Document Title:	**Quantitative Measures for Footwear Impression Comparisons**
Author(s):	**Martin Herman, Ph.D., Steven Lund, Ph.D., Hari Iyer, Ph.D., Gunay Dogan, Ph.D., Adam Pintar, Ph.D., Yooyoung Lee, Ph.D.**
Document Number:	**303985**
Date Received:	**December 2021**
Award Number:	**DJO-NIJ-17-RO-0202**

FINAL REPORT

Interagency Agreement between National Institute of Justice, Office of Justice Programs and National Institute of Standards and Technology

Interagency Agreement Number: DJO-NIJ-17-RO-0202

Project Title: Quantitative Measures for Footwear Impression Comparisons

Principal Investigator: Dr. Martin Herman
National Institute of Standards and Technology
100 Bureau Drive, STOP 2000
Gaithersburg, MD 20899-2000
Phone: 301-975-2788
Email: martin.herman@nist.gov

Other Key NIST Personnel: Dr. Steven Lund, Dr. Hari Iyer, Dr. Günay Doğan, Dr. Adam Pintar, Dr. Yooyoung Lee

Award Recipient: National Institute of Standards and Technology

Agreement Period: 09/28/2017 – 06/30/2020 (includes 6-month extension)

Disclaimer

Certain commercial entities, equipment, or materials may be identified in this paper in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

The opinions, recommendations, findings, and conclusions presented in this paper do not necessarily reflect the views or policies of NIST or the United States Government.

Purpose of Research

In the U.S. judicial system, forensic footwear examiners provide expert opinions and interpretations that arise from personal knowledge, training, and experience. Such opinions and interpretations rely heavily on the examiner's subjective judgement rather than on an empirically demonstrable basis. Recognizing this, the U.S. National Academy of Sciences [1] and the President's Council of Advisors on Science and Technology [2] set forth quantitative and automation goals for forensic pattern disciplines including footwear impression evaluations. Despite recent progress in automated footwear impression comparison for database retrieval, there is no end-to-end system available today that can be deployed in casework to provide quantitative support for examiner conclusions.

Given a questioned impression (crime scene) and a test impression (obtained from a shoe of interest), an important question to triers of fact is "did the shoe that made the test impression also make the questioned impression?" In actual casework, the answer to this question is uncertain and different examiners will have different levels of confidence regarding whether or not the two impressions were made by the same shoe. This personal uncertainty highlights the importance of identifying objective information available to shape one's perception. When using a computer algorithm to assign an ordinal similarity score between two impressions, such objective information can be recognized by considering questions such as:

- How similar are the test and questioned impressions?
- How similar to the questioned impression are test impressions from other shoes?
- What similarity levels have we seen in the past when comparing test and questioned impressions known to have come from the same shoe?
- What similarity levels have we seen when comparing test and questioned impressions from two different shoes of arbitrary make, model and size?
- What similarity levels have we seen in the past when comparing test and questioned impressions from two different shoes of the *same make, model and size* (which we call 'close non-matches')?

Though the answers to these questions, by themselves, do not tell us whether or not the questioned impression was made by the shoe of interest, their answers form an empirical basis for subsequent subjective interpretation and can help investigators, lawyers, judges, and jurors alike with the decisions they make throughout the investigative and judicial processes.

Though research has produced many similarity metrics and scores for quantifying pattern comparisons, pattern evidence disciplines have long experienced difficulty in applying their methods to casework in a manner that withstands scientific scrutiny. For

instance, score-based likelihood ratios (SLR), the most common approach to weight of evidence assessments based on algorithmic similarity scores, have been, rightfully, the subject of criticism [3].

The goal of our effort has been to research and develop quantitative methods for two-dimensional footwear evidence assessment to assist examiners by providing empirical support for their findings. The effort's purpose has been to provide a path forward for footwear impression evidence evaluation that focuses additional attention on the body of available empirical information. This effort will change the question from "What is the weight of evidence?" to "What relevant information is available to help assess the weight of evidence?" We view this transition as a significant advance for both the footwear impression discipline, and the entire field of pattern evidence.

Project Design and Methods

Our approach has been to develop an end-to-end system called Footwear Impression Comparison System (FICS). The system is based on a workflow that reduces the potential for bias by eliminating side-by-side human evaluation of the questioned impression and test impressions, parallels the analytic sequence examiners already follow, and provides quantitative support for examiner findings in harmony with the Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD) conclusion scale [4]. The first version of this end-to-end workflow was made into a proof-of-concept graphical user interface (GUI) application, which is described in [5, 6, 7]. An additional paper [8] details subsequent work to replace each component of the initial version focusing on performance and casework utility.

Throughout the remainder of this document we abbreviate "questioned impression" as **Q** and "test impression" as **K**.

Preliminary testing of the existing workflow components was performed using several sets of data, including impressions obtained with the Everspry EverOS scanner [9], from the West Virginia University (WVU) Footwear Impression Data [10], from the CSAFE longitudinal footwear database [11], eight realistic mock crime scene impressions (**Q**s) collected in our lab, and other impressions collected in our lab. The eight mock crime scene impressions are intended to represent a variety of case-like conditions and appear in the top panel of Figure 1. Each **Q** was made from one of 6 different size-9.5 Nike Dual Fusion St 2 shoes (3 pairs, which had been worn by the same person) and compared to **K**s from each of the 6 shoes. When being compared to a **Q** from the opposite foot, a **K** was flipped to appear as belonging to the same foot as **Q**. All **K**s, which are originally loaded as grayscale images, are thresholded and binarized after alignment, to obtain binary contact/noncontact maps. The lower panel of Figure 1 shows an example **K** from each of the six shoes after removing a Nike symbol in the heel of each impression that would have shown when **K** had been flipped, along with an illustration of the region of interest (ROI) for an impression,

which is automatically produced using an alpha hull region completion [12] applied to a thresholded version of **K** or to a binary contact markup of **Q**.

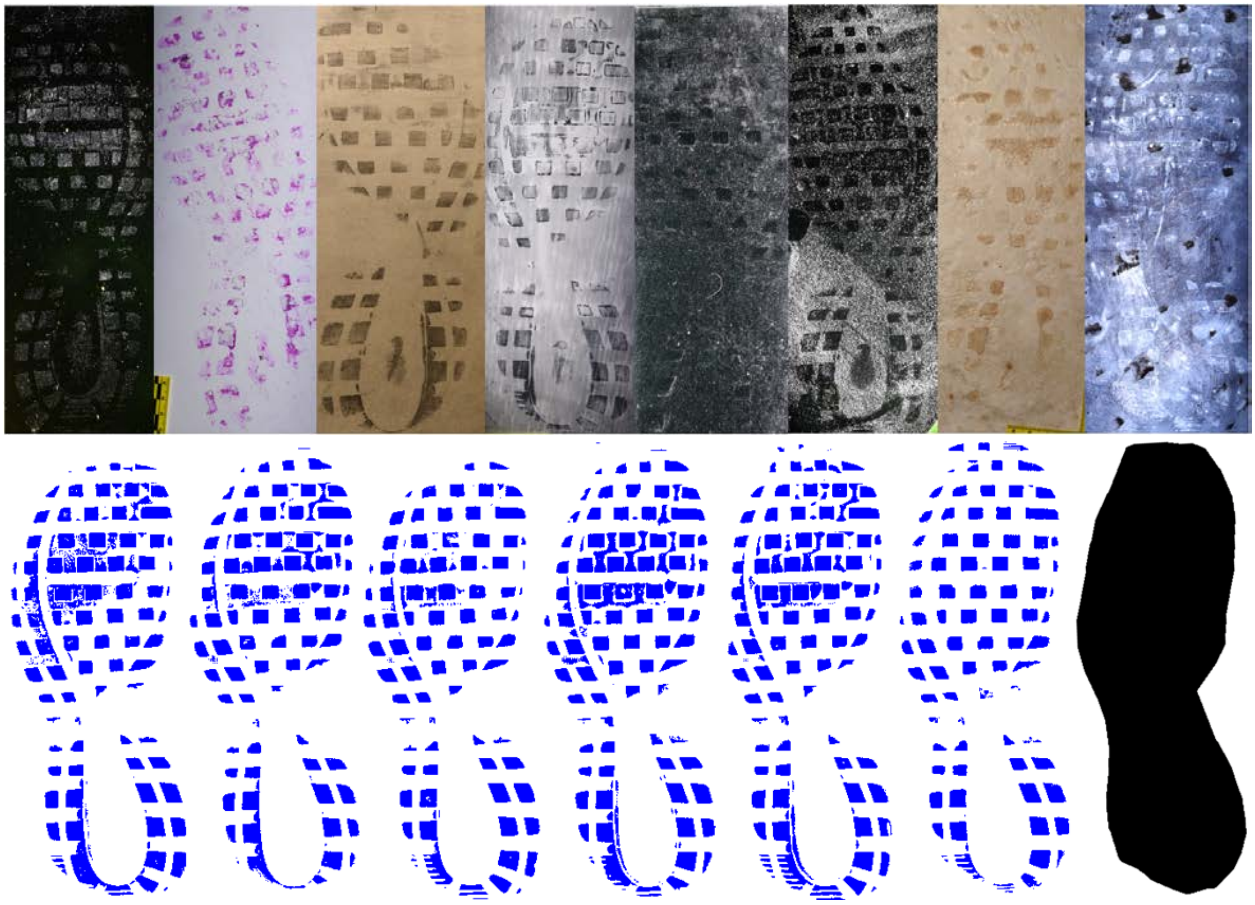

Figure 1: *Top:* **Q**s used in preliminary workflow tests. From left to right: Electrostatic lift from paper, gel lift of synthetic blood on tile treated with acid fuchsin protein, wet impression on desktop treated with magnetic powder, gel lift of wet impression on textured wood treated with aluminum powder, dust impression on glass, gel lift of wet impression on dusty tile, synthetic blood on tile, gel lift of wet impression on tile treated with aluminum powder. *Bottom:* (Blue) Example **K**s from each of 6 different shoes used in preliminary workflow tests; (Black) Example "region of interest" (ROI)

Results

Annotation: We rely on human pattern interpretation to independently annotate **Q** and **K**. Specifically, one user (normally the examiner) examines **Q** and annotates (1) regions of apparent contact, (2) clarity, or how confident the user is that the contact and non-contact annotations would match the features in a **K** made from the shoe that left **Q**, and (3) apparent Randomly Acquired Characteristics (RACs), if any are seen. (These are features

on the footwear outsole due to wear such as cuts, scratches, tears, holes, stone holds, etc.) While RAC and contact markups are binary, clarity is currently expressed using a four tiered color scale. From low to high, the colors and their intended meanings are: red - low confidence that a mated **K** would correspond to the markup in this region; orange - confident a mated **K** would mostly correspond, except near feature edges; yellow - confident a mated **K** would correspond, even near feature edges; green - like yellow, but also any prominent RACs are expected to have left a detectable signature in **Q**. Figure 2 shows example contact and clarity markups. Note that blue is not used to indicate a clarity level, but to fully exclude regions that could indicate an impression has been flipped during algorithm testing.

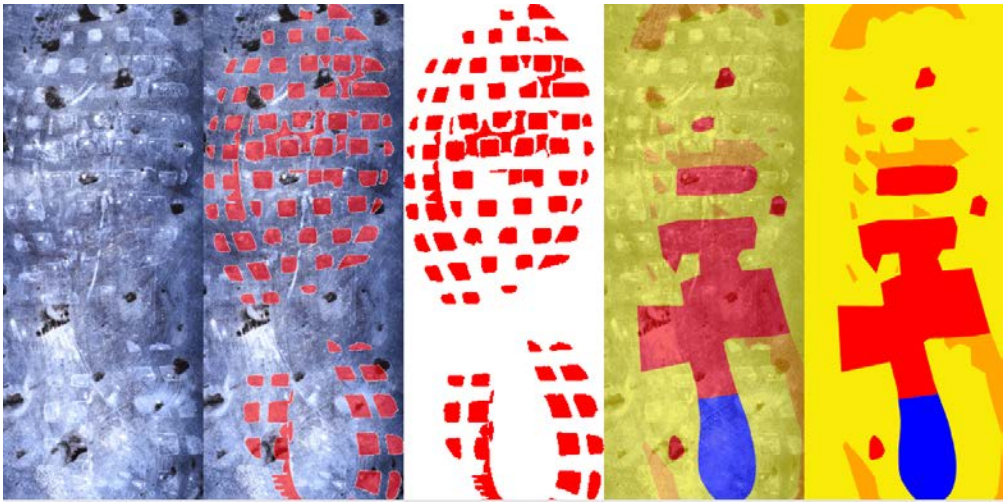

Figure 2: Example markup of **Q**. From left to right: **Q** (a gel lift of a wet impression on tile treated with aluminum powder); overlay of **Q** and contact markup; contact markup; overlay of **Q** and clarity markup; clarity markup.

Ideally, a second user will annotate any **K**s for (1) apparent RAC regions and (2) apparent wear regions, with the aid of looking at the corresponding physical outsole.

Rigid and flexible alignment: We developed rigid and flexible alignment routines to facilitate similarity scores based on pixel to pixel comparison. Rigid alignment is used to find an initial overlay of **K** on the binary contact markup of **Q** and produces an ordered list of potential ways **Q** could fit into **K**. In preliminary testing using cropped regions of the eight **Q**s, the rigid alignment methods effectively placed partials representing at least 20% of the outsole area [8]. After rigid alignment, we apply a flexible alignment routine to accommodate small amounts of distortion that generally occur between any two impressions from the same shoe. Each phase of alignment produces a score that can be used to construct preliminary assessments of the similarity between **Q** and **K** in the sense that two dissimilar outsole designs will have low alignment scores, regardless of how one is positioned relative to the other. We combine the alignment score from rigid alignment with the size score below as an initial assessment of the design and size correspondence between

Q and **K**.

Size score: We evaluate shoe size agreement based on the extent to which flexibly aligning to **Q** causes **K** to expand or contract. Negative size scores occur when **K** contracts and positive scores occur when **K** expands. To test FICS' potential in evaluating size differences, we marked up two questioned impressions from the CSAFE footwear database [11] and compared each questioned to **K**s from 20 size-10 shoes and 20 size-10.5 shoes with the same design. Combining the size and alignment scores showed good ability to discriminate between half-size differences, producing a receiver operating characteristic (ROC) curve with area under the curve (AUC) greater than 94% [8].

Pattern score: Having a pattern score that captures differences in wear is critical for moving towards casework applications, where RAC information is often not available in questioned impressions. We built models to predict the chance that the contact label of a given pixel in **Q** would differ from the label of the corresponding pixel in **K** based on the clarity of the pixel in **Q**. The models also consider how close the pixel is to an edge in **K**, whether it is a contact pixel in **K**, and, if so, whether that contact is labeled as wear. For each pixel, we compute the ratio of predictions from models trained on mated reference pairs and close-non-match reference pairs, respectively. Our pattern score is the sum of the logarithm of these ratios across pixels in the overlapping ROIs of **Q** and **K**. We used this pattern score to compare each of the eight staged **Q**s (Figure 2) to five **K**s from each shoe (30 total **K**s). In every case, the five mated **K**s produced the top five pattern scores [8], indicating strong discrimination among the six close-non-match shoes across a variety of case-like conditions.

RAC scores: When RACs have been marked in **Q** or in **K**, two separate metrics are computed for each marked RAC region. The first is the Jaccard index or Intersection over Union (IoU) [13], which works well when the RAC signature in **Q** is prominent enough to be recognized, or at least suspected, by the user without having seen the outsole [7]. However, it is generally difficult to isolate potential RACs among the undesired noise inherent in most **Q**s. For this reason, we have developed a second metric that only relies on the examiner detecting the RAC region in either **K** or **Q** [7]. This metric is designed to assess the extent to which the RAC region marked in one impression is reflected in the pixel intensities of the other impressions.

Tool for data visualization and score evaluation: For any scoring schemes or metrics, there are always alternatives. For large data sets, it can be tedious and time-consuming to analyze the scoring schemes in detail. To facilitate such analyses, we developed VEMOS (*Visual Explorer for Metrics of Similarity*) [14]. During score development we can use data visualization techniques available in VEMOS to assess whether chosen metrics match intuition or to spot unexpected behavior. Multiple scoring metrics can be combined with VEMOS to create fused scores with improved discrimination performance. VEMOS data visualization tools will have even greater value during case evaluation, allowing users to easily demonstrate and interrogate the information available in the set of reference

comparisons regarding the meaning of scores obtained in casework evaluation. These easy-to-use tools focus additional attention on empirical results, thereby further bolstering the scientific validity of the evidence evaluation process.

Mapping scores to SWGTREAD conclusion scale: Most examiners conduct comparisons using a sequential analysis with stages for design, size, wear, and RACs. We have implemented a parallel sequence for FICS that includes mapping outputs to SWGTREAD conclusions [4]. This provides a way to walk through a data-based decision process for how the considered **Q** and **K** lead an examiner to a given conclusion. Such a presentation rightfully draws attention to the collection of reference comparisons that have been conducted, any of which could be demonstrated if requested, and can help inform the meaning of the pattern evidence. Details for how scores from various FICS components are placed in different contexts of reference scores and related to conclusion levels can be found in [7].

Implications for Criminal Justice Policy and Practice in the U.S.

To address concerns expressed within the broader scientific community (2009 NAS report [1], 2016 PCAST report [2]), examiners require a collection of reproducible results capable of serving as the basis of interpretation. Black box studies seek to provide such a collection for expert human comparison. While clearly valuable, results based exclusively on human experts face limitations of within- and between-expert variability and restrictions on the number of test comparisons an examiner can undertake. If computer algorithms could perform sufficiently well, they could provide impartial, quantitative support for examiner conclusions. Unfortunately, computer algorithms of today have not been demonstrated to have similar ability as human experts to discriminate between mated and close non-match footwear impression pairs.

The long-term goal of the research described here is to develop a system that would eventually be used in casework. This system uses a workflow that parallels the scheme examiners already follow, has the potential to reduce bias, and provides quantitative and empirical support for examiner findings in harmony with the SWGTREAD (or similar) conclusion scale.

Successfully completing the long-term research would be a transformative landmark in pattern evidence evaluation. When asked to explain their conclusions, experts could first present a flow chart where the direction taken at each node is related to a demonstrable body of empirical results involving design, size, wear and RAC scores for the current case comparison, where those scores lie in relation to distributions of reference ground-truth-known comparisons, and what that means in terms of eliminating or not eliminating the shoe of interest on the basis of these design, size, wear and RAC scores. If

pressed, examiners could continue to provide a visual tour through the empirical results obtained in the current case and the body of information provided by the available reference comparison collection, including showing the image sets from the reference comparisons most relevant to the current case. To reduce concerns of bias, the developed system could be applied in such a way that the examiner who interprets and annotates the **Q** has never seen the shoe of investigative interest or its **K**s. Evidence evaluations will have greater repeatability and reproducibility and any disagreement could be traced back to differences in how the original impressions were marked up. Once annotated, a single interpreted image can be used in many automated comparisons. Taken to the extreme, within a reference database every annotated **Q** could be automatically compared to every annotated **K**. Automated comparison systems can complete more comparisons than can be accomplished by experts, who must limit participation due to caseloads, conducting fully manual comparisons. In the future, we envision that standard practice for casework would include comparing **Q**s with an entire blind lineup of **K**s from many shoes, including a shoe of investigative interest. Before that time, achieving the research goals identified in this work will have immediate impact by assisting experts to provide powerful reports and testimony that accurately reflect the information available and withstand scientific scrutiny. This practice will serve as an example to other pattern evidence disciplines for the scientific reporting of comparison algorithm results in casework.

Products Resulting from Project

Research Papers

- Gautham Venkatasubramanian, Vighnesh Hegde, Sarala Padi, Hari Iyer and Martin Herman, "Comparing Footwear Impressions that are Close Non-Matches Using Correlation-Based Approaches," *Journal of Forensic Sciences*, 2021;66(3):890–909. https://doi.org/10.1111/1556-4029.14658

- Gautham Venkatasubramanian, Vighnesh Hegde, Steven P. Lund, Hari Iyer and Martin Herman, "Quantitative Evaluation of Footwear Evidence: Initial Workflow for an End-to-End System," *Journal of Forensic Sciences*, 2021;66(6):2232–2251. https://doi.org/10.1111/1556-4029.14802

- Eve Fleisig and Günay Doğan, "VEMOS: A GUI for Evaluation of Similarity Metrics on Complex Data Sets," NIST Technical Note 2160, June 2021. https://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.2160.pdf

- Vighnesh Hegde and Steven P. Lund, "Expanding the NIST Footwear Evidence Framework: Revised RAC Scores, Weighting Reference Comparisons, and Mapping Results to Examiner Conclusions," 2021, *Presubmission Manuscript*.

- Steven P. Lund, "Footwear Impression Comparison System (FICS): A Workflow Built while Focusing on Casework Performance," 2021, *Presubmission Manuscript*.

Talks

- Steven Lund, "Quantitative Measures for Footwear Impression Comparisons," *2020 NIJ Forensic Science R&D Symposium,* February 18, 2020, Anaheim, CA. Available at: `https://forensiccoe.org/workshop/2020-nij-forensic-science-rd-symposium/`.
- Steven Lund, "NIST Footwear Impression Comparison System," *Forensics@NIST 2020,* November 5-6, 2020, NIST, Gaithersburg, MD. Available at: `https://www.nist.gov/news-events/events/forensicsnist-2020`.

Software

- Eve Fleisig and Günay Doğan, "VEMOS (Visual Explorer for Metrics of Similarity)," 2020. Software available at: `https://github.com/usnistgov/VEMOS`.

References

1. National Research Council. Strengthening forensic science in the United States: a path forward. Committee on Identifying the Needs of the Forensic Sciences Community. Washington, DC: National Academies Press, 2009. Document No.: 228091.

2. President's Council of Advisors on Science and Technology (PCAST). Report to the President – Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. Washington, DC: Executive Office of the President of the United States, 2016 Sep.

3. Neumann C. Defence Against the Modern Arts: the Curse of Statistics–FRStat. arXiv preprint arXiv:190801408, 2019.

4. SWGTREAD. Range of Conclusions Standard for Footwear and Tire Impression Examinations. Scientific Working Group for Shoeprint and Tire Tread Evidence, 2013 Mar. https://www.nist.gov/system/files/documents/2016/10/26/swgtread_10_range_of_conclusions_standard_for_footwear_and_tire_impression_examinations_201303.pdf.

5. Venkatasubramanian G, Hegde V, Lund S, Iyer H, Herman M. Quantitative Evaluation of Footwear Evidence: Initial Workflow for an End-to-End System. J Forensic Sci 2021;66(6):2232–2251. doi: org/10.1111/1556-4029.14802

6. Venkatasubramanian G, Hegde V, Padi S, Iyer H, Herman M. Comparing Footwear Impressions that are Close Non-Matches Using Correlation-Based Approaches. J Forensic Sci 2021;66(3):890-909. doi: 10.1111/1556-4029.14658.

7. Hegde V, Lund S. Expanding the NIST Footwear Evidence Framework: Revised RAC Scores, Weighting Reference Comparisons, and Mapping Results to Examiner Conclusions. Presubmission Manuscript, 2021.

8. Lund S. Footwear Impression Comparison System (FICS): a Workflow Built while Focusing on Casework Performance. Presubmission Manuscript, 2021.

9. EverOS V1.0. Dalian Everspry Sci & Tech Co., Ltd., 2020. http://www.everspry.com/en/products/products_03.htm

10. Richetelli N, Lee MC, Lasky CA, Gump ME, Speir JA. Classification of Footwear Outsole Patterns using Fourier Transform and Local Interest Points. Forensic Sci Int. 2017; 275:102–109. doi: 10.1016/j.forsciint.2017.02.030.

11. Center for Statistics and Applications in Forensic Evidence (CSAFE). Longitudinal Shoe Outsole Impression Study; 2019. Available from: https://forensicstats.org/shoeoutsoleimpressionstudy/.

12. Edelsbrunner H, Kirkpatrick D, Seidel R. On the Shape of a Set of Points in the Plane. IEEE Transactions on Information Theory. 1983; 29(4):551–559.

13. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019; 658–66.

14. Fleisig E, Doğan, G. VEMOS: A GUI for Evaluation of Similarity Metrics on Complex Data Sets. NIST Technical Note 2160, June 2021. https://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.2160.pdf