foundation of Yunnan Province of China	i. The dataset was few samples, and the distribution of samples was imbalance.
7	32 samples from four origins. Suining, Sichuan--15 Wenshan, Yunnan--8 Zhaotong, Yunnan--3 Honghe, Yunnan--6	HPLC.	Young Science and Technology Innovation Team of Sichuan Province, Natural Sciences Foundation of China Science	i. The dataset was few samples.

To mitigate sample selection bias, it is essential to establish clear sample selection criteria and procedures to ensure sample representativeness and diversity. This may involve random selection of samples from different geographical origins and batches to ensure dataset diversity. Furthermore, documenting and reporting the details of the sample selection process can provide transparency and traceability[15].

In *Panax notoginseng* near-infrared spectroscopy data collection, different research teams or laboratories may employ various models or brands of spectroscopy instruments. These instruments may vary in performance and parameters, including resolution, wavelength range, and noise levels. Such instrument differences can lead to data inconsistency, making it challenging to compare and merge different datasets.

To address this issue, calibration and calibration methods can be employed to ensure consistency in data generated by different instruments. Calibration involves using standard samples for adjustment to align the instrument's output data for compatibility with other instruments. Furthermore, instrument differences should be considered during data processing to eliminate variations introduced by the instruments. These methods aid in enhancing data comparability and credibility.

Inconsistencies in calibration and processing methods for *Panax notoginseng* near-infrared spectroscopy data can lead to issues concerning data quality and accuracy. Different research teams may employ varying calibration methods, preprocessing steps, and data processing procedures, complicating the comparison and merging of datasets. For example, some teams may use different data smoothing techniques, baseline correction methods, or outlier handling strategies, and these differences can affect the analysis outcomes.

To ensure data consistency, researchers should establish clear data processing procedures and calibration standards. This includes ensuring that all data undergo preprocessing and cleaning using the same steps to eliminate noise and variations. Additionally, documenting and reporting the details of data processing methods are crucial to ensure that other researchers can replicate the study.

The quality and consistency of *Panax notoginseng* near-infrared spectroscopy data are paramount for the credibility and reliability of

research. However, potential challenges exist due to various factors during data collection:

Noise: Near-infrared spectroscopy data often contain noise stemming from various factors such as instrument errors, environmental conditions, and sample preparation. Noise can affect data accuracy and credibility, necessitating noise removal and correction during data processing.

Data Inconsistency: Data inconsistency may exist between different datasets, attributed to differences in sample selection, instrument variations, or data processing methods. This inconsistency makes it challenging to compare and merge datasets, diminishing research replicability.

Sample Variability: The chemical composition of *Panax notoginseng* samples may be influenced by various factors like growth conditions, seasonal variations, and collection methods. This sample variability may lead to data instability, necessitating statistical analysis and modelling for resolution.

To overcome these challenges, rigorous data quality control and validation measures should be implemented. This includes regular maintenance and calibration of instruments, random sample selection, and replicate measurements to ensure data quality and stability. Additionally, statistical methods should be employed to assess data quality and reliability, ensuring the credibility of research results.

Panax notoginseng near-infrared spectroscopy data may contain proprietary or sensitive information, necessitating careful consideration of data privacy and security. Researchers and institutions must ensure secure data storage and transmission to prevent unauthorized access and leakage.

To address data privacy and security concerns, clear data management and protection policies should be established. This includes measures such as data encryption, access control, authentication, and audit trails to ensure data confidentiality and integrity. Additionally, clear permissions for data usage and sharing should be defined to prevent misuse or improper handling of data.

In conclusion, there are several potential issues in the collection of *Panax notoginseng* near-infrared spectroscopy data, including sample selection bias, instrument differences, inconsistencies in calibration and processing methods, and challenges related to data quality and consistency. Resolving these issues requires the establishment of clear data management and quality control standards to ensure research credibility and reliability. Furthermore, data privacy and security concerns must be prioritized to safeguard sensitive information.

IV. PRINCIPLES OF DATASET CONSTRUCTION

In constructing a dataset for *Panax notoginseng* near-infrared spectroscopy, standardization of sample selection and collection methods is a fundamental principle to ensure data quality and comparability. Standardizing sample selection and collection methods helps reduce sample selection bias and ensures that the dataset is representative and diverse[16], as shown in Figure 1. Here are some key principles and recommendations:

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 5, Sep.-Oct., 2023 pp: 78-88 www.ijemh.com

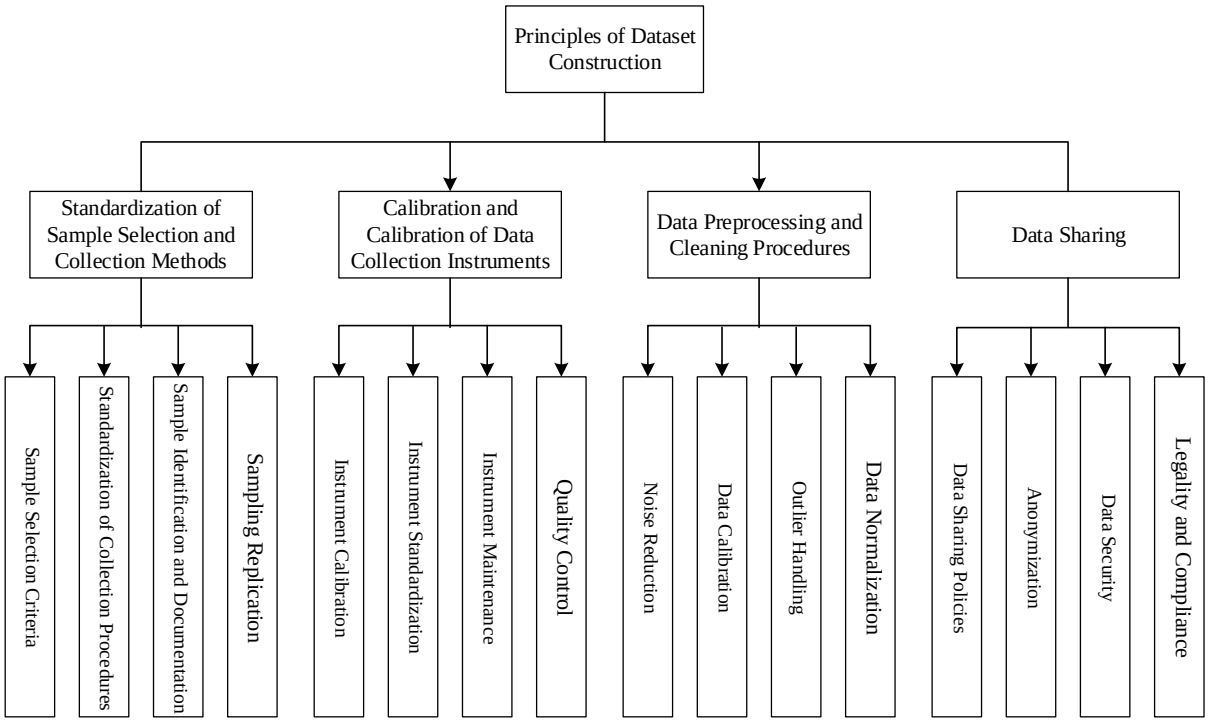

FIGURE 1: SCHEMATIC DIAGRAM OF THE PRINCIPLE OF DATASET CONSTRUCTION

Sample Selection Criteria: Clear criteria for sample selection should be established, including factors such as geographical origin, plant parts, and growth environments. Sample selection should be based on scientific principles to ensure diversity and representativeness of the dataset.

Standardization of Collection Procedures: The process of collecting samples should be standardized, including aspects such as the timing of sample collection, collection methods, and storage conditions. This helps reduce variability among different samples.

Sample Identification and Documentation: Each sample should have a unique identifier, and detailed information about the samples, such as collection date, location, and growth environment, should be thoroughly documented. This aids in traceability and data management.

Sampling Replication: Collecting multiple identical or similar samples and conducting replicate measurements helps assess data stability and consistency. This reduces sampling errors and enhances data credibility.

Calibration and calibration of data collection instruments are crucial steps in ensuring data accuracy and consistency. Different instruments may exhibit performance and parameter variations, necessitating measures to eliminate instrument-induced variability[16], as shown in Figure 1. Here are some key principles and recommendations:

Instrument Calibration: Regular calibration of each instrument should be performed to ensure the accuracy of its output. Calibration should involve the use of standard samples to adjust instrument parameters and correct deviations.

Instrument Standardization: Calibration should be tailored to the instrument model and performance characteristics, ensuring compatibility with other instruments.

Instrument Maintenance: Instruments should undergo routine maintenance and upkeep to ensure their proper functioning. Maintenance includes cleaning optical components, calibrating detectors, and ensuring instrument stability.

Quality Control: Quality control checks should be conducted during the data collection process, including monitoring and calibration using standard samples to ensure data accuracy and stability.

Data preprocessing and cleaning procedures are critical for ensuring data quality and credibility. These steps help eliminate noise, correct biases, and extract useful information[17], [18], as shown in Figure 1. Here are some key principles and recommendations:

Noise Reduction: Noise reduction techniques such as smoothing and waveform

correction should be applied to reduce noise in the data. This enhances data clarity and readability.

Data Calibration: Data should be calibrated using known standard samples to adjust biases and ensure accuracy. Calibration methods should be clearly documented and reported.

Outlier Handling: Identifying and handling outliers is part of data cleaning to prevent outliers from interfering with analysis results. Outlier handling methods should be selected based on data characteristics.

Data Normalization: If the dataset includes data from multiple instruments or batches, data normalization should be performed to ensure consistency and comparability among different data sources.

Data sharing is essential for advancing scientific research, but it must be balanced with the protection of data privacy and security[19], [20] , as shown in Figure 1. Here are some key principles and recommendations:

Data Sharing Policies: Clear data sharing policies should be established, including guidelines for data accessibility, sharing permissions, and usage regulations. This promotes data sharing and collaboration.

Anonymization: When sharing data, anonymization measures should be taken to protect the privacy of individuals or institutions. Identifying information can be removed or blurred to safeguard data privacy.

Data Security: Data should be stored in secure environments with appropriate data encryption, access control, and audit trail measures to prevent unauthorized access and leakage.

Legality and Compliance: Data sharing should comply with relevant laws, regulations, and ethical standards to ensure the legality and compliance of the data.

Constructing a dataset for *Panax notoginseng* near-infrared spectroscopy requires adherence to a set of principles, including standardization of sample selection and collection methods, calibration and calibration of data collection instruments, data preprocessing and cleaning procedures, and data sharing and privacy protection strategies. These principles contribute to ensuring data quality, credibility, and accessibility, thereby advancing the field of geographical origin research of *Panax notoginseng* and fostering collaboration.

V. STANDARD AND STRUCTURE OF IDEAL DATASET

To address the access limitations of the private dataset and propose a standard for achieving an ideal dataset for geographical origin research of *Panax Notoginseng*, consider the following steps and recommendations:

Promote Data Sharing Culture: Encourage researchers and institutions to share geographical origin datasets of *Panax Notoginseng* publicly or with a broader research community. Highlight the benefits of data sharing, such as increased visibility, collaboration opportunities, and accelerated research progress.

Data Standardization: Develop and establish data standards and protocols for geographical origin datasets of *Panax Notoginseng*. These standards should cover data collection, formatting, and documentation. Ensure that standardized data includes essential attributes like sample ID, geographical origin, plant part, collection date, altitude, temperature, humidity, pH level, and chemical composition measurements.

Quality Assurance and Control: Implement quality control measures during data collection to ensure accuracy and reliability. Regularly calibrate and maintain data collection instruments (e.g., spectrometers) to minimize measurement errors. Conduct duplicate measurements and replicate sampling to assess data consistency.

Data Preprocessing and Cleaning: Define clear data preprocessing and cleaning procedures to remove outliers, correct errors, and standardize data formats. Document all steps taken during data preprocessing to ensure transparency and reproducibility.

Metadata Documentation: Create detailed metadata records for each dataset, including information on data sources, collection methods, and any transformations applied to the data. Metadata should also provide context for variables and measurements, helping other researchers understand the dataset.

Open Data Repositories: Consider depositing standardized *Panax Notoginseng* geographical origin datasets in open data repositories or platforms dedicated to plant science or spectroscopy data. This ensures long-term accessibility. Follow best practices for data documentation when submitting datasets to repositories.

Data Governance and Privacy Protection: Establish data governance policies to ensure data security, privacy protection, and compliance with relevant laws and regulations. Anonymize or de-identify data to protect sensitive information about the origin of the samples.

Collaborative Research and Multi-Center Studies: Promote collaborative research projects that involve multiple research teams and institutions. This approach can pool resources, data, and expertise to create comprehensive datasets.

Encourage multi-center studies that collect data from various geographical locations to enhance dataset diversity.

By following these steps and recommendations, it can work towards overcoming the access limitations of private datasets and create an ideal dataset for geographical origin research of *Panax Notoginseng*. This standardized dataset will enhance the quality, comparability, and accessibility of data in the field, ultimately advancing scientific understanding and collaboration in *Panax Notoginseng* research.

The structure of a typical dataset for *Panax Notoginseng* geographical origins includes essential information about each sample, such as sample ID, geographical origin, plant part, collection date, altitude, temperature, humidity, soil pH level, and the concentrations of two chemical components. These columns help researchers understand the source, growth environment, and chemical characteristics of *Panax Notoginseng* samples. It's important to note that the actual structure of the dataset may vary depending on the specific research objectives and collected data. This dataset structure facilitates the capture of crucial information and supports research and analysis of *Panax Notoginseng* geographical origins.

Here's an example of the first 10 columns of a dataset:

Sample ID: A unique identifier for each sample.

Geographical Origin: The geographic location or region from which the *Panax Notoginseng* sample was collected.

Plant Part: The specific part of the *Panax Notoginseng* plant used for analysis (e.g., root, leaf, stem).

Collection Date: The date when the sample was collected.

Altitude: The altitude at which the plant was grown or collected.

Temperature: The average temperature at the collection site.

Humidity: The relative humidity at the collection site.

pH Level: The pH level of the soil or growing medium.

Chemical Composition: The concentration of a specific chemical compound or component, such as ginsenoside Rg1, Rb1.

Infrared Spectroscopy: It provides information of *Panax Notoginseng* samples about chemical composition, quality, geographical origin, and plant variety.

VI. CONCLUSION

This paper delves into the issues surrounding private datasets in the study of *Panax notoginseng* geographical origins. Through an analysis of existing research, it becomes evident that many researchers tend to collect data independently, resulting in private datasets they are unwilling to share publicly. This phenomenon gives rise to several problems, including limited data comparability, challenges in ensuring data quality, and hindrances to collaboration and knowledge exchange. The prevalence of private datasets restricts the depth and breadth of research in the field of *Panax notoginseng* geographical origins, impeding further progress.

To address the issues associated with private datasets, emphasis is placed on the principles and best practices of dataset construction. First and foremost, standardizing sample selection and data collection methods is essential to ensure dataset diversity and representativeness. Regular calibration and calibration of data collection instruments are necessary to eliminate the influence of instrument variations. Data preprocessing and cleaning procedures should be well-defined to enhance data quality and consistency. Additionally, strategies for data sharing and privacy protection are crucial to ensure data availability and security.

Given the challenges posed by private datasets, the following recommendations are put forth to drive the future development of research in *Panax notoginseng* geographical origins:

Promote Data Sharing: Researchers and institutions should actively promote a culture of data sharing by making data public or sharing it with other researchers. This will enhance data availability and comparability.

Establish Data Standards: For *Panax notoginseng* near-infrared spectroscopy data, data standards and specifications can be established to ensure data consistency and comparability.

Multi-Center Collaboration: Multi-center research projects can pool the resources and data of multiple research teams to address data scarcity issues and facilitate research progress.

Utilize Public Datasets: In research, public datasets can be utilized for model evaluation and validation, thereby improving research credibility and replicability.

In conclusion, the issues associated with private datasets hold significant implications in the field of *Panax notoginseng* geographical origins. By adhering to the principles and best practices of dataset construction, promoting data sharing, and taking measures such as collaboration and multi-center research, these issues can be better addressed, driving further advancements in the study of *Panax*

notoginseng geographical origins. In the future, data sharing will continue to play a pivotal role in advancing scientific research and achieving groundbreaking outcomes.

ACKNOWLEDGEMENTS

This work was supported by Science and Technology Project of Chongqing Municipal Education Commission (Grant No. KJZD-K202303301, KJQN202203309 and KJQN202203308), Planning Project of Chongqing Municipal Education Science (Grant No. K22YG313315) and School-Level General Project of Chongqing City Management College (Grant No. 2022SZZX11).

REFERENCES

[1] G. Ming-Liang *et al.*, 'A gradient-based discriminant analysis method for process quality control of carbonized TCM via Fourier transform near infrared spectroscopy: A case study on carbonized Typhae Pollen', *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 265, p. 120363, Jan. 2022, doi: 10.1016/j.saa.2021.120363.

[2] Z. Cui, C. Liu, D. Li, Y. Wang, and F. Xu, 'Anticoagulant activity analysis and origin identification of *Panax notoginseng* using HPLC and ATR- FTIR spectroscopy', *Phytochemical Analysis*, vol. 33, no. 6, pp. 971–981, Aug. 2022, doi: 10.1002/pca.3152.

[3] S. A. Fitia, R. Khathir, and Z. Zulfahrizal, 'Aplikasi Teknologi Near Infrared Reflectance Spectroscopy Dengan Metode Partial Least Square Untuk Prediksi Kadar Patchouli Alkohol Minyak Nilam', *JIMFP*, vol. 6, no. 4, pp. 627–636, Nov. 2021, doi: 10.17969/jimfp.v6i4.18127.

[4] Y. Lu *et al.*, 'Chemometric discrimination of the geographical origin of licorice in China by untargeted metabolomics', *Food Chemistry*, vol. 380, p. 132235, Jun. 2022, doi: 10.1016/j.foodchem.2022.132235.

[5] Y. Yang *et al.*, 'Comprehensive evaluation of Dendrobium officinale from different geographical origins using near-infrared spectroscopy and chemometrics', *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 277, p. 121249, Sep. 2022, doi: 10.1016/j.saa.2022.121249.

[6] J. Xie *et al.*, 'Determination of Cultivation Regions and Quality Parameters of Poria cocos by Near-Infrared Spectroscopy and Chemometrics', *Foods*, vol. 11, no. 6, p. 892, Mar. 2022, doi: 10.3390/foods11060892.

[7] Y. Yang *et al.*, 'Determination of geographical origin and icariin content of Herba Epimedii using near infrared spectroscopy and chemometrics', *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 191, pp. 233–240, Feb. 2018, doi: 10.1016/j.saa.2017.10.019.

[8] C. Ji *et al.*, 'Determination of the Authenticity and Origin of *Panax Notoginseng* : A Review', *Journal of AOAC INTERNATIONAL*, vol. 105, no. 6, pp. 1708–1718, Oct. 2022, doi: 10.1093/jaoacint/qsac081.

[9] L.-X. Tian, J.-H. Li, L. Zhang, B. Ahmad, and L.-F. Huang, 'Discrimination of five species of Panax genus and their geographical origin using electronic tongue combined with chemometrics', *World J Tradit Chin Med*, vol. 7, no. 1, p. 104, 2021, doi: 10.4103/wjtcm.wjtcm_80_20.

[10] X. Liu *et al.*, 'Geographical region traceability of Poria cocos and correlation between environmental factors and biomarkers based on a metabolomic approach', *Food Chemistry*, vol. 417, p. 135817, Aug. 2023, doi: 10.1016/j.foodchem.2023.135817.

[11] L. Li, Z. Zuo, and Y. Wang, 'Identification of geographical origin and different parts of *Wolfiporia cocos* from Yunnan in China using PLS- DA and ResNet based on FT- NIR', *Phytochemical Analysis*, vol. 33, no. 5, pp. 792–808, Jul. 2022, doi: 10.1002/pca.3130.

[12] J. Bai *et al.*, 'Identification of geographical origins of Panax notoginseng based on HPLC multi-wavelength fusion profiling combined with average linear quantitative fingerprint method', *Sci Rep*, vol. 11, no. 1, p. 5126, Mar. 2021, doi: 10.1038/s41598-021-84589-9.

[13] C. Ji *et al.*, 'Multi-Element Analysis and Origin Discrimination of Panax notoginseng Based on Inductively Coupled Plasma Tandem Mass Spectrometry (ICP-MS/MS)', *Molecules*, vol. 27, no. 9, p. 2982, May 2022, doi: 10.3390/molecules27092982.

[14] C. Liu, F. Xu, Z. Zuo, and Y. Wang, 'Network pharmacology and fingerprint for the integrated analysis of mechanism, identification and prediction in *Panax notoginseng*', *Phytochemical Analysis*, p. pca.3195, Dec. 2022, doi: 10.1002/pca.3195.

[15] Y. Zhou, Z. Zuo, F. Xu, and Y. Wang, 'Origin identification of Panax notoginseng by multi-sensor information fusion strategy of infrared spectra combined with random forest', *Spectrochimica Acta Part A: Molecular and*

Biomolecular Spectroscopy, vol. 226, p. 117619, Feb. 2020, doi: 10.1016/j.saa.2019.117619.

[16] H. Zhang *et al.*, 'Rapid determination of *Panax notoginseng* origin by terahertz spectroscopy combined with the machine learning method', *Spectroscopy Letters*, vol. 55, no. 9, pp. 566–578, Oct. 2022, doi: 10.1080/00387010.2022.2125017.

[17] Z. Zhang *et al.*, 'Rapid Geographical Origin Identification and Quality Assessment of Angelicae Sinensis Radix by FT-NIR Spectroscopy', *Journal of Analytical Methods in Chemistry*, vol. 2021, pp. 1–12, Jan. 2021, doi: 10.1155/2021/8875876.

[18] Y. Xu, W. Yang, X. Wu, Y. Wang, and J. Zhang, 'ResNet Model Automatically Extracts and Identifies FT-NIR Features for Geographical Traceability of Polygonatum kingianum', *Foods*, vol. 11, no. 22, p. 3568, Nov. 2022, doi: 10.3390/foods11223568.

[19] C. Liu, Z. Zuo, F. Xu, and Y. Wang, 'Study of the suitable climate factors and geographical origins traceability of Panax notoginseng based on correlation analysis and spectral images combined with machine learning', *Front. Plant Sci.*, vol. 13, p. 1009727, Feb. 2023, doi: 10.3389/fpls.2022.1009727.

[20] S. Zhang *et al.*, 'Untargeted Metabolomics Analysis Revealed the Difference of Component and Geographical Indication Markers of Panax notoginseng in Different Production Areas', *Foods*, vol. 12, no. 12, p. 2377, Jun. 2023, doi: 10.3390/foods12122377.