Forensic Science International 322 (2021) 110753

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Collecting data on textiles from the internet using web crawling and web scraping tools

Cyril Muehlethaler[a,b,c,*], René Albert[c]

[a] University of Quebec at Trois-Rivières, Canada
[b] Laboratoire de Recherche en Criminalistique, Trois-Rivières, Canada
[c] Centre International de Criminologie Comparee, Montreal, Canada

ARTICLE INFO

Article history:
Received 13 August 2020
Received in revised form 9 March 2021
Accepted 11 March 2021
Available online 15 March 2021

Keywords:
Forensic
Interpretation
Fiber
Fibre
Population study
Market study

ABSTRACT

Fibre population surveys are a necessary part of the forensic fibres examination field. They provide valuable information as to which fibres are the most popular and help estimate the likelihood of observing similar properties in a fibre unrelated to the event. The time needed to carry these types of studies is however a major obstacle to wider use. With the advent of e-commerce and digital computation, collecting information from digital sources and structuring it in a convenient way may provide meaningful information on fibres population. It has become more affordable for researchers who can now devote most of their time to extracting meaningful information from the structured data.

In this article, we have used a scrapy and kibana/elastic search interface to crawl and scrape a major online clothes retailer. In less than 24 h we have extracted 68 text-based field describing a total of 24,701 clothes to help provide precise estimations of fibres types and color frequencies. We were able to provide data that cotton, polyester, viscose and elastane are the 4 main types of fibres used in the textile industry. Elastane, while being very popular in garments, rarely accounts for more than 10% of the mass while cotton accounts for up to 80% of content. The most common colors are white, black, and blue, with important dependencies to the fibre type. Through further statistics and examples we demonstrate that web scraping techniques have the potential to provide near real-time population studies that can greatly benefit forensic practitioners.

1. Introduction

Fibre traces are commonly exploited in forensic science [1]. They are easily transferred whenever an activity provides a contact of sufficient intensity between a donor textile and any receiver surface (i.e. assaults, brawls, homicides, grand theft auto, or hit-and-runs). The number of fibres transferred is directly influenced by the intensity and the duration of contact, in addition to the characteristics of the donor fabric (i.e. garment, hat, scarf, sofa, seat), recipient surface (i.e. smooth, rough, textile or not), and time elapsed between the supposed transfer and the collection (persistence). The total number of fibres recovered and their physico-chemical characteristics are extremely important when assessing the case and determining the significance and likelihood of observing them if the suspected garment is truly at its origin rather than the fibres being present by chance on the recipient surface. Despite their polymorphism, it is not rare that a fibre present at random on a relevant recipient surface (i.e. background) possess similar non-differentiable class-characteristics as the fibres of interest collected at a crime scene. For these reasons, population studies, studies of target fibres, and color blocks studies have been important contributions in the forensic literature. These studies allow forensic scientists to evaluate the frequency of occurrence of particular combinations of type and color, and help estimate the likelihood of observing similar properties in a fibre unrelated to the event. Previously published studies have been intrinsically limited to a geographical area such as Europe [2–6], Australia [7,8], or North America [9,10], a particular type of recipient substrate such as public seats [6,7], outdoor surfaces [4,11], car seats [5,12], or living subjects [3,10,13], or a particular type of fibre such as cotton, polyester, or acrylic [14–16]. For a recent and comprehensive review of all relevant fibre population studies refer to Schnegg et al. [17]. Although extremely useful in particular scenarios, these studies may sometimes suffer from the representativity

* Correspondence to: University of Quebec at Trois-Rivières, Department of Chemistry, Biochemistry and Physics, CIPP-2134, 3351 Boulevard des Forges, Trois-Rivières, QC G9A5H7, Canada.
E-mail address: Cyril.Muehlethaler@uqtr.ca (C. Muehlethaler).

https://doi.org/10.1016/j.forsciint.2021.110753
0379-0738/

of their data because they permit forensic examiners to extrapolate to other regions or other substrates under special circumstances only, and rapid changes in fashion diminish their accuracy.

For the lack of better solution, forensic scientists keep relying on these studies and realize population surveys that are both time and resources consuming. A population study that ranges in the thousands of fibres, already represents weeks of work. At best, a few new population studies accounting for new types of fibres appear each year, while fashion, production, developments and trends are supposedly evolving far quicker. For these reasons, the field of fibres examination constantly lacks a bit of reactivity, questioning the relevance of quantitative interpretation with outdated studies.

With that in mind, we have been looking for a simpler way to comprehend the fibres frequencies and colors. Efforts to retrieve and store composition information have already been attempted manually. In 1988, Carroll et al. reported on a computerized database for assistance in the forensic identification of fibres [18]. They compiled generic category names, tradenames, and birefringence properties of around 1000 different fibres. Biermann and Grieve put together 20,786 records of clothes advertised in independent mail order companies [19–21]. This colossal work was accumulated manually within 9 months at a rate of approximately 150 records/day. Their work has been the first of this type and an important contribution to estimating fibre frequencies. This procedure permits to obtain a snapshot of the population at a specific time and location and allows for evaluation of compositions and colors frequencies quite easily. This type of purchase from catalog is no longer used nowadays, and the 21st century technological equivalent is now to browse clothes and shop on the internet.

The last decade has seen a major trend in collection, processing, and interpretation of big data, including other statistical treatments such as machine learning and artificial intelligence. Gathering information from digital sources and the ability to conveniently structure it has become more affordable for researchers. The tools are often freely distributed on the internet, tutorials explain how to code and extract the relevant information, and computing times have considerably diminished over the years. Collecting the data is made easy, convenient, and affordable to almost anyone, and practitioners can instead focus on extracting the meaningful information from the structured data.

Among these computing techniques, a whole category relies on robots that are coded to browse (bot crawler) and collect data (bot scraper) on the internet. These tools are extremely useful for obtaining relevant data in a systematic and automated way, which can help creating structured databases [22–24]. The ethics and legality of these procedures has been questioned recently [25]. Thus far, these reservations are mostly targeted towards individuals that profit from the collected data, gain any financial interest, or use personal information on their own behalf. Guidelines on the use of web scraping for research have been published in the recommendations from the Ethical Decision-Making and Internet Research committee [26]. In short, the courts have ruled against the use of web scraper for obtaining financial information (e.g. price comparator/aggregator) or personal information (e.g. social media/dating sites). For other applications, web scraping is a legal effective tool. Large companies are even aware of web scrapers and explicitly state their allowed (respectively disallowed) fields in the web page that could be extracted by a robot in a *robots.txt* file, which can generally be accessed through the internet browser at *<website name>/robots.txt*.

In this paper, we have used a web crawler and web scraper for browsing a major internet clothes retailer website. In less than 24 h, we have collected information on 24,000 clothes, including composition, color, and size, together with more than 60 other fields. Statistics about clothes types, colors, and compositions are presented. Differentiations of the statistics are made based on the number of different fibres types used in the composition (i.e. pure vs blended). Finally, other particularly relevant extracted fields are discussed, together with perspectives of such studies for the forensic community, in terms of comparing different countries or different periods of the year.

2. Materials and methods

2.1. Data collection

The web crawler was developed in-house based on open source tools Anaconda Navigator 1.9.7 (Python data science platform), Scrapy 1.5.2 (web spider for crawling websites) and Elasticsearch 6.7.0 / Kibana 6.7.0 (monitoring of collected data). The first part of developing any web scraper is to analyse the target website pages for their structure and isolate elements we wish to collect. We used web developer tools integrated in web browsers, which enabled us to identify the XPath expressions required to extract the desired fields. We designed the item structure and all its relevant components to match the details page of the target website (i.e. material composition, color, type, …). In order to obtain a sample that is an accurate representation of textile distribution in a given time frame, it is important to have a systematic approach to record acquisition. Starting from the most recent records, the bot will then move to older pages in a systematic way. Scrapy was used to write the data collection application. It offers a sophisticated package that has a variety of pipeline options to adapt to various data collection strategies. For data storage and acquisition monitoring, we used Elasticsearch and Kibana. Kibana offers real-time monitoring of the collected data with a rich visual interface and live statistics, adaptable to any type of extracted fields (Fig. 1).

We finally wrote a Python script to enable us to extract the data in a CSV format for further analyses with other statistical packages.

Batch scripts were created to easily initiate the data collection and extraction. These scripts launch the required servers and applications needed for the data acquisition, with a level of abstraction that allows users to access the tools without needing any technical knowhow. Some maintenance of the spider bot scrape rules are required from time to time, to match the occasional changes in the page structure. Fig. 2.

2.2. Data classification

A number of modifications were needed in order to prepare the data for statistics computation. For example, both nylon and polyamide fibre types were present in the initial text-fields collected. All nylon entries were converted to polyamide for harmonization of the categories. The same holds true for other fields, which were manually grouped under the same name due to multi-lingual translations (e.g. cotton and coton) or various commercial names or synonyms (e.g. viscose and rayon). The most complicated and laborious classification field was the codification of the color. Many entries had a unique color brand name, often far from suggesting the color it really represented, like *Antique* (white), *Abyss* (light blue), or *Alabaster* (beige). In case of doubts, the URL was systematically consulted and the color decided from the available pictures of the garment. Other names were more easily classified as the color was completed by a descriptive adjective, for example: *Allure blue*, *Alpine green*, or *Amaranth purple*. The classification was completed by searching and replacing all occurrences of the specific color names by their more generic equivalent. An automated function would be possible to implement by training an algorithm with keywords but was not deemed necessary at this stage of the project.

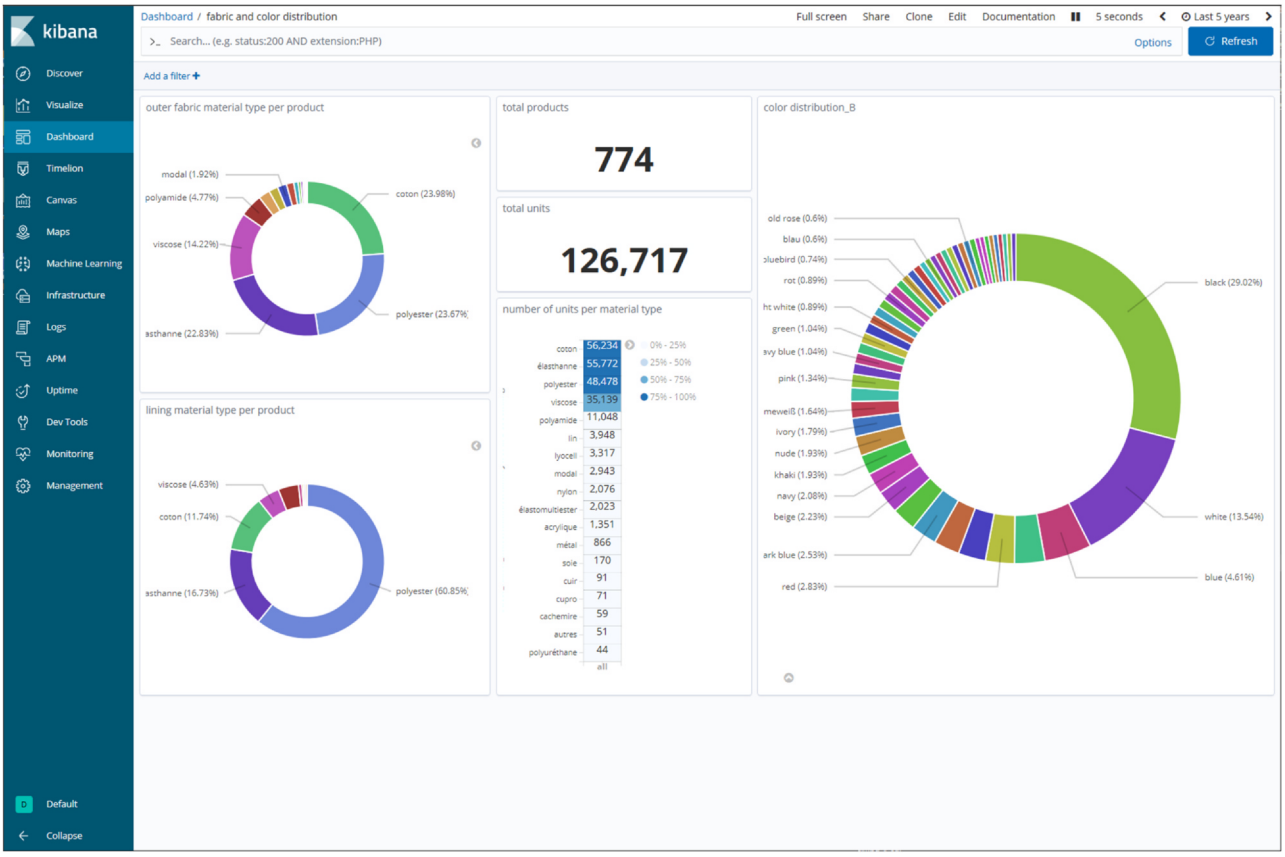

Fig. 1. Example of the Kibana 6.7.0 live interface, all statistics and graphs are updated every 5 s. From upper left to bottom right, the panels are respectively: 1) outer fabric materials type per product, 2) lining material type per product, 3) total products, 4) total units, 5) number of units per material type, 6) and color distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

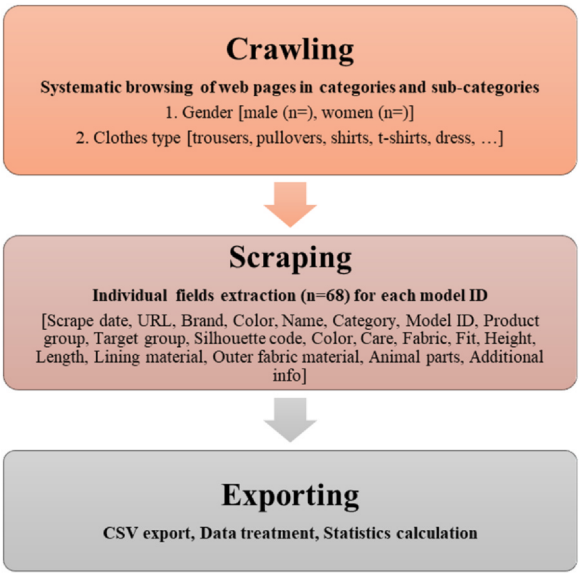

Fig. 2. Diagram of the crawling and scraping procedure based on Scrapy and ElasticSearch/Kibana.

2.3. Data treatment

The database of records was exported as a unique CSV file of 24,702 × 68 text-based fields and managed with Microsoft Excel version 16.0 (2016). Statistics and graphics were obtained from Excel, RStudio using the GGPlot2 package, and OriginPro 2019 9.6.0 (OriginLab Corp.).

3. Results and discussion

Due to the number of data collected, this article can only present a selection of relevant statistics and graphs. For readers interested in other frequencies and numbers, the data set is available upon request.

3.1. Composition

The statistics presented in the next sections correspond to a total amount of 24,701 clothes, 12,168 of which are categorized under women garments, and 12,533 under men garments. Among these clothes, 50% are of pure composition (i.e. one type of fibre only), and 39% are composed of 2 types of fibres (e.g. coton/polyester) (Fig. 3). Complex clothes with at least 3 different types of fibres are only a minority (11%).

Fig. 4 presents the most common types of clothes, divided by gender categories. The proportion of t-shirts, trousers, and pullovers is much more important in men, while dress and skirts are almost exclusively listed for women. Other categories such as bikinis, bras, nightdress, or beach accessories were kept for comprehensive estimation of the statistics. Although unlikely, it cannot be stated beforehand that these products will not show up in very specific forensic investigations someday. The categories listed are the most common with at least 0.05% of occurrence, other groups that were in minority were listed under the *other* category.

Important calculation distinctions need to be clarified regarding the main fibre types frequencies. Fig. 5 presents the frequencies of appearance of each fibre type in the listed composition of garments from our database. This helps answering questions of the type “how

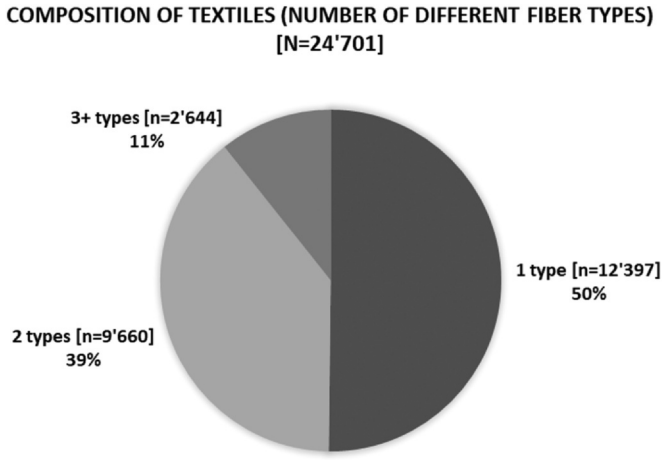

Fig. 3. Number of different fibre types used in the composition of textiles (n = 24,701).

many clothes containing cotton fibres are present in the database"? These values are independent of the fibre proportion in the garment (i.e. 10, 20, …, or 100%) and based solely on the reported composition. Fig. 6 provides a similar estimation of the most common fibre types but this time weighted by their proportions in blended garments. These values help answer questions of the type "how common are cotton fibres among garments?" and provide a more accurate estimate of the available fibre population at a time. Weighted proportions are needed to more precisely estimate the occurrence of fibres that might transfer, or simply be present as background. Taking an example, cotton and elastane are both common appearing at respectively 56.33% and 32.07% in garments. Elastane however rarely accounts for more than 10% of the mass while cotton is often found in amounts up to 80%. Their weighted frequencies of appearance are then 1.61% and 49.24% respectively.

The most common fibre types are cotton (49.24%) and polyester (28.65%), more than three times ahead of other frequent types such as viscose (9.77%) and polyamide (5.13%). Polyester is present in 41.23% of the garments, often as a blend of lower percentage. As a consequence, it represents only 28.65% of the fibres population. The opposite holds true for cotton which represents 56.33%. Apart from these 4 types, all other fibres do not represent a proportion higher than 1.6% in the general population. While the textile industry predicts a huge increase in polyester use in the next decade, the clothes on the market right now do not reflect this change. China produces currently around 70% of the polyester on the market, and is expect to increase in the following decade. These figures would suggest this trend to be occurring in the apparel industry as well (see annexe for a more detailed comparison of the worldwide fibre market). However, cotton is still the most popular fibre type as of 2019 in a typical European market. Overall, natural fibres account for 49.98% in the population, with an overwhelming majority being cotton (49.24%), and only 0.74% for other natural fibres such as linen or jute. Animal fibres account for 2.22% in the total population, wool (1.33%), leather (0.43%) and silk (0.33%) being the most important. The leather category comprises leather used as a fabric in garments production, but also contains denim jeans possessing a small leather patch at the back of the belt area. Other animal fibres include the goat cashmere (0.08%) and mohair (0.03%), and alpaca fibres (0.02%). All these types are very minor with proportions lower than 0.1%. 81 clothes among the 24,701 contained metal fibres in their composition (0.03%), although never present at more than 10% in each item according to their label. These metallic fibres are based on gold, silver, steel or aluminum materials, either pure or as a coating for synthetic fibres, and provide aesthetic and conductive properties (i.e. static electricity free) for technical fabrics. Elastane comes directly after the 4 most common fibre types (coton, polyester, viscose and polyamide) with 1.61% of the fibres population. While never used in any of the garments as pure fibre type (i.e. 100% elastane), and often being a minority component in blended fabrics (less than 10%), it still appears in 32% of clothes (7921 out of 24,701) and often used in 2-types blended fabrics.

Fig. 7 presents the most common 100% pure clothes, by making a further distinction between men and women. This separation becomes particularly interesting for distinguishing the clothes made of viscose, which are very common for women (20.44%), and much less for men (2.4%). The opposite trend is visible for cotton clothes, which represent 62.63% of men clothes, and 41.96% of women. A plausible explanation is the very high proportion of 100% cotton t-shirts and shirts among the men's category, while the 100% viscose tops and dresses are very popular women's clothes. Other types of fibres do not show significant differences. Altogether, cotton and polyester clothes represent 81.8% of the 12,397 pure garments considered.

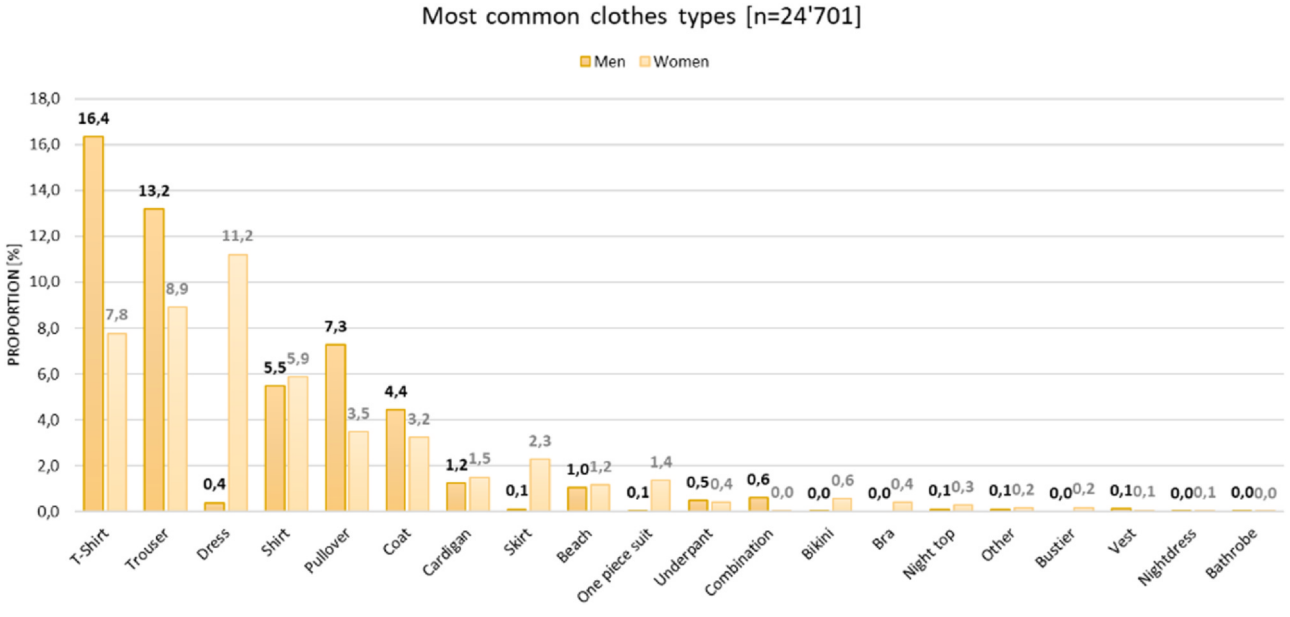

Fig. 4. Statistics on the most common clothes types according to men and women populations (n = 24,701).

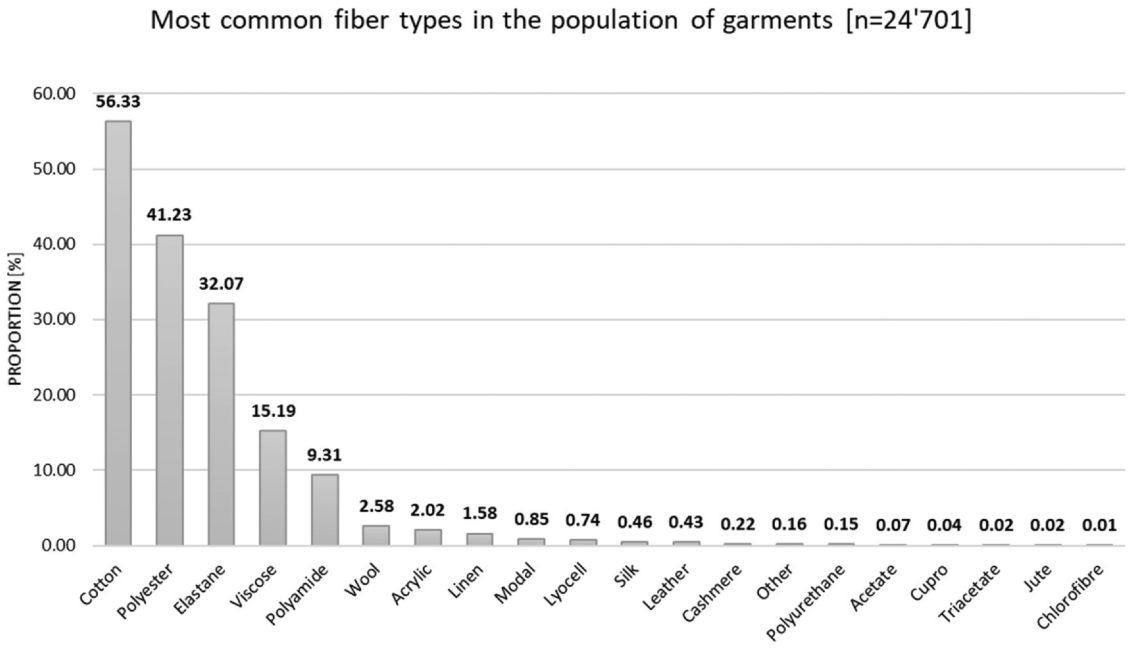

Fig. 5. Statistics on the most common fibre types as listed in the composition of the garments in the database (n = 24,701). Includes pure and blended garments.

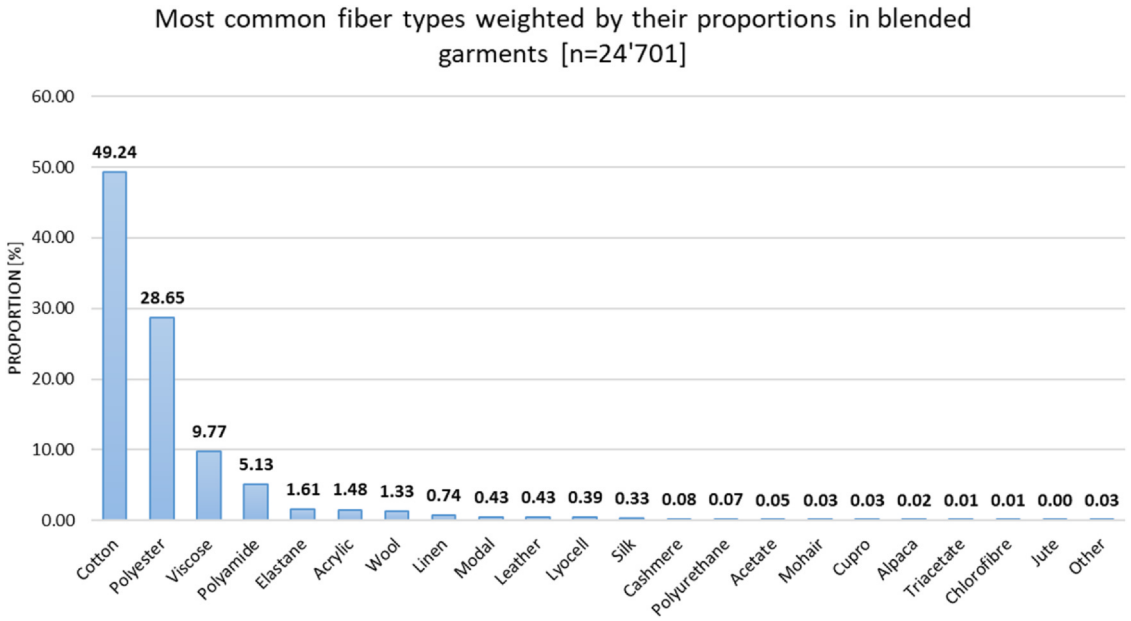

Fig. 6. Statistics on the most common fibre types used in garments, weighted by their individual proportions in blended fabrics. Population of fibres was estimated by summing up all individual compositions (proportions) for each garment in the study (n = 24,701).

As Fig. 3 suggested, 50% of the clothes on the market are composed of blended fabrics of at least two types of different fibres. Among these blends, some combinations are far more frequent than others as demonstrated with Fig. 8. For computing reasons, the categories are not mutually exclusive, which means that a few items containing cotton and elastane might also appear under the cotton and polyester category if their composition is cotton/polyester/elastane. Unsurprisingly, the three most common blends of fibres are cotton/elastane (23.3%), polyester/elastane (21%), and cotton/polyester (19%). The light green area at the top of the first three columns indicate the 5.81% of clothes that have a composition including all of these three most common fibre types in blended garments. Except for viscose and polyamide, all other types of fibres are a minority in blended garments. For ease of visualization, Fig. 9 and Fig. 10 were produced in order to help understanding the proportion of 2-types versus 3+ types of fibres blending.

Fig. 9 indicates the separation of all the clothes composed of cotton and polyester fibres based on their proportion (n = 3′330). Each dot represents a particular composition that was observed during the study, their size being correlated to the number of times they appeared. Values on the diagonal indicate garments made of these two types of fibres only (i.e. cotton and polyester sum is 100%), while values below the diagonal indicate that additional fibres are present in the composition (i.e. cotton and polyester sum is less than 100%). It is possible to notice two important things in Fig. 9: 1) the cotton proportion is seemingly higher than the polyester proportion (i.e. denser and larger dots on the right-end of the graph), and 2) cotton and polyester most commonly appear as a 2-components

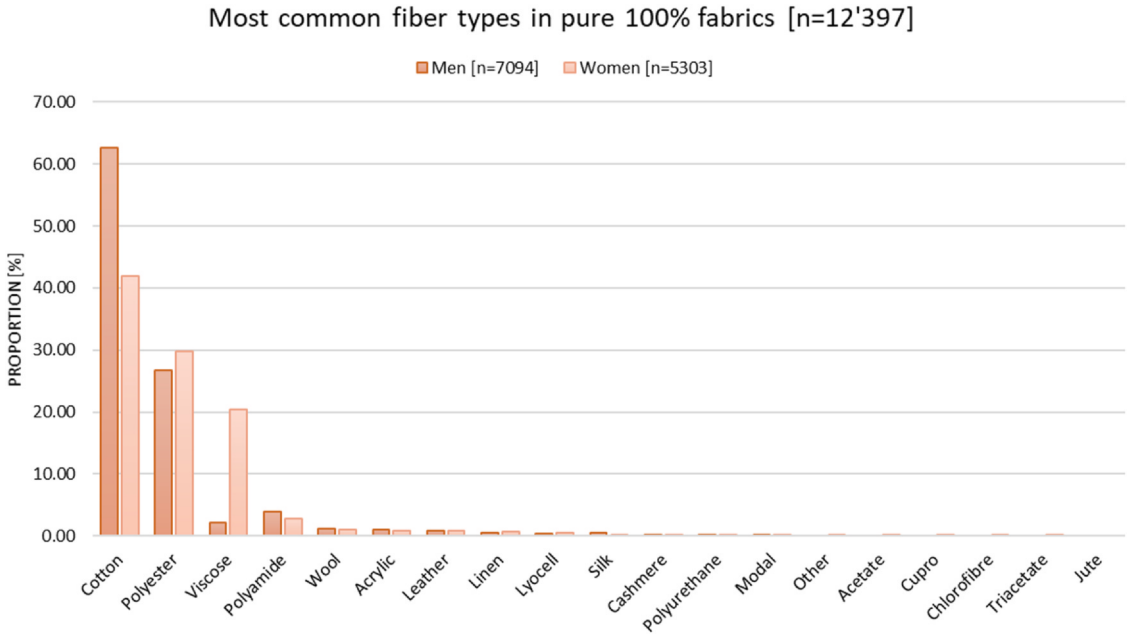

Fig. 7. Statistics on the most common pure textiles in men and women (i.e. 100% composed of a single fibre type) (n = 12,397).

blended garment, rather than having 3 or more different types of fibres in the composition (i.e. denser and larger dots on the diagonal). The main dots on the graph are respectively (for cotton/polyester): 35/65 (4.86%), 50/50 (5.19%), 60/40 (9.24%), 70/30 (5.22%), and 80/20 (11.11%).

Fig. 10 was constructed in the same way as Fig. 9 for the other categories representing the five most frequent blends, respectively: a) cotton-elastane, b) polyester-elastane, c) polyester-viscose, and d) viscose-elastane. Fig. 10 a) reveals that 71.99% of the cotton and elastane blends are based on 90% or more of cotton, with the most popular one being 98% cotton and 2% elastane (26.01%). Elastane is rarely found at more than 10% as illustrated by the downward tightening of the values. The same holds true for Fig. 10 b) and Fig. 10 d), with 16.32% of the polyester-elastane clothes being 95/5 blends, and 12.81% of the viscose and elastane clothes being of 95/5 blends.

3.2. Color

The estimation of the color frequencies was the most challenging due to variety of existing ways in describing the shade (common and generic names, multiple colors). The color categories correspond to the main color as indicated on the label and do not include the fact that minor parts of the garment can be of another color. Garments composed of multiple colors for which it was not specified, or those that were more difficult to assess (no dominant color, or blend of too many colors) were categorized as *other*. A few garments were

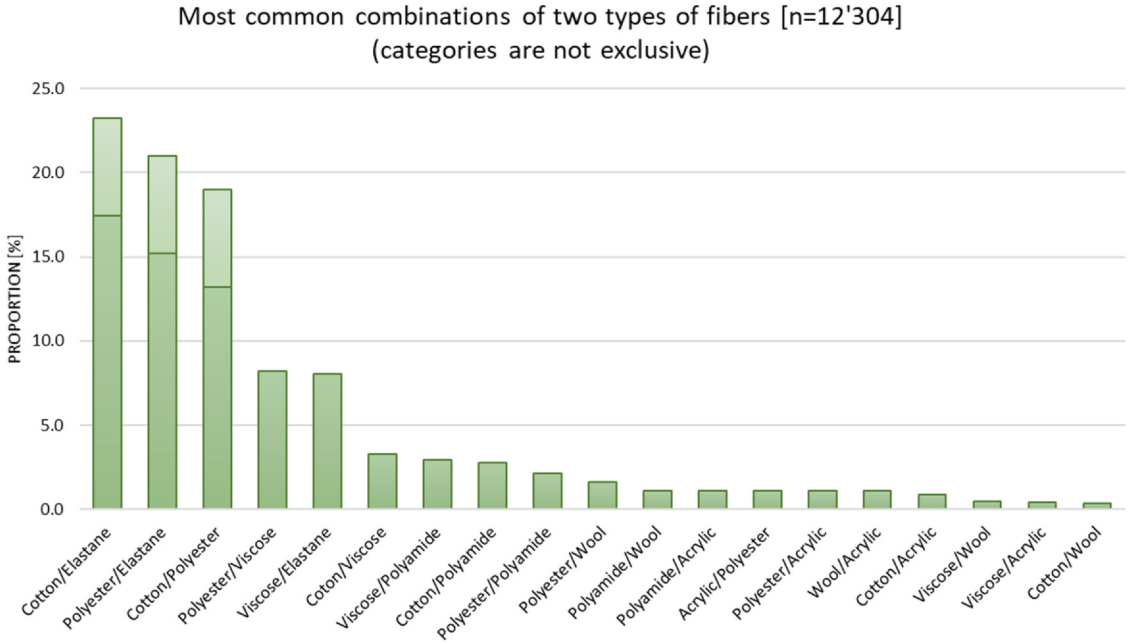

Fig. 8. Most common combinations of two types of fibres as appearing in blended garments (n = 12,304). These statistics include 2-types (39%) and 3+ types (11%) of clothes. The categories are not mutually exclusive, and a 3+ types of fibres garment might appear in multiple categories at the time, as illustrated by the light green (5.81%) corresponding to all three (cotton/elastane/polyester). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

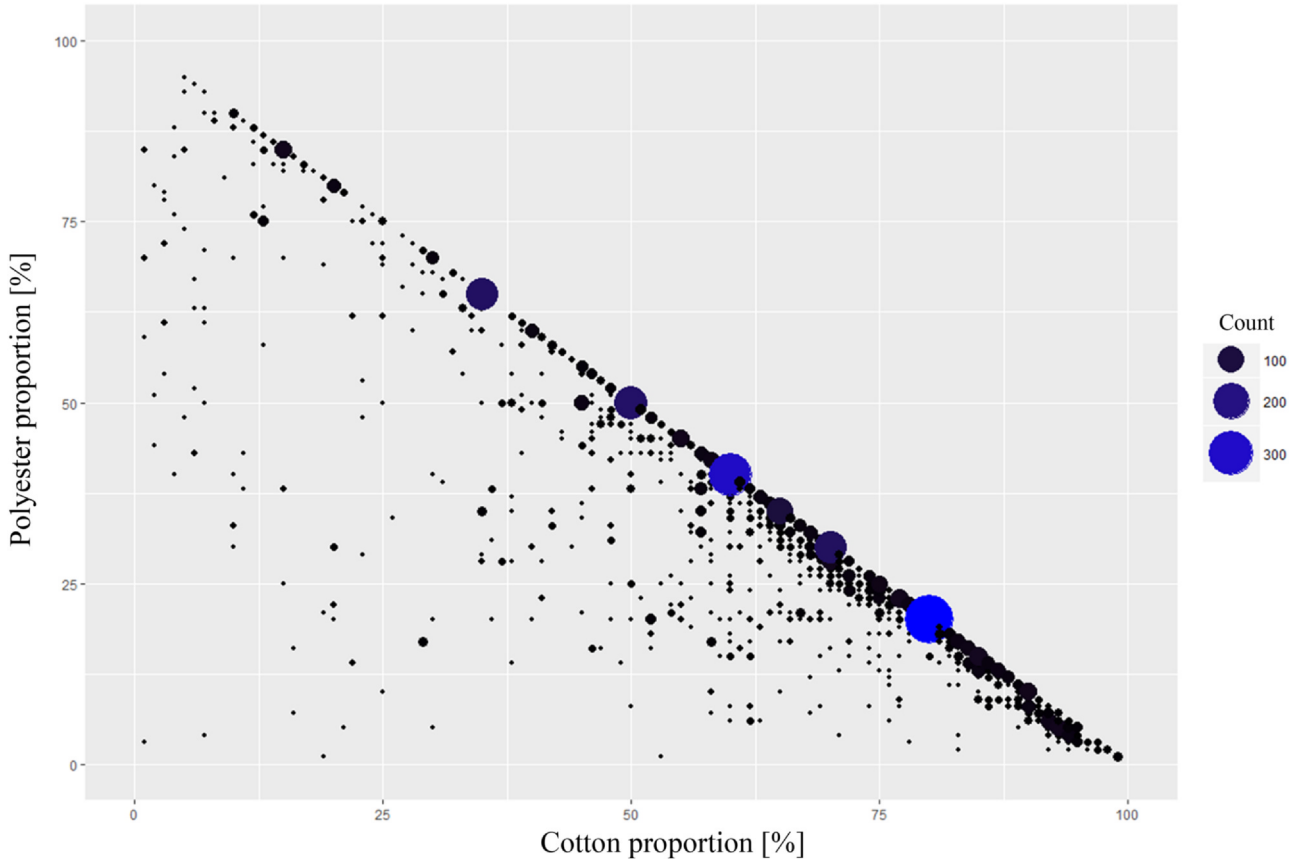

Fig. 9. Illustration of the variety and frequency of occurrence of blended cotton and polyester clothes (n = 3330). The size and color of the dots varies accordingly to their frequency. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

specifically labeled as *multicolor* on the website and the category was kept as is. In total, the garments appear in one of 14 different categories corresponding to the most common colors. Arbitrary decisions in cases of doubt have been decided beforehand (i.e. bourgogne (burgundy) was classified as a red, sand or beach as a beige, coral as an orange, etc).

Fig. 11 presents the most common colors in the complete database of clothes. Black (26.1%) and blue (25.8%) are the two most common fibre colors, and represent more than 50% of all clothes. White and grey come third (13.2%) and fourth (7.7%) respectively, slightly ahead of more colored fibres such as green (6%) and red (5.6%). Surprisingly the results show a relatively high proportion of pink, beige, and yellow fibres, which would normally be considered a minority and very specific colors. We tried comparing these colors to the fashion trends that were in effect at the time of the study, and it is very interesting to note the similarities between the two (Fig. 11). Pink, beige, and yellow are all among the top colors for the spring-summer fashion trends. One hypothesis that demands to be verified in future works is that the frequencies of the four main color blocks (black, blue, white, grey) will remain stable year after year and be invariant to fashion, while the succeeding moderately common colors are much more volatile and the subject of quick change in regards to seasons and fashion trends. To add to this reflection, considering the life span of a garment and the moderate proportion of new garments within the population, it seems difficult to predict if these trends will reflect in the background population of fibres considered for evaluation at all, and if they might already appear after a short amount of time.

Fig. 12 presents the color distribution associated with each type of the most common fibres. This figure shows that cotton fibres have a different color distribution than any other type of fibre. This difference is particularly visible in the highest occurrence of blue, white and grey cotton fibres (i.e. blue denim, white and grey t-shirts). Other types of fibres (i.e. polyester, viscose, …) show a distribution of their colors that is somewhat similar, with the exception of black viscose fibres (i.e. suits, formal clothing). Other than cotton, polyester, and viscose, all other fibres' type and color combinations have a frequency of appearance less than 1%. The probability of observing them as background on any type of surface is therefore pretty low. It is particularly interesting to compare the accuracy of these results with previously published studies. Roux and Margot [5] in particular reported the same 1% conservative estimate for colored synthetic fibres. Compared to Grieve and Biermann study [4] we can grasp the huge increase in polyester use worldwide, going from approximately 1.8% in 1997 to 30% in 2020.

One of the most common use of population studies and frequencies figures is during the evaluation of evidence through likelihood ratio, and more particularly on the estimation of the frequency parameter (*f*). It is sometimes suggested that, in absence of more accurate data, we can consider the fibre type and fibre color as independent and simply multiply them. Independent events would respect the following equation:

P (Type AND Color) = P(Type) * P(Color)

While in the case of dependent events, the probability calculation should be in the form:

*P (Type AND Color) = P(Type) * P(Color=Type)*

Fig. 13 demonstrates that the independency of the two parameters cannot be considered true, by comparing both dependent and independent situations for each of the conditions in Fig. 12. Considering the type of fibre and its color as independent events will

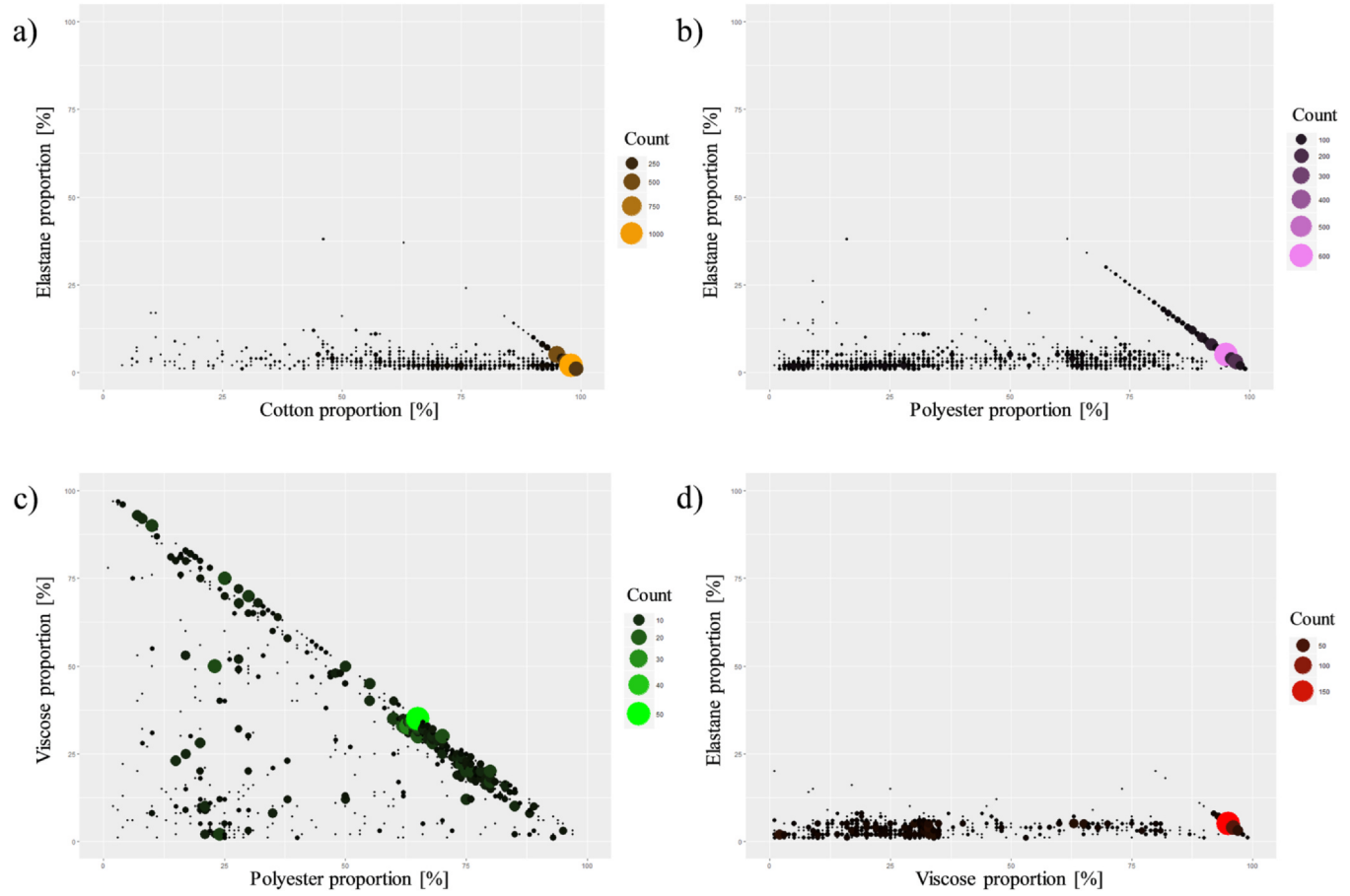

Fig. 10. Illustration of the variety and frequency of occurrence of blended garments of: a) cotton and elastane (n = 4078), b) polyester and elastane (n = 3687), c) polyester and viscose (n = 1444), and d) viscose and elastane (n = 1412). The size and color of the dots varies accordingly to their frequency. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

have a tendency to underestimate the real value. The correlation between the values is only 0.7673 with a linear tendency (red line) showing values about ¼ below their theoretical value if both events were independent (red dashed line). This is for the vast majority explained by the over-representation of black and blue viscose fibres (i.e. formal clothing and suits) that indirectly impacts the proportion of white viscose. These three values have the strongest dependency of color versus type. Interestingly the blue cotton fibres are on par with the theoretical values for independent events, while white cotton fibres slightly deviate from the theoretical distribution. By removing blue, black, and white cotton and viscose fibres, the correlation increases to 0.83. With the exception of viscose and cotton, the rule-of-thumb that we can consider the fibre type and fibre color as independent and simply multiply them, although not recommended, remains a conservative estimate. By calculating the frequency with the dependency we generally obtain a lower probability that gives a higher LR in favor of the accusation proposition.

3.3. Other categories

Among the 68 fields that are scraped by the internet bot are all the information about brand, models, extraction date, and URL, but also other interesting information such as price ranges, categories of clothes, target customers, units, sizes, sleeves length, technology,

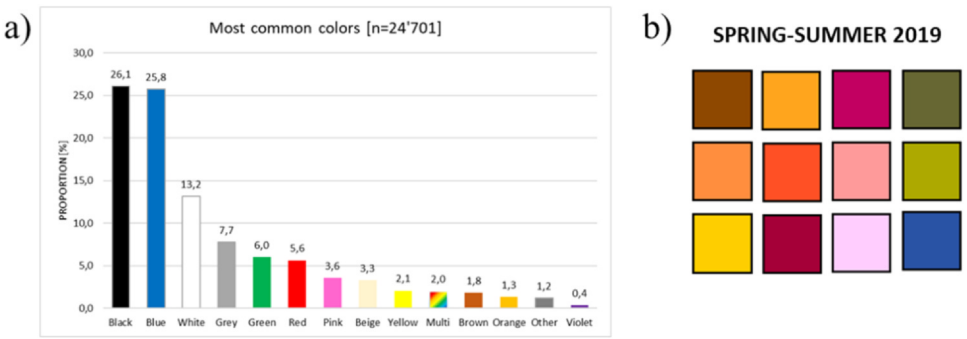

Fig. 11. a) Most frequent colors observed in the garments. b) Comparison with colors listed as trending for the spring-summer 2019 season during which the study was realized. Colors names from upper left to bottom right are respectively: toffee, turmeric, pink peacock, terrarium moss, mango mojito, fiesta, living coral, pepper stem, aspen gold, jester red, sweet lilac, princess blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

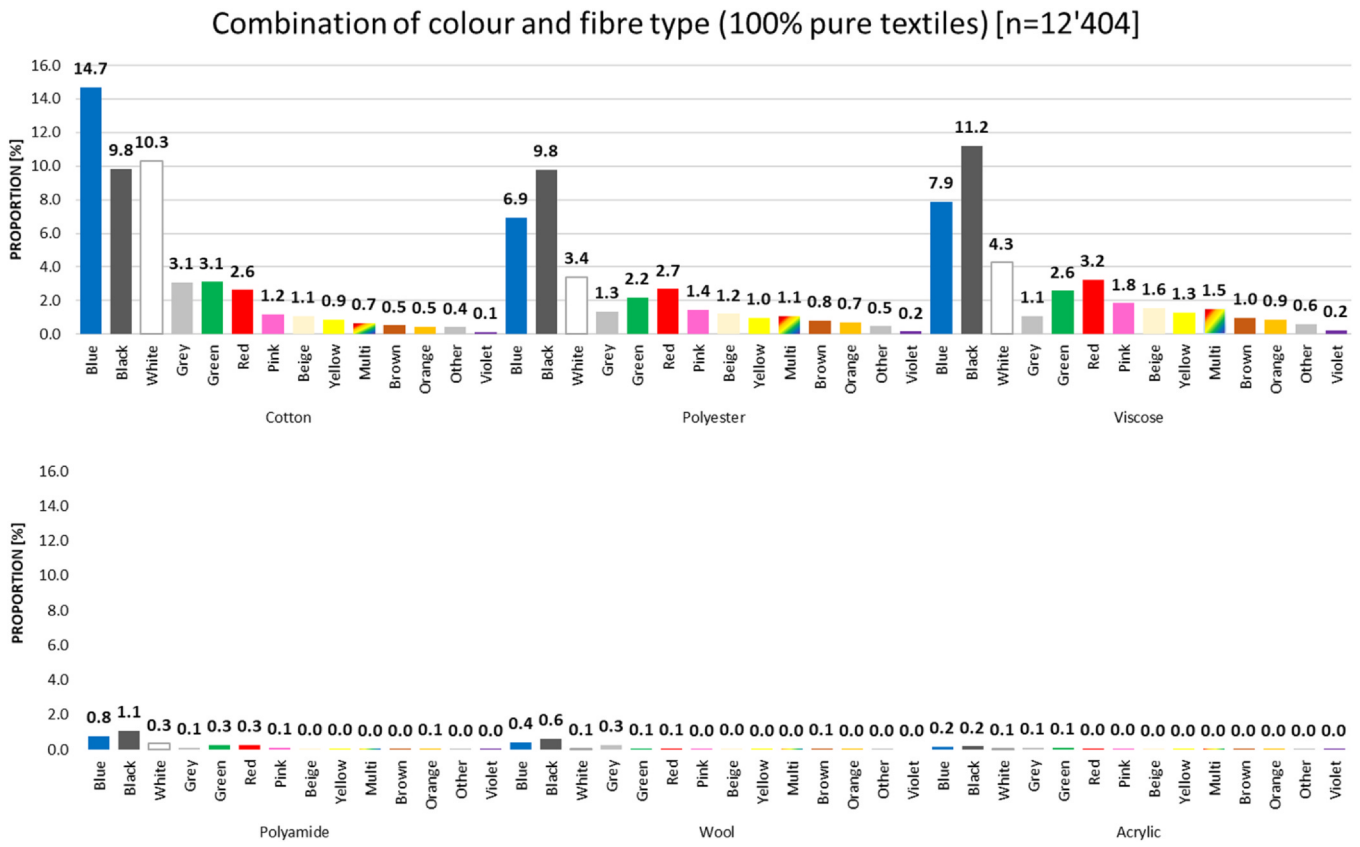

Fig. 12. Color distribution for the most common fibre types demonstrating the dependency of the two parameters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

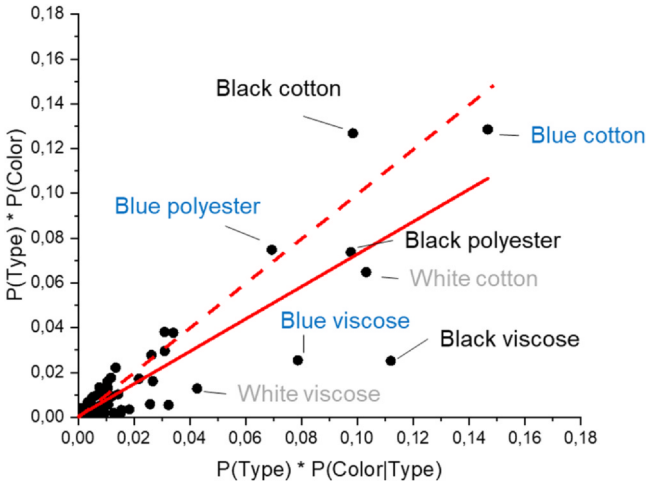

Fig. 13. Dependency testing between types and colors of fibres. The full red line indicates the linear tendency with a 0.76 R^2, while the dashed red line indicated the theoretical linear tendency in the case of completely independent events. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

transparency, etc. All of these fields are not necessarily pertinent for forensic scientists in order to extract statistics but are nonetheless worthy to be aware of. A couple of informative statistics on other fields are given here-below.

428 clothes (out of 1944 coats and vests in the study = 22%) mention a filling fibre in their composition. About 80% of those are composed of polyester, 15% of down and feathers, and 4% of cotton. The remaining 1% being three clothes that are combinations of the aforementioned compositions and two made of 100% viscose.

In total, 1089 clothes (4.4%) contain either a fur component, some down and feathers filling, leather, pearls, shells, or bones. Manufacturers are legally obliged by the 2012 Textile Products Regulation to mention that their products contain some animal parts on the label, usually in form of a sentence such as “non-textile parts of animal origin”. No details about which animal parts are used were obtainable, the field is a binary yes/no category. Wool and other animal hair do not fall under this regulation, as they are considered textile materials.

3.4. Representativity of the data

The figures we obtained compare well to previously published population studies. A few differences are noted with the Biermann and Grieve study of 1996–1998 on sales catalogues [19–21]. First, on the most common garments included in the database: while their study was dominated by pullovers and blouses, we obtained an overwhelming majority of t-shirts and trousers. The difference in fashion between the 90s and 2020 cannot explain these differences by itself. The fact that one of their mail order catalogues was dominated by socks and underwear suggests an over-representation depending on the catalogue’s theme. The same bias might apply to our situation as well. The web scraping technique works on a certain number of rules that unfortunately can introduce biases in our study. The first and most obvious is that the extraction is first based on the date the garment was added to the website. This means that all of the most recent items will be extracted first, providing us with a great estimate of seasons and trends, while the oldest items, no matter how popular they are, will appear in the statistics only if we scrape tens of thousands of clothes. Another important point to emphasize in our population study is that each model of clothing is treated once only, and does not represent its popularity on the

website, nor its purchase count. These information are generally part of the data restricted by the companies, not freely available to scraper bots. One possibility to get around this problem would be to consult the SKUs (stock keeping units), which can be extracted by the scraper. It would seem logical to assume that the clothes with the largest stock are the most popular. This potential avenue of research will also indirectly integrate clothing sizes (S, M, L, XL) variation, as they influence the available SKUs. Lastly, the frequency data were obtained for the entire garment, and a further step is required in order to estimate the quantity (and type) of fibres collected as traces in cases. The shedding capacity of a garment has a very strong influence on the number of fibres that will be transferred. It is also known by forensic scientists that blended garments have a differential shedding where one type of fibres is more easily transferred than the other one, independently of its proportion in the textile [27].

In our opinion, this represents the next level of skill that web scraper robots will be able to achieve in order to provide the forensic scientist with an immediate access to virtually infinite statistics on clothes from the internet. A possible semi-automated knowledge-based platform, scraping different websites in different countries at regular intervals, providing practitioners with up to date statistics on any questions they might have does not seem out of reach. We might soon be able to answer questions as specific as, what was the prevalence of red linen fibres in summer 2019 in Australia? How likely is it to find orange viscose fibres in trousers? Is it uncommon to find pieces of leather on clothes in the USA?

4. Conclusion

Through web crawling and web scraping, we have extracted 68 text-based fields that describe 24,701 clothes from a major online retailer website. These data were extracted at a rate of approximately 1000 piece of clothing per hour and managed through a Kibana/Elasticsearch interface. The frequency figures we obtained compare well to previously published studies based on a physical collection of the fibres, demonstrating the reliability of this procedure. While these studies take time and resources, we have performed our extraction in less than 24 h. We demonstrated that cotton, polyester, viscose and elastane are the 4 main types of fibres used in the textile industry, with an important predominance of cotton and polyester. Elastane, while being very popular in garments (appearing in 32% of the garments), rarely accounts for more than 10% of the mass while cotton is often superior to 80%. Their weighted frequencies of apparition are then 1.61% and 49.24% respectively. The color statistics also demonstrate the preponderance of black, blue, and white, with frequency dependencies to fibre types. More work on the trends and seasonality of color statistics are needed and in reach with such tools. Through this article, we have demonstrated the feasibility, utility, and incredible ease with which web scraper robots are providing real time statistics on fibre populations to assist forensic practitioners in their routine work. The possibilities to develop similar tools in other related fields seems a major avenue of research, potentially fulfilling the drawbacks and time constraints of handmade population studies.

5. Annexes

5.1. Worldwide fibre production

Polyester is often said to be the most common fibre type worldwide. In an attempt to better understand the reported values, we attempted to combine statistics from various sources. A majority of polyester fibres are produced and used for industrial and consumer fabrics (i.e. diapers, filters, tents, ropes, sails, fish nets, bags, …) and home furnishings (i.e. carpets, sofas, curtains, draperies, …). Polyester is known to be everywhere around us, though only 40% of the synthetic fibres produced worldwide are used in clothes, while this proportion goes up to 57% for cotton. In contrast to polyester, cotton is primarily used for apparels (57%), followed by interiors (33%), and industrial applications (10%). Cotton also has multiple household use, mostly towels and bed linen, but is predominantly used in apparels. Cotton still remains one of the most important textile fibres in the world, accounting for around 27.9% of total world fibre use [28]. Fig. 14.

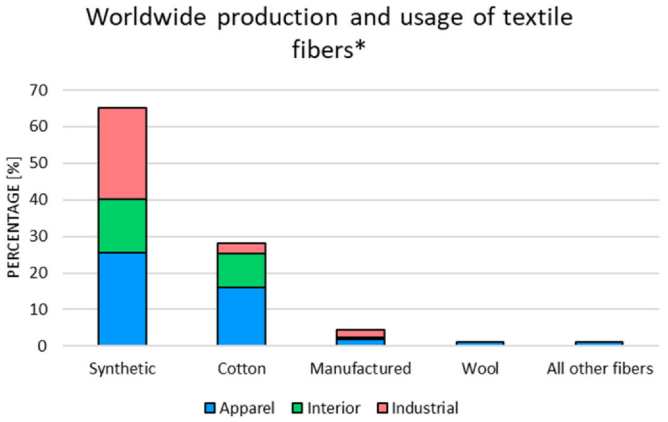

Fig. 14. Worldwide production of textile fibres and their repartition among apparel, interior or industrial usages. The synthetic fibre category comprises polyester, polyamide, acrylic, etc. The manufactured fibres category comprises all the regenerated cellulose fibres such as acetate, viscose, etc. * Precise numbers were not available at the time of redacting this paper.
The percentages represent estimations from multiple sources (textileword.com, thefibreyear.com, statista.com) and must be interpreted with caution.

CRediT authorship contribution statement

Cyril Muehlethaler: Conceptualization, Investigation, Methodology, Formal analysis, Software, Visualization, Writing - original draft. **René Albert:** Conceptualization, Methodology, Software development.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to the Centre International de Criminologie Comparée (CICC)), and the National Science and Engineering Research Council of Canada (NSERC, RGPIN-2019-04827) for financial support.

References

[1] J. Robertson, C. Roux, K.G. Wiggins, Forensic Examination of Fibres, third ed., CRC Press, Boca Raton, 2018.
[2] M.C. Grieve, A survey on the evidential value of fibres and on the interpretation of the findings in fibre transfer cases. Part 1 - fibre frequencies, Science & Justice 40 (2000) 180–200.
[3] R. Palmer, H.J. Burch, The population, transfer and persistence of fibres on the skin of living subjects, Science & Justice 49 (2009) 259–264.
[4] M.C. Grieve, T.W. Biermann, The population of coloured textile fibres on outdoor surfaces, Science & Justice 37 (1997) 231–239.
[5] C. Roux, P. Margot, The population of textile fibres on car seats, Science & Justice 37 (1996) 25–30.
[6] J. Was-Gubala, Comparative population studies of fibres secured in Poland, Czech Republic, and Germany, Problems of Forensic Sciences 60 (2004) 58–77.
[7] S. Cantrell, C. Roux, P. Maynard, J. Robertson, A textile fibre survey as an aid to the interpretation of fibre evidence in the Sydney region, Forensic Science International 123 (1) (2001) 48–53.

[8] R. Watt, C. Roux, J. Robertson, The population of coloured textile fibres in domestic washing machine, Science & Justice 45 (2) (2005) 75–83.
[9] W. Fong, S.H. Inami, Results of a study to determine the probability of chance match occurrence between fibres known to be from different sources, Journal of Forensic Science 31 (1) (1986) 65–72.
[10] S.J. Dignan, K. Murphy, Fibre evidence from fingernail clippings, Canadian Society of the Forensic Science Journal 35 (2002) 17–21.
[11] V. Akulova, D. Vasiliauskiene, D. Talaliene, Further insights into the persistence of transferred fibres on outdoor clothes, Science & Justice 42 (3) (2002) 165–171.
[12] T. Coyle, J. Jones, C. Shaw, R. Friedrichs, Fibres used in the construction of car seats - an assessment of evidential value, Science and Justice 52 (2012) 259–267.
[13] C.M. Ashcroft, S. Evans, I.R. Tebbett, The persistence of fibres in head hair, Journal of the Forensic Science Society 28 (1988) 289–293.
[14] M.C. Grieve, T.W. Biermann, Wool fibres - transfer to ninyl and leather vehicle seats and some observations on their secondary transfer, Science & Justice 37 (1996) 31–38.
[15] J. Jones, T. Coyle, Synthetic flock fibres: a population and target fibre study, Science & Justice 51 (2011) 68–71.
[16] M.C. Grieve, T.W. Biermann, K. Schaub, The individuality of fibres used to provide forensic evidence: not all blue polyesters are the same, Science & Justice 45 (1) (2005) 13–28.
[17] M. Schnegg, R. Palmer, G. Massonnet, Les paramètres clés de l'interprétation des fibres textiles en sciences criminelles. Partie I: occurrence et bruit de fond, Canadian Society of the Forensic Science Journal 51 (1) (2018) 1–25.
[18] G.R. Carroll, W.C. Lalonde, B.D. Gaudette, S.L. Haewley, R.S. Hubert, A computerized database for forensic textile fibres, Canadian Society of Forensic Science Journal 21 (1988) 1–10.
[19] T.W. Biermann, M.C. Grieve, A computerized data base of mail order garments: a contribution toward estimating the frequency of fibre types found in clothing. Part 1: the system and its operation, Forensic Science International 77 (1–2) (1996) 65–73.
[20] T.W. Biermann, M.C. Grieve, A computerized data base of mail order garments: a contribution toward estimating the frequency of fibre types found in clothing Part 2: the content of the data bank and its statistical evaluation, Forensic Science International 77 (1–2) (1996) 75–91.
[21] T.W. Biermann, M.C. Grieve, A computerized data base of mail order garments - a contribution toward estimating the frequency of fibre types found in clothing. Part 3: the content of the data bank - is it representative? Forensic Science International 95 (2) (1998) 117–131.
[22] S. Khalil, M. Fakir, RCrawler: an R package for parallel web crawling and scraping, SoftwareX 6 (2017) 98–106.
[23] E. Ferrara, P. De Meo, G. Fiumara, R. Baumgartner, Web data extraction, applications and techniques: a survey, Knowledge-based systems 70 (2014) 301–323.
[24] J. Myllymaki, Effective web data extraction with standard XML technologies, Computer Networks 39 (2002) 635–644.
[25] United Stats District Court for the Northern District of California United States, Court of Appeals for the Ninth Circuit 2019 HiQ Labs, Inc v Linkedin Corp San Fancisco.
[26] A. Markham, E. Buchanan, Ethical Decision-Making and Internet Research: recommendations from the AoIR Ethics Working Committee (ver. 2.0), 2012.
[27] L. Skokan, A. Tremblay, C. Muehlethaler, Differential shedding: a study of the fiber transfer mechanisms of blended cotton and polyester textiles, Forensic Science International 308 (2020) 110181.
[28] S.J. Kadolph, S.B. Marcketti, Textiles, twelfth ed., Pearson, Boston, 2016.