Qualitative evaluation of ASR adaptation in a lecture context: Application to the PASTEL corpus

Salima Mdhaffar[⋆]*, Yannick Estève*[‡]*, Nicolas Hernandez*[‡‡]*, Antoine Laurent*[⋆]*, Richard Dufour*[‡]*, Solen Quiniou*[‡‡]

[⋆] LIUM - University of Le Mans, France
[‡] LIA - University of Avignon, France
[‡‡] LS2N - University of Nantes , France

firstname.lastname@{univ-lemans,univ-avignon,univ-nantes}.fr

Abstract

Lectures are usually known to be highly specialised in that they deal with multiple and domain specific topics. This context is challenging for Automatic Speech Recognition (ASR) systems since they are sensitive to topic variability. Language Model (LM) adaptation is a commonly used technique to address the mismatch problem between training and test data. In this paper, we are interested in a qualitative analysis in order to relevantly compare the accuracy of the LM adaptation. While word error rate is the most common metric used to evaluate ASR systems, we consider that this metric cannot provide accurate information. Consequently, we explore the use of other metrics based on individual word error rate, indexability, and capability of building relevant requests for information retrieval from the ASR outputs. Experiments are carried out on the PASTEL corpus, a new dataset in French language, composed of lecture recordings, manual chaptering, manual transcriptions, and slides. While an adapted LM allows us to reduce the global classical word error rate by 15.62% in relative, we show that this reduction reaches 44.2% when computed on relevant words only. These observations are confirmed with the high LM adaptation gains obtained with indexability and information retrieval metrics.

Index Terms: Automatic speech recognition, Language model adaptation, Word error rate, Individual word error rate, Indexability metric, Educational applications

1. Introduction

Over these last years, automatic speech recognition systems (ASR) got significant improvements thanks to Deep Neural Networks (DNN) for both acoustic and language models. Nevertheless, such ASR systems are still sensitive to topic variations. In the framework of the ANR PASTEL (Performing Automated Speech Transcription for Enhancing Learning) research project[1] started in 2017, we focus on the capabilities of speech transcription technology in a human learning environment.

In this paper, we target on processing ASR transcriptions of lectures that have been filmed in order to make them indexable, but also capable to be a source of relevant requests with the intention of binding a lecture to external pedagogical resources. Since the PASTEL project also aims to assist teachers to create numerical resources, for instance to create a SPOC (Small Private Online Course), automatic matching recommendations to enrich the pedagogical content with external resources are expected by exploiting automatic transcriptions.

[1] http://www.agence-nationale-recherche.fr/Project-ANR-16-CE33-0007

Assuming that the Word Error Rate (WER) metric is not relevant enough to compare the ASR system performance for such specific tasks [1, 2], we explore the use of more relevant evaluation metrics to analyse the effects of the ASR language model adaptation. Language Model (LM) adaptation of spoken lectures is a well-known issue in the literature [3, 4, 5, 6, 7, 8, 9]. In 2002, [10] authors already demonstrated that the use of a topic-related vocabulary improves speech recognition and indexing for video lectures. The performance of LM adaptation of these works is evaluated using WER or perplexity, but WER does not allow to differentiate between general and domain words in the transcript, and does not take into account the impact of the error according to the final task [11]. And perplexity, computed only on text, does not give real information about the final ASR performance. Some works such as [6, 7] have also used the standard information retrieval metrics (Precision, Recall, F-mesure). However, it has been demonstrated in [12] that the combination of precision and recall with the harmonic mean decreases the importance of errors of deletion and insertion. In addition to the exploration of relevant evaluation metrics to compare ASR outputs in this human learning environment, we also present the PASTEL corpus, that will be very soon distributed[2] under an open source license. This corpus contains video lectures in French language with manual and automatic annotations described in Section 2.

2. The PASTEL corpus

We present here the corpus we used for our experiments, and give a brief overview of the annotation guidelines we followed to extend the data. The data was collected from the project CominOpenCourseware (COCo)[3] which provides a number of videos with potential resources (video, slides, time alignment of the video with the slide changes) and from the canal-U platform[4] which is an online digital video library of higher education. All the videos were manually transcribed by an expert human annotator using the Transcriber[5] tool. The conventions used for the evaluation of transcription campaign [13] served as a guide for manually transcribing registered lectures.

2.1. Topic segmentation

The problem of topics segmentation of lectures is not trivial since it deals with the problem of segmenting mono-thematic material. The main objective of such a task is to automatically

[2] https://github.com/mdhaffar/PASTEL
[3] http://www.comin-ocw.org
[4] https://www.canal-u.tv
[5] http://trans.sourceforge.net

chapter the lecture video to facilitate content access and navigation, but also to help the matching between some parts of the lecture video and external resources. We assumed that a topic boundary can only be located in the vicinity of a slide change during the lecture. Therefore, for each change of slide, a human expert annotated: 1) If there is a topic shift, 2) the exact moment of the topic shift defined as being positioned between two words, 3) the granularity of the topic shift (1 or 2) or if the segment type is an interruption.

Granularity 1 marks that a new notion is started while staying in the same global topic. Granularity 2 is used when a global topic shift occurs which allows us to split the lecture into chapters, each chapter consisting of at least one "granularity 1" segment.

Out of these topic granularities, interruptions, corresponding to moments of public management or technical problems (e.g. video-projector troubleshouting), have been annotated. The annotations were performed with the ELAN software[6].

2.2. Keywords annotation

In-domain words were manually extracted from both manual transcriptions and presentation slides. The underlying objective was to determine how well these specific words were recognized with and without LM adaptation. We consider as in-domain words the linguistic expressions which refer to concepts, objects or entities being essential for the understanding of the current slide or a given transcription. We have included all the scientific and technical terms as well as acronyms and expressions allowing us to go further in the course topic. This annotation was made to courses for which slides were provided.

2.3. Corpus Statistics

The global annotated corpus includes 9 lectures[7]. The total duration of the corpus is about 10 hours. Table 1 presents some statistics of our corpus. The second, third, and fourth columns of the table represent the numbers of "granularity 1" , "granularity 2" and "interruption" segments, respectively. The columns 5 and 6 represent the number of keywords annotated for both transcriptions and slides, respectively. The last one contains the duration of each lecture. The number of speakers in this corpus is 7. As said in previous section, note that 3 lectures ((7), (8) and (9)) were made without slides.

3. Language Model adaptation and ASR description

Adaptation to a new domain requires data from this domain. Assuming that a domain is represented by the material we hold for a lecture, two main issues must be addressed: 1) Where to collect the data? 2) How to collect the relevant one?

We based our work on [14], and use the web as a source for domain data. For automatic speech recognition of lectures, texts of presentation slides are expected to be useful for adapting a language model. Slide titles are essential for giving listeners a quick idea of the content of a course part. This is often the main information on which a listener relies to search and to point out in the course. So, the idea is to use slides title as queries. Queries are submitted to a web search engine (Google) and the returned page links are downloaded. We have limited the search to 400 web pages for each query. The main textual content of a web page must be extracted. Language models have to be trained on cleaned corpora to ensure a certain quality level. In our case where the source of the data is the web, Web pages need to be cleaned to a plain text before a proper analysis is performed on the text. We adapt LM by linear interpolation between an existing LM and the LM trained by web data.

The ASR system is based on the Kaldi toolkit [15]. Acoustic models were trained on about 300 hours of French broadcast news speech with manual transcriptions, using the chain-TDNN training recipe [16]. A sMBR discriminative training [17] was performed on top of chain nnet3 system. The generic (n-gram) language models were trained on these manual transcriptions of speech, but also on newspaper articles, for a total of 1,6 billions of words. The vocabulary of the generic language model contains around 160k words. More details about language models can be found in [18].

Table 1: *Corpus statistics: Duration, number of Granularity 1 units (G1), Granularity 2 units (G2), Interruptions (I), keywords in transcripts (Kw_t) and keywords in slides (Kw_s).*

Lecture	G1	G2	I	Kw_t	Kw_s	Duration
(1)	31	2	2	65	59	1h 04m
(2)	38	10	3	30	37	1h 17m
(3)	35	3	3	121	79	1h 14m
(4)	42	7	7	74	97	1h 05m
(5)	72	5	3	316	158	1h 16m
(6)	52	5	5	131	107	1h 09m
(7)	49	7	0	-	-	1h 21m
(8)	12	7	1	-	-	0h 41m
(9)	57	0	1	-	-	1h 08m
Total	388	46	25	734	537	10h25m

4. Metrics based on word error rate

We report here one of the most used metric to assess the performance of ASR systems and one of its variant which aims at measuring performance on some specific aspects.

4.1. Word Error Rate (WER)

The most used metric to evaluate the quality of an ASR system is the WER [19]. This metric consists of counting the errors according to the predefined types of insertion, deletion and substitution derived by a Levenshtein [20] alignment between manual (reference) and automatic (hypothesis) transcriptions. WER is computed as:

$$WER = \frac{I + S + D}{N} \quad (1)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the number of words in the reference. WER assigns a global error rate on transcriptions.

4.2. Individual Word Error Rate (IWER)

In their study on ASR errors, [21] have analyzed the effects of lexical, prosodic, contextual, and disfluency features for two conversational speech recognition systems. Each feature is composed by a set of words. For example, disfluency features concerns words occurring before and after repetitions, filled

[6]https://tla.mpi.nl/tools/tla-tools/elan

[7]Courses's Name in English *(1): Introduction to computer science* , *(2): Introduction to algorithms*, *(3): fuctions*, *(4): Social networks and graphs*, *(5): Distributed algorithms*, *(6): Natural language processing*, *(7): Republic Architecture*, *(8): Traditional methods*, *(9): imagery*

pauses and fragments. To perform their analysis, they have introduced the Individual Word Error Rate (IWER). For deletion and substitution errors, the principle is the same as WER. We attribute 1 or 0 by comparing the hypothesis and the reference. But for insertion errors, there may be two adjacent reference words that could be responsible and since we have no way to know which word is responsible, we simply assign equal partial responsibility for any insertion errors to both of the adjacent words. So, for the i^{th} reference word, the IWER is calculated as:

$$IWER(w_i) = del_i + sub_i + \alpha.ins_i \quad (2)$$

where $del_i = 1$ if w_i is deleted, $sub_i = 1$ if w_i is substituted and $ins_i =$ number of insertions adjacent to w_i . The parameter α is computed as follows:

$$\alpha = \frac{I}{\sum_{w_i} ins_i} \quad (3)$$

where I is the number of insertions in all the corpus (the total penalty for insertion errors is the same as when computing WER). The IWER for a set of words is the average IWER for individual words:

$$IWER(w_1...w_n) = \frac{1}{n}\sum_{i=1}^{n} IWER(w_i) \quad (4)$$

Usually, the IWER is applied to a specific word list. In this work, we propose to use the IWER to evaluate LM adaptation by targeting in-domain words. In the case of lecture transcription, each lecture has its own domain words that are different from one lecture to another. Consequently, we propose to generalize the metric to make it possible to obtain a global score on a whole corpus of lectures transcripts that does not share the same domain words and not only on a single transcription. This proposition is described in the next section.

5. Intrinsic evaluation: Adaptation performance

The IWER metric offers a way to measure the recognition performance for a given feature on one domain document. We propose a new use case of the IWER metric by considering the in-domain words as feature and we extend it to a more general version, namely the $IWER_{Average}$, to obtain a global score over a corpus of multi-domain transcriptions following the next formula:

$$IWER_{Average} = \frac{1}{\sum_{y=1}^{m} n_m}\sum_{j=1}^{m}\sum_{i=1}^{n_m} IWER(w_i) \quad (5)$$

where m is the number of lecture transcripts and n_m is the number of domain words in the lecture transcript of m.

Experimental results are summarized in Table 2. Four feature configurations are reported. The first line considers all the words in the lexicon and corresponds to the WER result. The next two lines are IWER values computed on in-domain terms from the lecture slides. The last one is respectively IWER values computed on keywords manually extracted from the manual transcripts (*Cf.* Section 2.2).

Results presented in Table 2 show that while an adapted language model permits to reduce the global word error rate by 15.62% (19.46% to 16.42%), we show that this reduction reaches 44.2% when computed on relevant words only (IWER computed on keywords manual extracted from manual transcriptions: from 31% to 17.30%).

Results in Table 2 are computed by adapting LM and by enriching the base dictionary with new words extracted from new downloaded in-domain data (see Section 3). In Table 3, we highlight the impact of the automatic integration of new (expected in-domain) words into the ASR language model vocabulary. After enrichment, the Out-Of-Vocabulary (OOV) rate of words decreases from 0.86% to 0.11%. While global WER shows a relative reduction of 1.91% (16.74% to 16.42%) when new in-domain words are taken into account in the adapted language model, this reduction achieves 19.42% relative (21.47% to 17.3%) with the IWER measure applied to manual transcript keywords. This illustrates that the IWER measure is able to better express the gain provided by the integration of in-domain words in the ASR vocabulary than the global WER.

Table 2: *(%) $IWER_{Average}$ score for 4 features: all the words (=WER), two manual keyword annotations (from slides/transcripts), automatic keyword extraction (slide titles).*

	ASR w/o adaptation	ASR w/ adaptation
All words (= WER)	19.46	16.42
Slides title words	29.52	14.05
Manual slide keywords	32.31	14.52
Manual transcript keywords	31	17.30

Table 3: *(%) $IWER_{Average}$ score for LM adaptation with generic/enriched vocabulary.*

	generic vocabulary	enriched vocabulary
All words (= WER)	16.74	16.42
Manual transcript keywords	21.47	17.30

As a conclusion, if we consider that all errors do not share the same gravity, and that in-domain words are the most important words occurring in a lecture, the $IWER_{Average}$ better expresses the gain brought by LM adaptation than the WER could.

6. Extrinsic evaluation: Document retrieval and indexability

IWER provides an intrinsic evaluation. It is important to know not only the accuracy of an ASR but how errors affect other tasks. This is the goal of an extrinsic evaluation, where the system is evaluated on the tasks based on automatic transcriptions.

6.1. Performance on document retrieval task

One of the PASTEL project's goals is to enrich the transcriptions with external resources. Therefore, it is important to evaluate the impact of the transcription on a document retrieval task. The idea is to generate Web search queries from transcription segments in order to observe how relevant the retrieved documents are. We set as relevant the documents which are retrieved from queries built on the manual transcriptions. We then seek to compare the relevant documents retrieve with the manual (reference) transcription to the one extracted using the automatic transcriptions, with and without adaptation. In practice, we compute the average covering rate on average. Queries were built for each topic segment of "granularity 1" by considering the most salient words of the segments. Saliency was computed based on the words's TF-IDF weights. We experi-

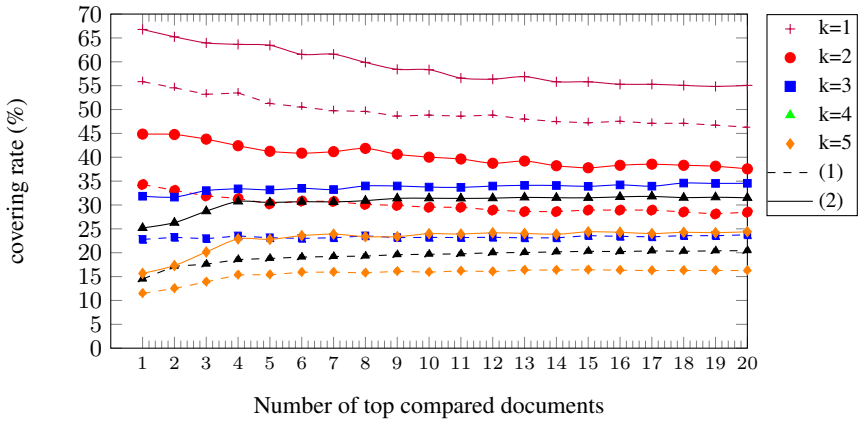

Figure 1: *Retrieval task: Comparison of the covering rate between queries built from manual transcription segments and respectively automatic transcriptions (1) without (bullet marker) and (2) with adaptation (cross marker). Result given on the top 20 retrieved documents, considering k most salient words in queries.*

mented queries with 1 to 5 most salient words. Only the first 20 retrieved documents from using the Google search engine were considered in the comparison. Results show consistently that transcription with adaptation outperforms the transcription without adaptation in terms of retrieving relevant resources for all queries. Results shows consistently that transcription with adaptation outperforms the transcription without adaptation in terms of retrieving relevant resources for all queries.

6.2. Performance on indexability task

In this section, our aim is to evaluate the indexability of transcripts. In other words, we want to determine whether the quality of transcripts plays a role in its indexing and retrieval. Topic segments of "granularity 1" were indexed using the Lemur[8] search engine. Three sets of segments were considered: the ones from manual transcriptions, the ones from automatic transcriptions without adaptation, and the ones from automatic transcriptions with adaptation. Each segment set was searched with queries built on the keywords definitions we used for features in Section 5. Each query returns an ordered list of segments. To evaluate the indexability quality, we use the Spearman's coefficient [22] to measure the rank correlation between manual and automatic transcriptions (resp. without and with adaptation).

Table 4: *Indexability of transcriptions evaluation: Using the same base of queries, comparison of the retrieval results with the Spearman's Rank Correlation coefficient.*

	ASR w/o adaptation	ASR w/ adaptation
Slides' titles words	0.458	0.588
Manual transcripts keywords	0.288	0.516

Table 4 shows an average correlation score through the whole corpus. Results indicate a better indexability in favor of adaptation. Transcriptions with adaptation get the best results with all set of test queries.

[8]https://www.lemurproject.org

6.3. Discussion

We have seen in our experimental framework (Table 2) that the automatic adaptation of LM for speech recognition allows us to reduce the global relative WER by 15.6% (WER from 19.46% to 16.4%). These values, although interesting, do not highlight the impact related to the target tasks for which the automatic transcriptions are generated. In terms of information retrieval task for example, we find an increase in the coverage rate of retrieved documents (compared to documents that would have been found from requests extracted from manual transcriptions) that can exceed 28.5 % (k = 1, level 1, coverage ratio increasing from 56% to 67%). Finally, in terms of indexability, we show in this study that the Spearman correlation rate (compared to the indexation obtained by manual transcriptions) can increase by more than 79% (from 0.288 to 0.516) for the terms the most important documents through the adaptation of LM.

WER is not sufficient to measure the different facets of the quality of automatic transcriptions. For instance, as shown in these results, their use as information retrieval targets or indexed documents cannot be finely appreciated from WER. So, to better evaluate the gain provided by LM adaptation for both indexability and document retrieval, it seems particularly relevant to use specific measures, like the covering rate or the indexability measure based on the Spearman correlation we propose.

7. Conclusion

In this paper, we presented the PASTEL corpus, a new French corpus with manual annotations distributed under an open source licence. We showed that WER does not provide enough information to capture the impact of the LM adaptation in the context of video lecture processing. We suggest to use a variant of the IWER metric that focuses on a specific set of (in-domain) words. In addition, we proposed the use of extrinsic evaluation metrics that measure the capability of building relevant requests for document retrieval (covering rate) and that measure the indexability of automatic transcriptions. Even though we applied these evaluation approaches for measuring the impact of LM adaptation, the use of these metrics can be considered for more general ASR systems comparison in the context of video lecture processing.

8. References

[1] M. A. B. Jannet, O. Galibert, M. Adda-Decker, and S. Rosset, "How to evaluate ASR output for named entity recognition?" in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[2] B. Favre, K. Cheung, S. Kazemian, A. Lee, Y. Liu, C. Munteanu, A. Nenkova, D. Ochei, G. Penn, S. Tratz *et al.*, "Automatic human utility evaluation of ASR systems: Does WER really predict performance?" in *INTERSPEECH*, 2013, pp. 3463–3467.

[3] P. Cerva, J. Silovsky, J. Zdansky, J. Nouza, and J. Malek, "Real-time lecture transcription using ASR for Czech hearing impaired or deaf students," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[4] P. Bell, H. Yamamoto, P. Swietojanski, Y. Wu, F. McInnes, C. Hori, and S. Renals, "A lecture transcription system combining neural network acoustic and language models." in *INTERSPEECH*, 2013, pp. 3087–3091.

[5] A. Park, T. J. Hazen, and J. R. Glass, "Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1. IEEE, 2005, pp. I–497.

[6] H. Yamazaki, K. Iwano, K. Shinoda, S. Furui, and H. Yokota, "Dynamic language model adaptation using presentation slides for lecture speech recognition," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[7] T. Kawahara, Y. Nemoto, and Y. Akita, "Automatic lecture transcription by exploiting presentation slide information for language model adaptation," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4929–4932.

[8] A. Martínez-Villaronga, A. Miguel, J. Andrés-Ferrer, and A. Juan, "Language model adaptation for video lectures transcription," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8450–8454.

[9] J. Miranda, J. P. Neto, and A. W. Black, "Improving ASR by integrating lecture audio and slides," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8131–8135.

[10] W. Hürst, T. Kreuzer, and M. Wiesenhütter, "A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web." in *ICWI*. Citeseer, 2002, pp. 135–143.

[11] D. Luzzati, C. Grouin, I. Vasilescu, M. Adda-Decker, E. Bilinski, N. Camelin, J. Kahn, C. Lailler, L. Lamel, and S. Rosset, "Human annotation of ASR error regions: Is" gravity" a sharable concept for human annotators?" in *LREC*, 2014, pp. 3050–3056.

[12] J. Makhoul, F. Kubala, R. Schwartz, R. Weischedel *et al.*, "Performance measures for information extraction," in *Proceedings of DARPA broadcast news workshop*. Herndon, VA, 1999, pp. 249–252.

[13] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, and K. Choukri, "The ESTER evaluation campaign for the rich transcription of french broadcast news." in *LREC*, 2004.

[14] G. Lecorvé, G. Gravier, and P. Sébillot, "An unsupervised web-based topic language model adaptation method," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 5081–5084.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[16] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI." in *Interspeech*, 2016, pp. 2751–2755.

[17] J. Kaiser, B. Horvat, and Z. Kacic, "A novel loss function for the overall risk criterion based discriminative training of HMM models," in *Sixth International Conference on Spoken Language Processing*, 2000.

[18] A. Rousseau, G. Boulianne, P. Deléglise, Y. Estève, V. Gupta, and S. Meignier, "LIUM and CRIM ASR system combination for the REPERE evaluation campaign," in *International Conference on Text, Speech, and Dialogue*. Springer, 2014, pp. 441–448.

[19] D. S. Pallett, "A look at NIST's benchmark ASR tests: past, present, and future," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 483–488.

[20] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[21] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[22] T. D. Gauthier, "Detecting trends using Spearman's rank correlation coefficient," *Environmental forensics*, vol. 2, no. 4, pp. 359–362, 2001.