ISSN : 2165-4069(Online)
ISSN : 2165-4050(Print)

IJARAI

International Journal of
Advanced Research in Artificial Intelligence

Volume 4 Issue 11

www.ijarai.thesai.org

A Publication of
The Science and Information Organization

Editorial Preface

From the Desk of Managing Editor…

Artificial Intelligence is hardly a new idea. Human likenesses, with the ability to act as human, dates back to Geek mythology with Pygmalion's ivory statue or the bronze robot of Hephaestus. However, with innovations in the technological world, AI is undergoing a renaissance that is giving way to new channels of creativity.

The study and pursuit of creating artificial intelligence is more than designing a system that can beat grand masters at chess or win endless rounds of Jeopardy!. Instead, the journey of discovery has more real-life applications than could be expected. While it may seem like it is out of a science fiction novel, work in the field of AI can be used to perfect face recognition software or be used to design a fully functioning neural network.

At the International Journal of Advanced Research in Artificial Intelligence, we strive to disseminate proposals for new ways of looking at problems related to AI. This includes being able to provide demonstrations of effectiveness in this field. We also look for papers that have real-life applications complete with descriptions of scenarios, solutions, and in-depth evaluations of the techniques being utilized.

Our mission is to be one of the most respected publications in the field and engage in the ubiquitous spread of knowledge with effectiveness to a wide audience. It is why all of articles are open access and available view at any time.

IJARAI strives to include articles of both research and innovative applications of AI from all over the world. It is our goal to bring together researchers, professors, and students to share ideas, problems, and solution relating to artificial intelligence and application with its convergence strategies. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations.

We hope that this journal will inspire and educate. For those who may be enticed to submit papers, thank you for sharing your wisdom.

Editor-in-Chief
IJARAI
Volume 4 Issue 11 November 2015
ISSN: 2165-4069(Online)
ISSN: 2165-4050(Print)

Editorial Board

Reviewer Board Members

CONTENTS

Vicarious Calibration Data Screening Method Based on Variance of Surface Reflectance and Atmospheric Optical Depth Together with Cross Calibration

Kohei Arai 1
1Graduate School of Science and Engineering
Saga University
Saga City, Japan

***Abstract*—Vicarious calibration data screening method based on the measured atmospheric optical depth and the variance of the measured surface reflectance at the test sites is proposed. Reliability of the various calibration data has to be improved. In order to improve the reliability of the vicarious calibration data, some screenings have to be made. Through experimental study, it is found that vicarious calibration data screening would be better to apply with the measured atmospheric optical depth and variance of the measured surface reflectance due to the facts that thick atmospheric optical depth means that the atmosphere contains serious pollution sometime and that large deviation of the surface reflectance from the average means that the solar irradiance has an influence due to cirrus type of clouds. As the results of the screening, the uncertainty of vicarious calibration data from the approximated radiometric calibration coefficient is remarkably improved. Also, it is found that cross calibration uncertainty is poorer than that of vicarious calibration.**

Keywords*—Vicarious calibration; Surface reflectance; Atmospheric Optical Depth; Sky-radiometer; Terra/ASTER; Satellite remote sensing

I. INTRODUCTION

Visible and Near Infrared mounted on earth observation satellites and the short-wavelength infrared radiation thermometer, Alternative calibration using measurement data on the ground and onboard calibration by the calibration mounting system is performed. For example, Marine Observation Satellite-1 [1], Landsat-7 Enhanced Thematic Mapper Plus [2], SeaWiFS [3], High Resolution Visible: HRV/SPOT-1 and 2 [4], Hyperion [5], POLDER [6], etc. by ASTER [7]. The calibration results have been reported. Further, report according to reciprocity with a uniform ground surface [8] over a wide area such as desert radiometer each other overlapping of the observation wavelength range have been made [9].

Vicarious calibration are conducted with consideration of the influence due to the atmosphere obviously [11]. Furthermore, the well-known cross calibration through comparisons among visible to near infrared radiometers onboard same or the different satellites is effective for calibration of the visible to near infrared radiometers in concern [12]. [13], [14]. To conduct the error analysis in the vicarious calibration of visible and near infrared radiometer, Arai et al. made it clear dominant error factors of vicarious calibration accuracy [15]. According to it, most dominant error factors are the surface reflectance measurement followed by optical depth measurement that allows estimation of aerosol property. It is still difficult to estimate the aerosol characteristic and surface reflectance estimations. In order to estimate refractive index of aerosol particles, it is strongly suggested to use skyradiometer[1] or aureole-meters [16], [17]. Since April 2003, Arai et al. have been doing the observation of aerosol by skyradiometer, POM-1 that is manufactured by PREDE Co. Ltd. [18]. The skyradiometer allows measurement of solar direct, diffuse and aureole that results in estimation of refractive index and size distribution of aerosol particles [19]. Nakajima proposes a method of estimating the volume particle size distribution and the complex refractive index [20]. Arai proposes a method for using the Simulated Annealing: SA as inverse problem-solving [21]. Furthermore, improved Modified Langley method as the calibration method of sky-radiometer and as the method for estimation of extraterrestrial solar irradiance as well as atmospheric optical depth is proposed by Arai. The method is for estimating the top of atmosphere radiance with consideration of not only down-welling but also up-welling p and s polarized irradiance and radiance [21].

Reliability of the vicarious calibration data has to be evaluated. Vicarious calibration data are used to be suffered from atmospheric conditions, existing cirrus clouds, smoke from wild fire that happens nearby test sites, enthused gasses from automobiles that situated nearby the test sites, and so on. These are invisible mostly. Therefore, vicarious calibration data are suffered from these influences even if we conducted field campaigns with great care about these. It is possible to find such these influences through careful screening test with the measured data of surface reflectance and optical depth. The method proposed here is to make a screening the vicarious calibration data suffered from the influences for improvement uncertainty of the vicarious calibration data.

In the next section, the method and procedure of the proposed screening method is described followed by experimental data and estimated results. Then conclusion is described with some discussions.

[1] http://skyrad.sci.u-toyama.ac.jp/

II. PROPOSED METHOD

A. Vicarious Calibration Method

Flowchart of the reflectance based vicarious calibration method is shown in Figure 1. At the test site (relatively high surface reflectance of homogeneous area of desert which is situated at comparatively high elevation (thin atmosphere) is desired, field campaign is used to be conducted. At the field campaign, atmospheric optical depth, surface reflectance, column ozone, column water vapor is measured. From the measured atmospheric optical depth, size distribution can be estimated using Angstrome exponent together with extraterrestrial solar irradiance through Langley plot. Total atmospheric optical depth can be divided into Rayleigh scattering component due to atmospheric continuant, and Mie scattering component due to aerosol particles in the scattering components. On the other hand, absorption components due to water vapor, ozone and aerosols are also estimated. Rayleigh scattering component can be estimated with air-temperature and atmospheric pressure on the ground. Absorption due to ozone can be estimated with the absorption coefficient of ozone and the measured column ozone in unit of Dobson Unit. In the visible to near infrared wavelength region, major contribution is from atmospheric continuant (O_2, N_2), water vapor, ozone, and aerosols. Therefore, these contributions in the forms of scattering and absorption have to be taken into account. Through radiative transfer equation solving process (mostly MODTRAN code is used to use) with the estimated influencing aforementioned parameters, Top of the Atmosphere: TOA radiance (at sensor radiance) can be estimated. Then the estimated TOA radiance is compared with satellite sensor data (Digital Number; DN is converted to radiance). Thus gain can be calibrated. This gain degradation is called as Radiometric Calibration Coefficient: RCC. It is referred to Vicarious Calibration Coefficient: RCCvic. On the other hands, most of visible to near infrared radiometer onboard satellites has own onboard calibration system. It provides Onboard Calibration Coefficient (OBC).

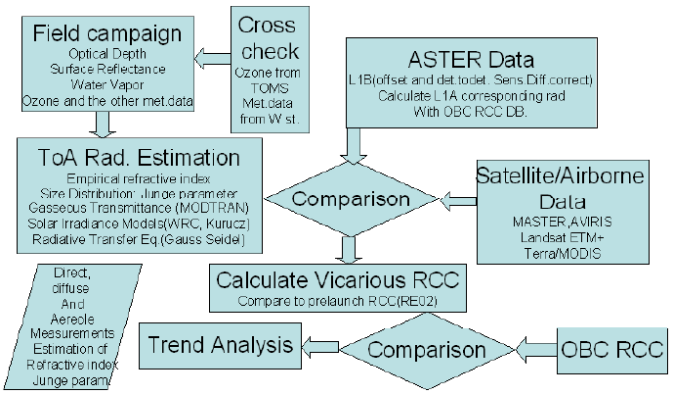


Fig. 1. Flowchart of reflectance based vicarious calibration method

Then it is possible to compare both coefficients.

B. Error Budget Analysis

There are 8 error sources for the vicarious calibration. The result from error budget analysis of the vicarious calibration method is shown in Table 1.

TABLE I. ERROR BUDGET FOR REFLECTANCE BASED VICARIOUS CALIBRATION METHOD

Error sources	Error Type	Error (%)
Optical depth	Random	1.5
Surface reflectance measurement instrument	Random	2
Standard plaque	Systematic	1
Averaging	Random	0
Refractive index	Random	1.8
Size distribution	Random	2
Radiative transfer code	Systematic	1
Registration	Random	1
RSS		4.06

Optical depth measurement has error of 1.5% while surface reflectance measurement instrument has 2% of error which includes Bi-Directional Reflectance Distribution Function: BRDF effect. Standard plaque is used as reference of reflectance and has 1% of error. Estimation accuracy of refractive index and size distribution is not high enough. 1.8 and 2 % of errors are suspected for each. Radiative transfer code has 1% of error while 1% of error is suspected due to registration of test site (location identification) in the satellite sensor image. Thus 4.06% error is suspected for vicarious calibration.

C. Vicarious Calibration Data Screening Method

Reliability of the vicarious calibration data has to be evaluated. Vicarious calibration data are used to be suffered from atmospheric conditions, existing cirrus clouds, smoke from wild fire which happens nearby test sites, enthused gasses from automobiles which situated nearby the test sites, and so on. These are invisible mostly. Therefore, vicarious calibration data are suffered from these influences even if we conduct field campaigns with great care about these. It is possible to find such these influences through careful screening test with the measured data of surface reflectance and optical depth. The method proposed here is to make a screening the vicarious calibration data suffered from the influences for improvement uncertainty of the vicarious calibration data. There are two major factors, optical depth and standard deviation of the measured surface reflectance. By using threshold, vicarious calibration data can be screened.

D. Uncertainty of Vicarious Calibration

Time series of RCCvic data can be approximated with appropriate function (Usually single exponential function of "a EXP(-b d) + c") in the sense of trend analysis. Let be RCCvic' denotes the approximated RCCvic. Then uncertainty of vicarious calibration can be expressed in equation (1).

$$U=SQRT(\Sigma(RCCvic-RCCvic')^2/n_i(n_i-p_i)) \quad (1)$$

where n and p denotes the number of vicarious calibration data and the condition number, respectively. Thus uncertainty of the vicarious calibration can be calculated.

III. EXPERIMENT

A. Trend of the Vicarious Calibration Data

One of examples of vicarious calibration data of ASTER/VNIR onboard Terra satellite is shown in Figure 2

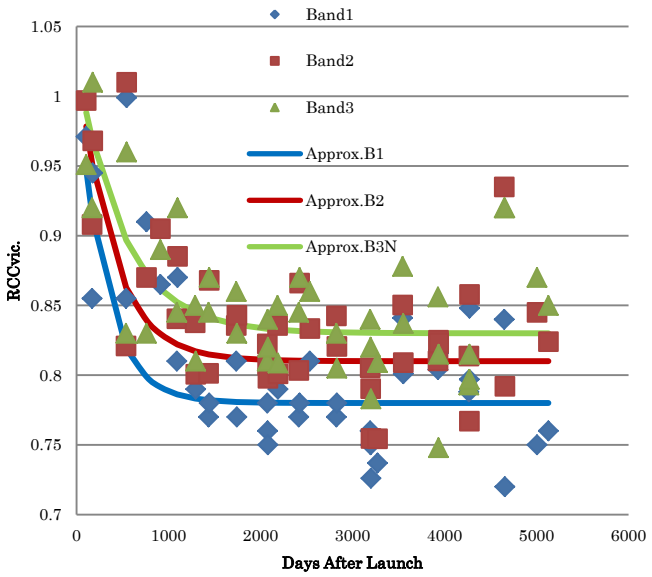


Fig. 2. Vicarious calibration data of ASTER/VNIR onboard Terra satellite

ASTER/VNIR onboard Terra satellite was launched in December 1999. This is approximately 15 years data. Solid lines are approximated function with the single exponential function. It is clear that vicarious calibration data are scattered because there are many data which are influenced by the smoke due to wild fire, exhausted gasses from automobile, cirrus, etc. VNIR has three spectral bands, 1 to 3 which are corresponding to green, red and near infrared bands.

Table 2 shows the coefficients of the approximation function of vicarious calibration data as a function of days after launch, x.

TABLE II. COEFFICIENTS OF THE APPROXIMATION FUNCTION OF VICARIOUS CALIBRATION DATA AS A FUNCTION OF DAYS AFTER LAUNCH, X

y = b0 * exp(-b1 * x) + b2	Band1	Band2	Band3
b0	0.2374657	0.2227815	0.194918
b1	0.0033325	0.0026667	0.0019807
b2	0.7551861	0.83	0.8468567

The difference between vicarious calibration data and the approximated vicarious calibration data is shown in Table 3 while uncertainty of vicarious calibration data defined in equation (1) is shown in Table 4, respectively.

TABLE III. DIFFERENCE BETWEEN VICARIOUS CALIBRATION DATA AND THE APPROXIMATED VICARIOUS CALIBRATION DATA

Band1	Band2	Band3
0.161	0.071	0.071

TABLE IV. UNCERTAINTY OF VICARIOUS CALIBRATION DATA DEFINED IN EQUATION (1)

Band1	Band2	Band3
0.0101579	0.0067297	0.0067447

B. Vicarious Calibration Data After the Screening

Figure 3 shows the vicarious calibration data trend after the screening.

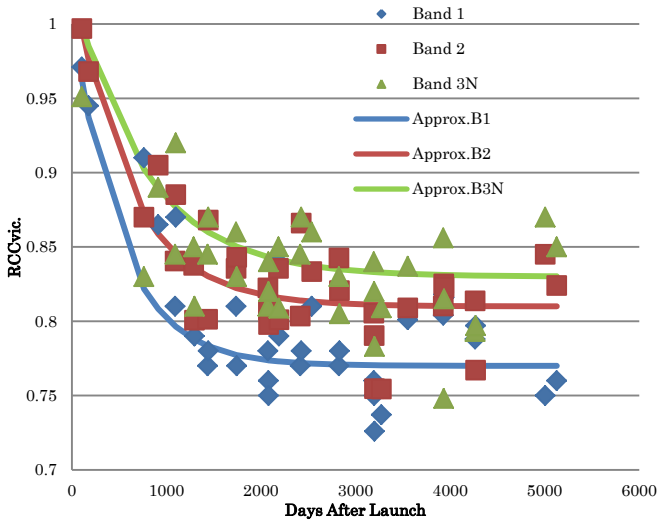
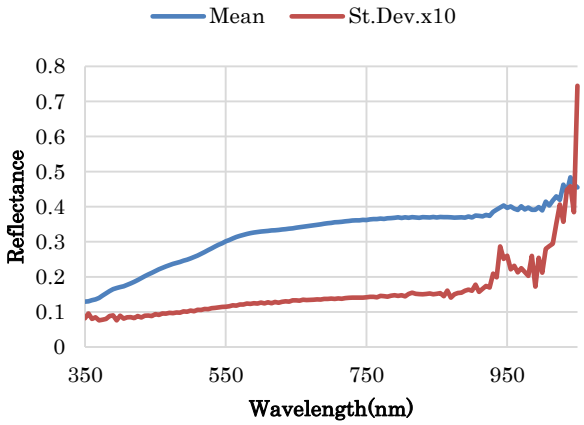
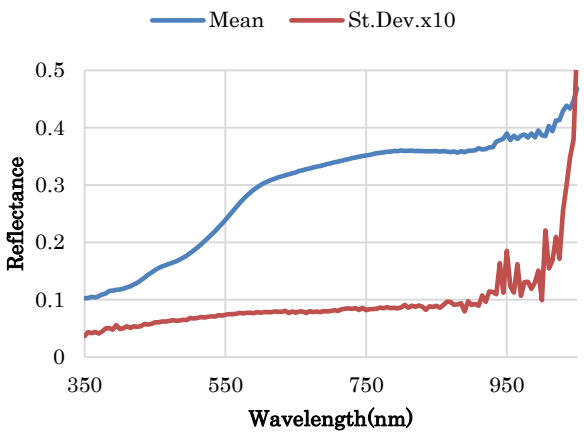


Fig. 3. Vicarious calibration data trend after the screening

One of the examples of the measured surface reflectance for test site of Railroad Valley Playa in Nevada, USA which is acquired on 25 September 2015 is shown in Figure 4 while one if examples of the measured surface reflectance without screening of the test site Ivanpah Playa in California, USA which is acquired on 18 September 2015 is shown in Figure 4 (b), respectively.



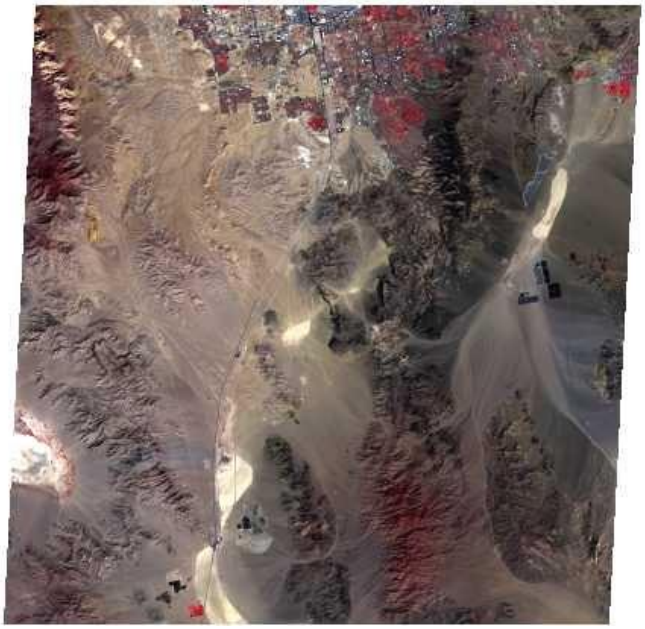
(a)Railroad Valley Playa



(b)Ivanpah Playa

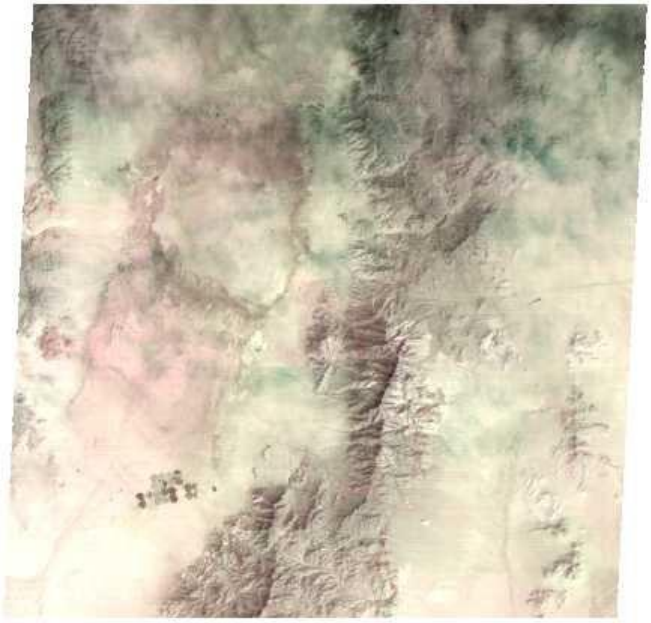
Fig. 4. Measured surface reflectance at the test sites

As shown in Figure 4, the measured surface reflectance between Ivanpah and Railroad Valley Playas are almost same while standard deviation of surface reflectance is quite different (standard deviation of Ivanpah playa is approximately 50% greater than that of Railroad Valley playa). Figure 5 (a) shows ASTER/VNIR image of Ivanpah playa while Figure 5 (b) shows that of Railroad Valley playa. Meanwhile, Figure 5 (c) shows ASTER/TIR image of Railroad Valley playa which shows the suspected existing cirrus. Although cirrus clouds cannot be seen in the ASTER/VNIR image of Railroad Valley playa, ASTER/TIR image shows existing of cirrus clouds almost all over the test site area. During the surface reflectance measurement, solar irradiance is changed a lot due to the existing cirrus. Therefore, the standard deviation of the measured surface reflectance is 50% much greater than that of Ivanpah playa. We would better to omit such unreliable vicarious calibration data.



(a)VNIR image of Ivanpah

(b)VNIR image of Railroad Valley



(c)TIR image of Railroad Valley

Fig. 5. ASTER/VNIR and TIR images of Ivanpah and Railroad Valley playas

Table 5 shows the coefficients of the approximation function of vicarious calibration data after the screening as a function of days after launch, x.

TABLE V. COEFFICIENTS OF THE APPROXIMATION FUNCTION OF VICARIOUS CALIBRATION DATA AFTER THE SCREENING AS A FUNCTION OF DAYS AFTER LAUNCH, X

y = b0 * exp(-b1 * x) + b2	Band1	Band2	Band3
b0	0.2374657	0.2227815	0.194918
b1	0.002	0.0016667	0.0013
b2	0.77	0.81	0.83

The difference between vicarious calibration data and the approximated vicarious calibration data is shown in Table 6 while uncertainty of vicarious calibration data defined in equation (1) is shown in Table 7, respectively.

TABLE VI. DIFFERENCE BETWEEN VICARIOUS CALIBRATION DATA AND THE APPROXIMATED VICARIOUS CALIBRATION DATA

Band1	Band2	Band3
0.040	0.024	0.036

TABLE VII. UNCERTAINTY OF VICARIOUS CALIBRATION DATA DEFINED IN EQUATION (1)

Band1	Band2	Band3
0.0061288	0.0047501	0.0058725

It is found that uncertainty of vicarious calibration can be improved remarkably in particular for Band 1.

C. Error Budget Analysis of Cross Calibration

MISR and MODIS sensors are onboard Terra satellite as well. The spectral coverage of MISR and MODIS are overlapped with ASTER/VNIR. Therefore, cross calibration can be done for VNIR and MISR (VNIR Band 1, 2, 3) and VNIR and MODIS (VNIR Band 2 and 3). Due to the fact that MODIS does not have the corresponding band for VNIR Band 1, cross calibration cannot be done. The results from error budget analysis are shown in Table 8. In the proposed cross calibration, it is conducted at the same dates for field campaigns because the vicarious calibration data can be used for cross calibration.

TABLE VIII. ERROR BUDGET FOR CROSS CALIBRATION

Error items	Error sources	Error (%)
Uncertainty of the instruments for comparison	MISR,MODIS	4.06
Registration	Uniformity of the surface reflectance	2
Spectral response	Surface reflectance	1.5
	Atmospheric effect	1
RSS		4.87

D. Cross Calibration Results with MISR

Figure 6 shows the cross calibration data trend derived RCC (RCCcross) with MISR. It is possible to approximate with the same function of the single exponential function as a function of days after launch, x. The coefficients of the approximation function are shown in Table 9 while the difference between cross calibration data and the approximated data are shown in Table 10. Uncertainty defined as equation (1) for cross calibration with MISR is shown in Table 11. In comparison to the uncertainty of the vicarious calibration, cross calibration accuracy is not better than vicarious calibration obviously. It, however, is useful to find the biases between ASTER/VNIR and MISR.

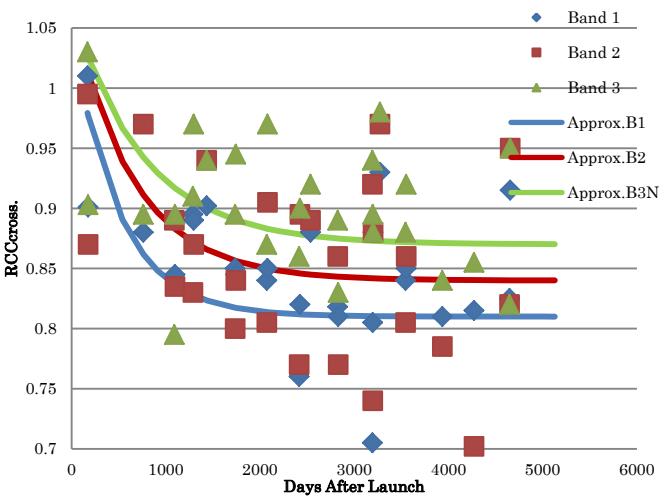


Fig. 6. Cross calibration data of ASTER/VNIR with MISR onboard Terra satellite

TABLE IX. COEFFICIENTS OF THE APPROXIMATION FUNCTION OF CROSS CALIBRATION DATA AFTER THE SCREENING AS A FUNCTION OF DAYS AFTER LAUNCH, X

y = b0 * exp(-b1 * x) + b2	Band1	Band2	Band3
b0	0.2374657	0.2227815	0.194918
b1	0.002	0.0015	0.0013
b2	0.81	0.84	0.87

TABLE X. DIFFERENCE BETWEEN CROSS CALIBRATION DATA AND THE APPROXIMATED VICARIOUS CALIBRATION DATA

Band1	Band2	Band3
8.985	9.844	10.508

TABLE XI. UNCERTAINTY OF CROSS CALIBRATION DATA DEFINED IN EQUATION (1)

Band1	Band2	Band3
0.0759393	0.0794899	0.0821237

E. Cross Calibration Results with MODIS

Figure 7 shows the cross calibration data trend derived RCC (RCCcross) with MODIS. It is possible to approximate with the same function of the single exponential function as a function of days after launch, x. The coefficients of the approximation function are shown in Table 12 while the difference between cross calibration data and the approximated data are shown in Table 13. Uncertainty defined as equation (1) for cross calibration with MODIS is shown in Table 14. In comparison to the uncertainty of the vicarious calibration, cross calibration accuracy is not better than vicarious calibration obviously. It, however, is useful to find the biases between ASTER/VNIR and MODIS.

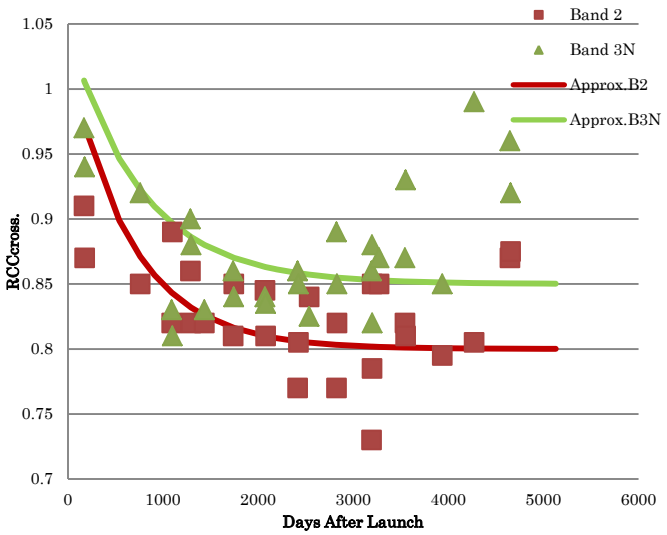


Fig. 7. Cross calibration data of ASTER/VNIR with MODIS onboard Terra satellite

TABLE XII. COEFFICIENTS OF THE APPROXIMATION FUNCTION OF CROSS CALIBRATION DATA AFTER THE SCREENING AS A FUNCTION OF DAYS AFTER LAUNCH, X

y = b0 * exp(-b1 * x) + b2	Band1	Band2	Band3
b0	-	0.222781456	0.194918032
b1	-	0.0015	0.0013
b2	-	0.8	0.85

TABLE XIII. DIFFERENCE BETWEEN CROSS CALIBRATION DATA AND THE APPROXIMATED VICARIOUS CALIBRATION DATA

Band1	Band2	Band3
-	8.884	10.035

TABLE XIV. UNCERTAINTY OF CROSS CALIBRATION DATA DEFINED IN EQUATION (1)

Band1	Band2	Band3
0	0.075513481	0.080253544

As the results from the uncertainty evaluations of the cross calibration between ASTER/VNIR and MISR as well as MODIS, it is almost same between cross calibrations of ASTER/VNIR and MISR as well as MODIS. It is also found that the uncertainty of cross calibration is poorer than that of vicarious calibration.

IV. CONCLUSION

Vicarious calibration data screening method based on the measured atmospheric optical depth and the variance of the measured surface reflectance at the test sites is proposed. Reliability of the various calibration data has to be improved. In order to improve the reliability of the vicarious calibration data, screening has to be made. Through experimental study, it is found that vicarious calibration data screening would be better to apply with the measured atmospheric optical depth and variance of the measured surface reflectance due to the facts that thick atmospheric optical depth means that the atmosphere contains serious pollution sometime and that large deviation of surface reflectance from the average means that the solar irradiance has influence due to cirrus type of clouds. As the results of the screening, the uncertainty of vicarious calibration data from the approximated radiometric calibration coefficient is remarkably improved. Also, it is found that cross calibration uncertainty is poorer than that of vicarious calibration.

ACKNOWLEDGEMENTS

Author would like to thank Dr. Akira Ono of National Institute of Advanced Industrial Science and Technology: AIST for his initiating of this research works. Also, author would like to thank Dr. Satoshi Tsuchida and his research staff, Japanese Space Systems: JSS members together with Dr. Kurtis Thome of NASA/GSFC as well as Prof. Dr. Stuart Biggar and his research staff for their contributions of the field experiments and valuable discussions.

REFERENCES

[1] Arai K., Preliminary assessment of radiometric accuracy for MOS-1 sensors, International Journal of Remote Sensing, 9, 1, 5-12, 1988.

[2] Barker, JL, SK Dolan, et al., Landsat-7 mission and early results, SPIE, 3870, 299-311, 1999.

[3] Barnes, RA, EEEplee, et al., Changes in the radiometric sensitivity of SeaWiFS determined from lunar and solar based measurements, Applied Optics, 38, 4649-4664, 1999.

[4] Gellman, DI, SF Biggar, et al., Review of SPOT-1 and 2 calibrations at White Sands from launch to the present, Proc. SPIE, Conf.No.1938, 118-125, 1993.

[5] Folkman, MA, S.Sandor, et al., Updated results from performance characterization and calibration of the TRWIS III Hyperspectral Imager, Proc. SPIE, 3118-17, 142, 1997.

[6] Hagolle, O., P.Galoub, et al., Results of POLDER in-flight calibration, IEEE Trans. On Geoscience and Remote Sensing, 37, 1550-1566, 1999.

[7] Thome, K., K. Arai, S. Tsuchida and S. Biggar, Vicarious calibration of ASTER via the reflectance based approach, IEEE transaction of GeoScience and Remote Sensing, 46, 10, 3285-3295, 2008.

[8] Cosnefroy, H., M.Leroy and X.Briottet, Selection and characterization of Saharan and Arabian Desert sites for the calibration of optical satellite sensors, Remote Sensing of Environment, 58, 110-114, 1996.

[9] Arai, K., In-flight test site cross calibration between mission instruments onboard same platform, Advances in Space Research, 19, 9, 1317-1328, 1997.

[10] Nicodemus, FE, "Directional Reflectance and Emissivity of an Opaque Surface", Applied Optic (1965), or FE Nicodemus, JC Richmond, JJ Hsia, IW Ginsber, and T. Limperis, "Geometrical Considerations and Nomenclature for Reflectance, ", NBS Monograph 160, US Dept. of Commerce (1977).

[11] Slater, PN, SFBiggar, RGHolm, RDJackson, Y.Mao, MSMoran, JMPalmer and B.Yuan, Reflectance-and radiance-based methods for the in-flight absolute calibration of multispectral sensors, Remote Sensing of Environment, 22, 11-37, 1987.

[12] Kieffer, HH and RL Wildey, Establishing the moon as a spectral radiance standard, J., Atmosphere and Oceanic Technologies, 13, 360-375, 1996.

[13] Arai, K., Atmospheric Correction and Residual Errors in Vicarious Cross-Calibration of AVNIR and OCTS Both Onboard ADEOS, Advances in Space Research, 25, 5, 1055-1058, 1999.

[14] Liu;. JI, Z. Li, YL Qiao, Y.-J. Liu, and Y.-X. Zhang, A new method for cross-calibration of two satellite sensors, Int J. of Remote Sensing, 25 , 23 5267-5281 , 2004 .

[15] Kohei Arai , error analysis of vicarious calibration of satellite visible and near infrared radiometer based KJThome, reflectance , Japan Photogrammetry Journal , Vol.39, No.2, pp.99-105, (2000) .

[16] Arai, K., Vicarious calibration for solar reflection channels of radiometers onboard satellites with deserted area of data, Advances in Space Research, 39, 1, 13-19, 2006.

[17] Arai, K. and X.Liang, Characterization of aerosols in Saga city areas, Japan withy direct and diffuse solar irradiance and aureole observations, Advances in Space Research, 39, 1, 23-27, 2006.

[18] Kohei Arai , vicarious calibration of ASTER / VNIR based on long-term observations of the optical properties of aerosols in Saga , Journal of the Remote Sensing Society of Japan , 28,3,246 over 255,2008

[19] Kohei Arai , applied linear algebra , modern science , Inc. , 2006

[20] Nakajima, T., M.Tanaka and T. Yamauchi, Retrieval of the optical properties of aerosols from aureole and extinction data, Applied Optics, 22, 19, 2951-2959, 1983.

[21] Kohei Arai , Xing Ming Liang, Estimation of complex refractive index of aerosol using direct solar direct, diffuse and aureole by simulated annealing, and polarized radiance -simultaneous estimation of particle size distribution and refractive index, Journal of the Remote Sensing Society of Japan , Vol.23, No.1, pp .11-20,2003 .

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission "A" of ICSU/COSPAR since 2008. He wrote 33 books and published 500 journal papers. He is now Editor-in-Chief of IJACSA and IJISA.

Method for Surface Reflectance Estimation with MODIS by Means of Bi-Section between MODIS and Estimated Radiance as well as Atmospheric Correction with Skyradiometer

Kohei Arai 1
1Graduate School of Science and Engineering
Saga University
Saga City, Japan

Kenta Azuma 2
2 Cannon Electronics Inc.
Tokyo, Japan

***Abstract*—Method for surface reflectance estimation with MODIS by means of bi-section algorithm between MODIS and estimated radiance is proposed together with atmospheric correction with sky-radiometer data. Surface reflectance is one of MODIS products and is need to be improved its estimation accuracy. In particular the location near the skyradiometer or aeronet sites of which solar direct, aureole and diffuse radiance are measured, it is possible to improve the estimation accuracy of surface reflectance. The experiment is conducted at the skyradiometer site which is situated at Saga University. There is Ariake Sea near the Saga University. It is rather difficult to estimate surface reflectance of the sea surface because the reflectance is too low in comparison to that of land surface. In order to improve surface reflectance estimation accuracy, atmospheric correction is mandated. Atmospheric correction method is also proposed by using skyradiometer data. Through the experiment, it is found that these surface reflectance estimation and atmospheric correction methods are validated.**

***Keywords*—Sea surface reflectance; Atmospheric correction; Sky-radiometer; MODIS; satellite remote sensing**

I. INTRODUCTION

Sea surface reflectance, water leaving radiance are fundamental characteristics and are importance parameters for the estimations of chlorophyll-a concentration, suspended solid, etc. Therefore, there is a strong demand to improve sea surface reflectance estimation accuracy. In order to improve surface reflectance, it is required to improve atmospheric correction accuracy. In the visible to near infrared wavelength region, the absorption components due to water vapor, ozone, aerosols, and the scattering due to atmospheric molecules, aerosols are major components. In particular, aerosol absorption and scattering (Mie scattering) is not so easy to estimate rather than scattering component due to atmospheric molecules (Rayleigh scattering). After the estimation of these components, radiative transfer equation has to be solved for the atmospheric correction. This is the process flow of the atmospheric correction [1]-[6]. Also, atmospheric component measurement, estimation, retrievals are attempted together with sensitivity analysis [7]-[17]. It is still difficult to estimate the aerosol characteristic estimation which results in difficulty on surface reflectance estimations.

The method proposed here is based on ground based Skyradiometer[1] which allows aerosol refractive index and size distribution through measurements of spectral optical depth through direct and aureole as well as diffuse solar irradiance. These are measured aerosol refractive index and size distribution, not estimated refractive index and size distribution. Therefore, it is expected that atmospheric correction can be done much precisely rather than estimation without sky radiometer data.

One of the examples are shown here for sea surface reflectance estimation with MODIS[2] data of Ariake Sea in Japan. Method for surface reflectance estimation with MODIS by means of bi-section algorithm between MODIS and estimated radiance is proposed together with atmospheric correction with sky-radiometer data. Surface reflectance is one of MODIS products and is need to be improved its estimation accuracy. In particular the location near the skyradiometer or aeronet sites of which solar direct, aureole and diffuse radiance are measured, it is possible to improve the estimation accuracy of surface reflectance. The experiment is conducted at the skyradiometer site which is situated at Saga University. There is Ariake Sea near the Saga University. It is rather difficult to estimate surface reflectance of the sea surface because the reflectance is too low in comparison to that of land surface. In order to improve surface reflectance estimation accuracy, atmospheric correction is mandated. Atmospheric correction method is also proposed by using skyradiometer data.

In the next section, the method and procedure of the experimental study is described followed by experimental data and estimated results. Then conclusion is described with some discussions.

[1] http://skyrad.sci.u-toyama.ac.jp/
[2] http://modis.gsfc.nasa.gov/

II. PROPOSED METHOD

A. The Proposed Method

Atmospheric correction is important for estimation of surface reflectance (Remote Sensing Reflectance [3]) in particular for estimation of sea surface reflectance estimation. The proposed atmospheric correction method is based on Skyradiometer data derived aerosol size distribution and refractive index. The aerosol refractive index and size distribution can be estimated by using SkyradPack[4] with direct and diffuse solar irradiance those are measured with skyradiometer. Scattering phase function, extinction as well as scattering and absorption coefficients and asymmetry index are then estimated by using mie2new software code with the estimated refractive index and size distribution. Meantime, geometric relation among the satellite sensor of MODIS onboard AQUA satellite is estimated with MODIS Level 1B product. These estimated values are set to the input parameters (Tape 5) of MODTRAN[5] of atmospheric radiative transfer code. Other input parameters are set at the default values. In the process of estimation of the Top of the Atmosphere: TOA Radiance, MODTRAN is used.

The well-known bi-section method is used for estimation surface reflectance because TOA radiance is getting large in accordance with sea surface reflectance. First, initial value of the sea surface reflectance is assumed. By using the initial sea surface reflectance together with the aforementioned input parameters, all the required input parameters are set for MODTRAN. Then TOA radiance can be estimated based on MODTRAN. The estimated TOA radiance is compared to MODIS Level 1B product derived at sensor radiance. The sea surface reflectance can be estimated by minimizing the difference between TOA radiance and the at sensor radiance by changing the sea surface reflectance. The proposed process flow is shown in Figure 1.

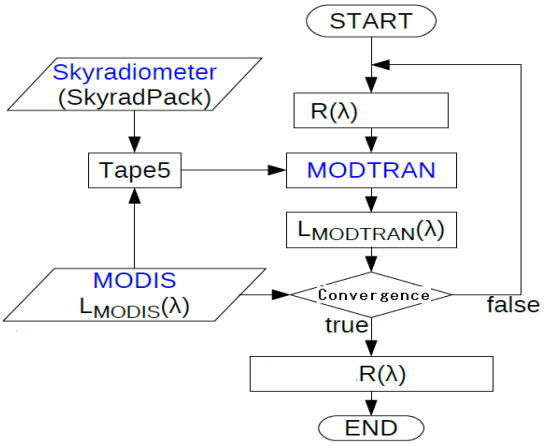


Fig. 1. Process flow of the proposed surface reflectance estimation method

[3] https://books.google.co.jp/books?id=Sy_4jIcRmvUC&pg=PA36&dq=remote+sensing+reflectance+SeaDAS&hl=ja&sa=X&ved=0CBwQ6AEwAGoVChMIip6IntLyxwIVoiimCh1LTAC0#v=onepage&q=remote%20sensing%20reflectance%20SeaDAS&f=false

[4] SkyradPack is available from the University of Tokyo, Nakajima, et al., 1996

[5] http://modtran5.com/

R denotes the remote sensing reflectance while $L_{MODTRAN}$ denotes MODTRAN derived radiance. L_{MODIS} of SeaDAS[6] defined standard product of sea surface reflectance derived from MODIS data is used

The bi-section process is converged within 10 times of iterations because there is only one unknown parameter. The accuracy of this iterative process is around 0.0009765.

B. The Intensive Study Areas

Figure 2 shows the intensive study areas in the Ariake Sea area, Kyushu, Japan.

Fig. 2. Intensive study areas

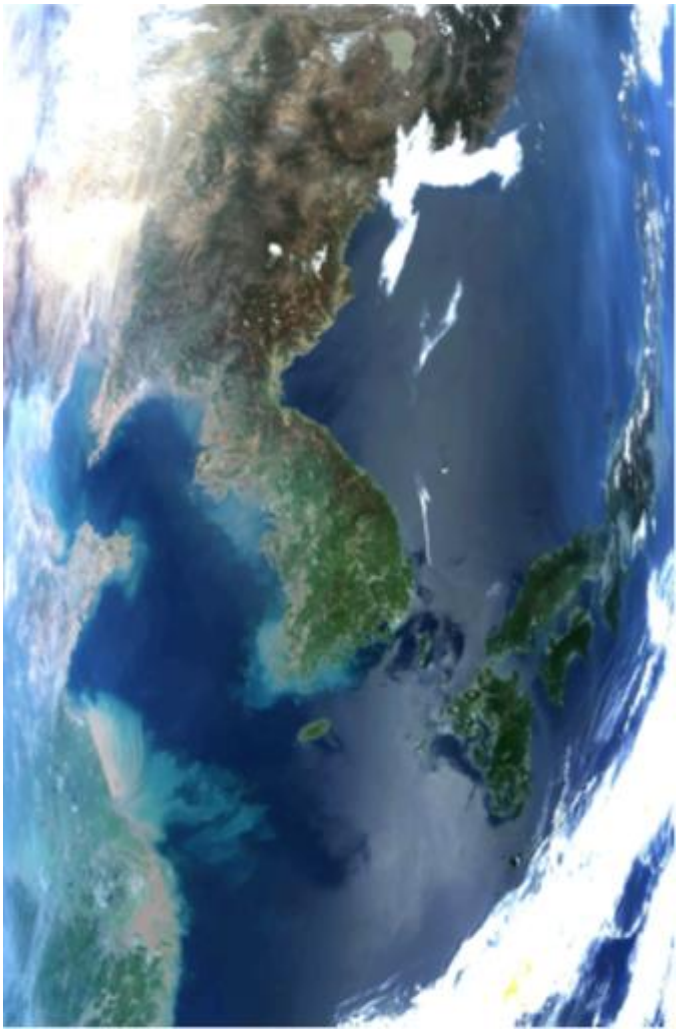
III. EXPERIMENT

A. The Data Used

Terra/MODIS Level 1B of Band 8 to 16 product of Ariake Sea (Latitude: 32.82-33.25 N, Longitude: 130.05-130.65 E), Japan which is acquired at 02:20 (GMT) on May 1 2003 is used. The number of pixel data of Ariake Sea is 638 pixels (Ground resolution of MODIS is 1 km). MODIS Level 1B imagery data is shown in Figure 3.

MODIS on Terra L1B, 2003/5/1 02:20
Latitude 32.82 - 33.25 degree, Lon 130.05 - 130.65 degree

[6] http://seadas.gsfc.nasa.gov/



(a)Portion of MODIS image

(b)MODIS image of the intensive study area of Ariake Sea

Fig. 3. MODIS image of the intensive study area of Ariake Sea acquired on May 1 2003

The locations of MODIS pixels of the intensive study area of Ariake Sea are shown in Figure 4.

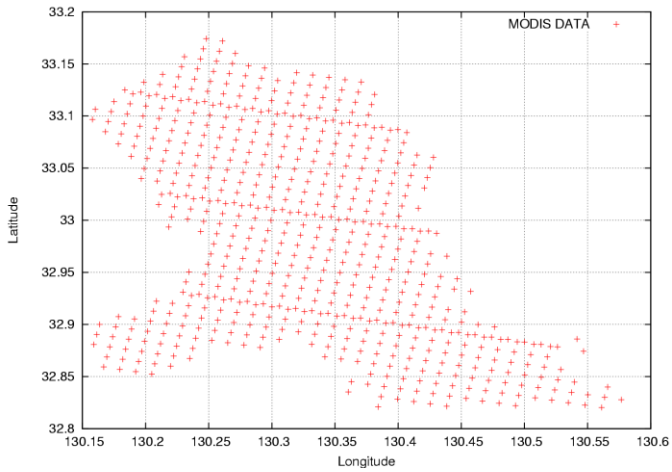


Fig. 4. Location of the MODIS pixels of the intensive study area

B. *The Experimental Results*

MODIS band number, center wavelength, Root Mean Square Difference: RMSD between MODIS standard product of surface reflectance and SeaDAS defined standard remote sensing reflectance, the estimated remote sensing reflectance by the proposed method with the default input parameters of the used MODTRAN (Mid-Latitude Summer), and the estimated remote sensing reflectance by the proposed method with the input parameters including phase function of aerosols are shown in Table 1.

TABLE I. ROOT MEAN SQUARE DIFFERENCE: RMSD COMPARISONS

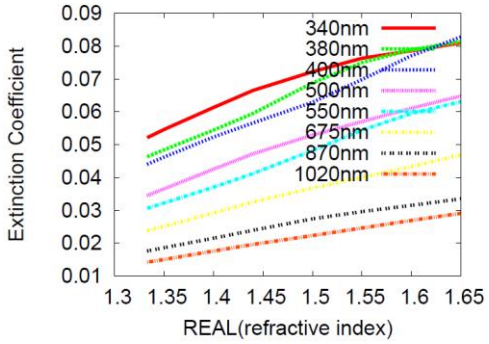
Band	Center Wavelength(nm)	RMSD(1/str)		
		SeaDAS	Default	Proposed
8	412	0.00283	0.00257	0.00156
9	443	0.00361	0.0019	0.00144
10	488	0.00552	0.00341	0.0022
11	531	0.00692	0.00444	0.00336
12	551	0.00677	0.00396	0.003
13	667	0.00221	0.0016	0.00149
14	678	0.00216	0.00154	0.00146
15	748	0.000495	0.000651	0.000739
16	869	0.000248	0.00082	0.000831

In the Table 1, “Default” denotes the proposed method with the default input parameters of the atmosphere without using skyradiometer data while “Proposed” denotes the proposed method with using skyradiometer data. From the table, it may say that the remote sensing reflectance by the proposed method is much closer than the others to the standard product of surface reflectance, L1B product derived remote sensing reflectance in particular for shorter wavelength rages from 412 to 678 nm. Meanwhile, SeaDAS defined remote sensing reflectance is much closer than the others for the longer wavelength ranges from 748 to 869 nm (Near infrared wavelength region). Therefore, it may say that it would be better to use the measured skyradiometer data for improvement of estimation accuracy of surface reflectance. Moreover, the TOA radiance (at sensor radiance) can be estimated simultaneously for vicarious calibration, in particular.

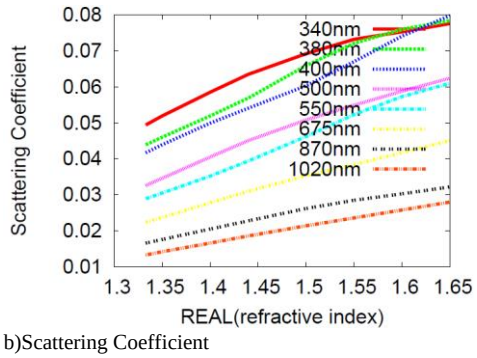
C. Sensitivity Analysis

The relations between aerosol refractive index (Real and Imaginary parts) and extinction coefficient, scattering coefficient, absorption coefficient, and asymmetry parameter are investigated at the wavelengths, 340, 380, 400, 500, 550, 675, 870, and 1020nm (relatively transparent wavelength). Figure 5 shows the relations for the real part of the refractive index of aerosol and extinction, scattering, absorption coefficients and asymmetry parameter while Figure 6 shows those for the imaginary part of the refractive index and extinction, scattering, absorption coefficients and asymmetry parameter.

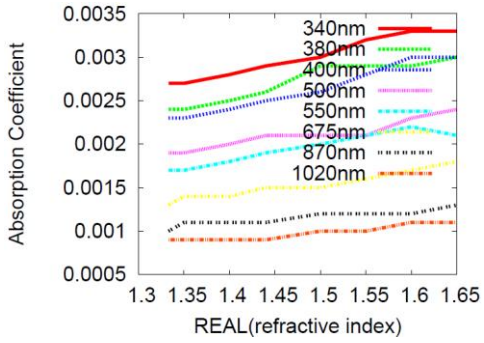
Extinction coefficient consists scattering and absorption coefficients of aerosol particles. On the other hand, asymmetry parameter is an asymmetric characteristic of aerosol scattering phase function. Rayleigh scattering phase function is symmetry while Mie scattering phase function is asymmetry (Forward scattering is dominant). Optical property of aerosol particles can be expressed with these coefficients and asymmetry parameter. Influencing components of aerosol particles to the optical property are refractive index and size distribution. Refractive index consists of real and imaginary parts, complex function. Real part represents refractive component of aerosol particles while imaginary part expresses absorptive component. There are some approximated size distribution functions of aerosol particles. Log-Normal distribution, Power Law distribution as well as Junge distribution functions are representatives. Therefore, the relations among these parameters are examined in these figures,



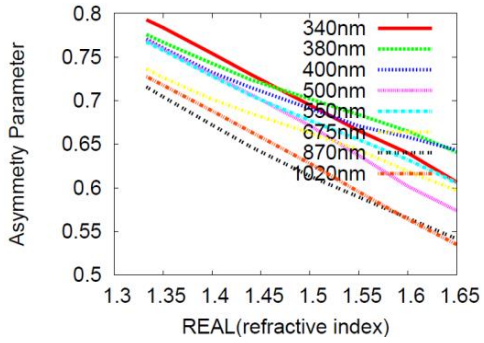
(a)Extinction Coefficient



b)Scattering Coefficient

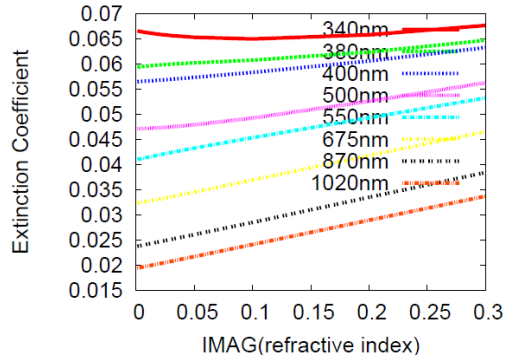


(c)Absorption Coefficient

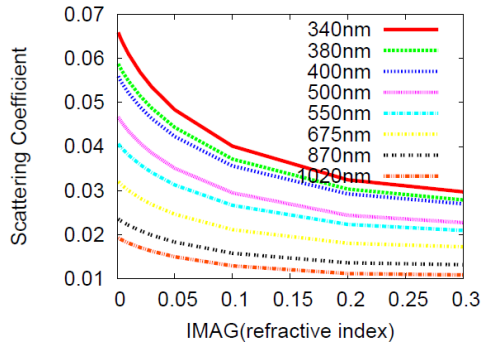


(d)Asymmetry Parameter

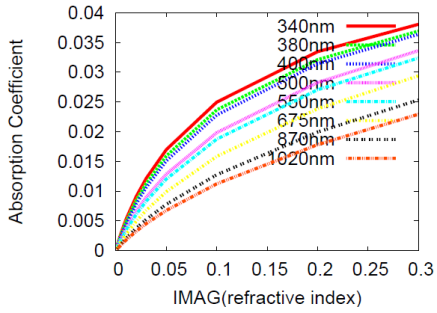
Fig. 5. Relations between real part of refractive index and extinction, scattering, absorption coefficients and asymmetry parameter



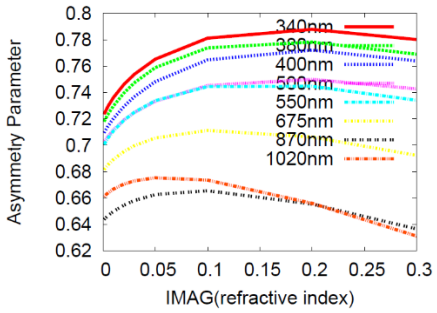
(a)Extinction Coefficient



(b)Scattering Coefficient



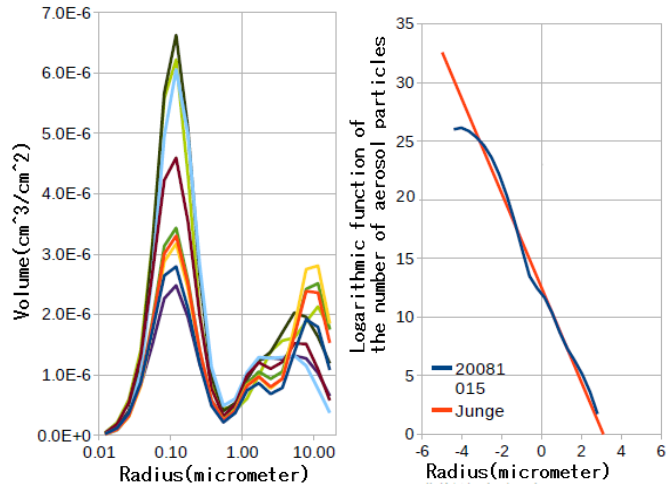
(c)Absorption Coefficient



(d)Asymmetry Parameter

Fig. 6. Relations between imaginary part of the refractive index and Extinction, Scattering, Absorption coefficients, and asymmetry parameter

Figure 7 (a) shows a typical size distribution function of volume spectrum while Figure 7 (b) shows a typical size distribution of number spectrum (logarithmic function of aerosol particle number). In the figures, dark blue size distributions are measured at Saga University on October 15 2008. Red colored linear function shows Junge distribution with Junge parameter v in the equations (1) and (2). As shown in these figures, in usual, size distribution can be expressed with bi-modal function of Log-Normal function and is based on Power Law expression.



(a)Volume spectra (b)Number spectrum

Fig. 7. Typical aerosol size distributions

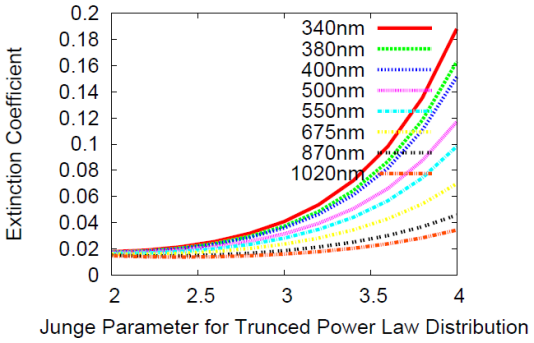
On the other hand, Figure 8 and 9 shows the relations between Junge parameter and extinction, scattering, absorption coefficients and asymmetry parameter as well as the coefficient "C" of the truncated Power Law Distribution function of aerosol size distribution (equations (1) and (2)) and extinction, scattering, absorption coefficients and asymmetry parameter, respectively. There are two appropriate aerosol distribution functions,

Power Law and Log-Normal Distributions. Meanwhile, there are four major atmospheric components, extinction, scattering, absorption coefficients and asymmetry parameter. Power Law Distribution function is as follows,

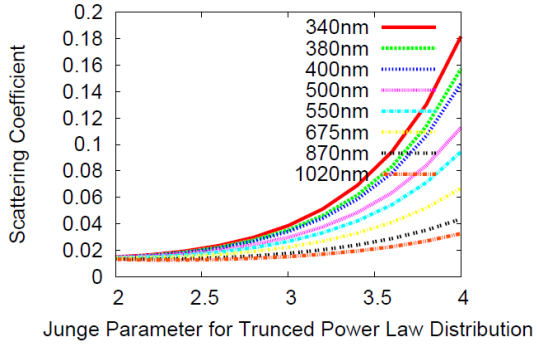
$$n(r) = C10^{v+1} \quad (r \leq 0.1\mu m) \qquad (1)$$

$$n(r) = Cr^{-(v+1)} \quad (r > 0.1\mu m) \qquad (2)$$

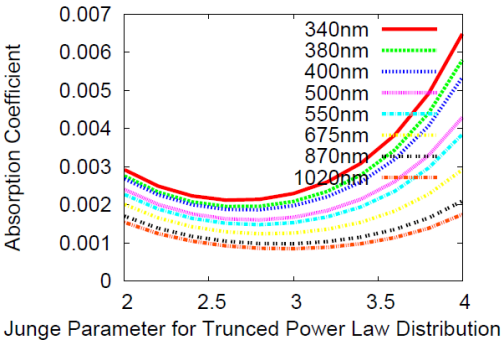
where n, r, C denotes the number of aerosol particles, radius of aerosol particles, and coefficient.



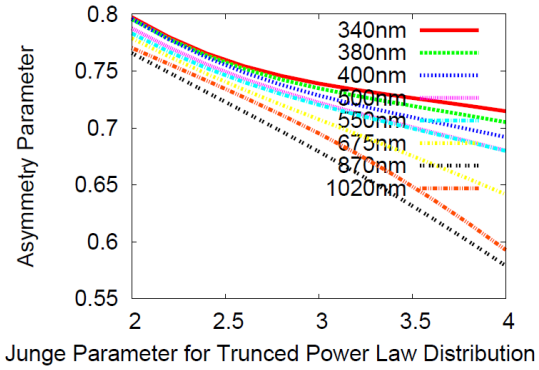
(a)Extinction Coefficient



(b)Scattering Coefficient

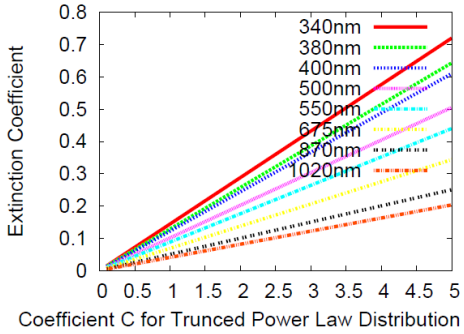


(c)Absorption Coefficient

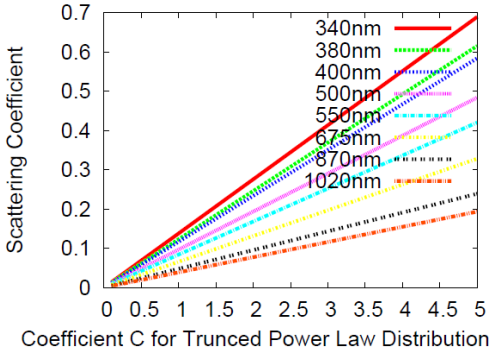


(d)Asymmetry Parameter

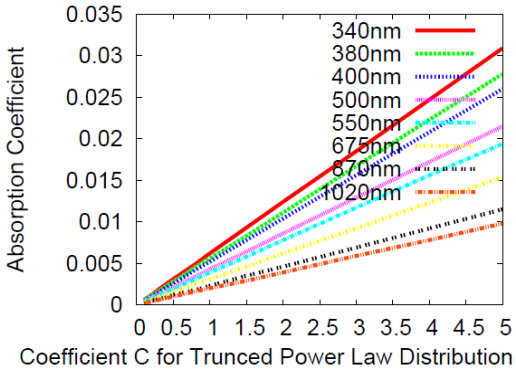
Fig. 8. Relation between Junge parameter for the truncated Power Law Distribution and extinction, scattering, absorption coefficients and asymmetry parameter



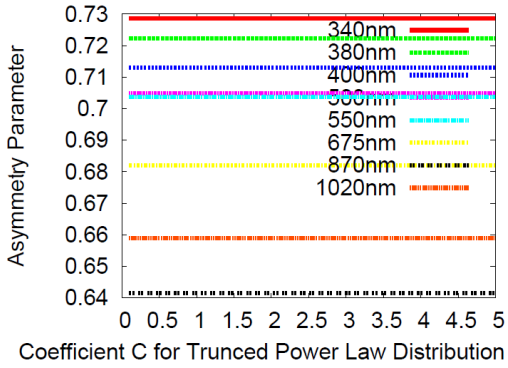
(a)Extinction Coefficient



(b)Scattering Coefficient



(c)Absorption Coefficient



(d)Asymmetry Parameter

Fig. 9. Relation between the coefficient C for truncated Power Law Distribution and extinction, scattering, absorption coefficients and asymmetry parameter

If the Log-Normal Distribution is assumed for aerosol size distribution, then the results from the sensitivity analysis are shown in Figure 10. Log-Normal Distribution function is as follows,

$$\log \sigma_g = (\Sigma n_i (\log D_i - \log D_g)^2/(N-1))^{-1/2} \quad (3)$$

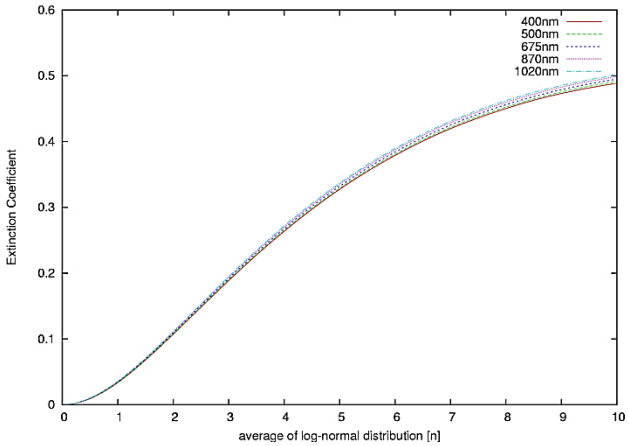
Where,

σ_g = geometric standard deviation (GSD)
D_i = midpoint particle diameter of the *i*th bin
n_i = number of particles in group *i* having a midpoint size *Di*
$N = \sigma_{ni,}$ the total

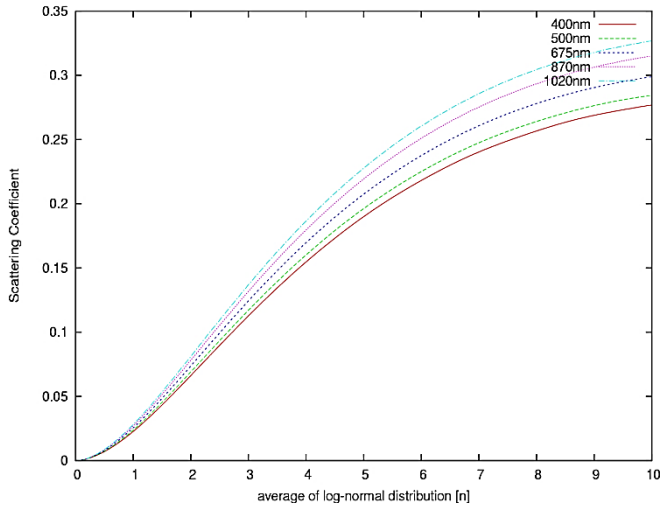
The parameter for the Log-Normal Distribution is as follows,

n=1.0 [cm^{-3}]
σ_g=0.4[micrometer]

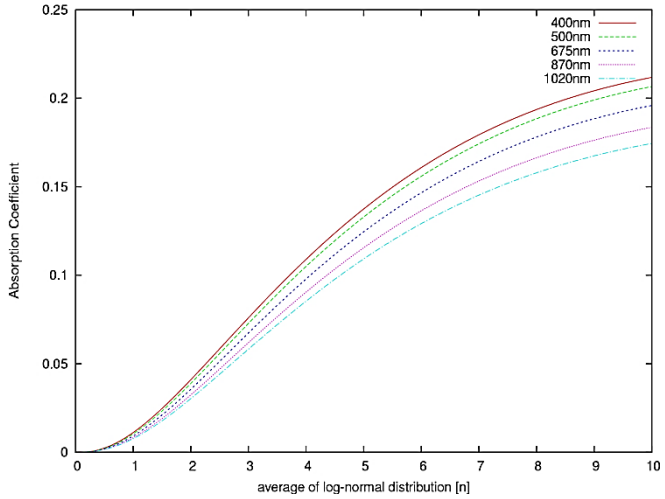
There is a parameter for the Log-Normal Distribution, averaged distribution of *n*. The sensitivity of extinction, scattering, and absorption coefficients as well as asymmetry parameter are varied by the averaged distribution as shown in Figure 10. It is necessary to care about these sensitivity as well as selection of aerosol size distribution function for the convergence process in the proposed process flow which is shown in Figure 1.



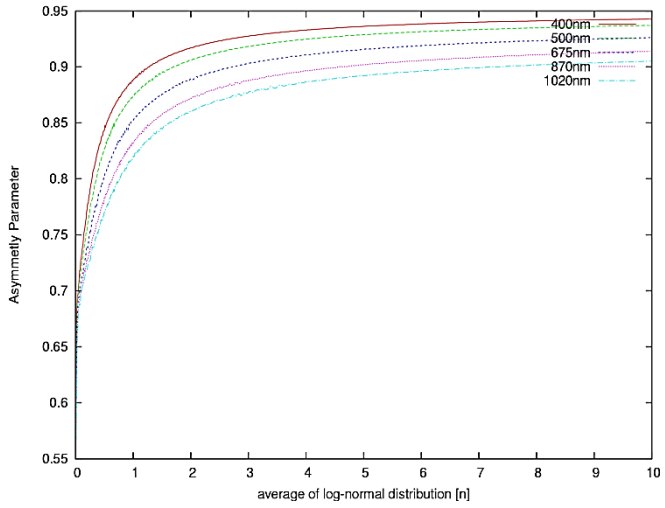
(a)Extinction Coefficient



(b)Scattering Coefficient



(c)Absorption Coefficient



(d)Asymmetry Parameter

Fig. 10. Results from the sensitivity analysis assuming Log-Normal Distribution for aerosol size distribution

IV. Conclusion

Through experiments with the standard surface reflectance product of MODIS and the estimated remote sensing reflectance based on SeaDAS processing software, and the proposed bi-section based convergence process of estimation method with skyradiometer data derived aerosol refractive index and size distribution, it is found that the proposed method with skyradiometer data is superior to the SeaDAS derived remote sensing reflectance.

Further investigations are required for selection of appropriate aerosol size distribution function. The experiment is conducted with the assumed Junge ditribution with the parametrization of Junge parameter. It, however would better to take the other aerosol size distribution functions, Log-Normal, and Power Law distributions from the results of the sensitivity analysis.

References

[1] Ramachandran, Justice, Abrams(Edt.),Kohei Arai et al., Land Remote Sensing and Global Environmental Changes, Part-II, Sec.5: ASTER VNIR and SWIR Radiometric Calibration and Atmospheric Correction, 83-116, Springer 2010.

[2] Kohei Arai, Atmospheric correction and vicarious calibration of ADEOS/AVNIR and OCTS, Advances in Space Research, Vol.25, No.5, pp.1051-1054, 2000.

[3] Kohei Arai, Atmospheric correction and residual error in vicarious calibration of AVNIR and OCTS both onboard ADEOS, Advances in Space Research, Vol.25, No.5, pp.1055-1058, 2000.

[4] K.Arai, Atmospheric correction and vicarious calibration of ADEOS/AVNIR and OCTS, Advances in Space Research, Vol.25, No.5, pp.1051-1054, (2000).

[5] K.Arai, Atmospheric correction and residual errors in cross calibrationof AVNIR and OCTS both onboard ADEOS, Advances in Space Research, Vol.25, No.5, pp.1055-1058, (2000).

[6] Chrysoulakis,Abrams, Feidas and Kohei Arai, Comparison of Atmospheric correction methods using ASTER data for the area of Crete, Greece, International Journal of Remote Sensing, 31,24,6347-6385,2010.

[7] K.Arai, Monte Carlo simulation of polarized atmospheric irradiance for determination of refractive index of aerosols, International Journal of Research and Review on Computer Science, 3, 4, 1744-1748, 2012.

[8] O.Uchino, T.Sakai, T.Nagai, I.Morino, K.Arai, H.Okumura, S.Takubo, T.Kawasaki, Y.mano, T.Matsunaga, T.Yokota, On recent stratspheric aerosols observed by Lidar over Japan, Journal of Atmospheric Chemistry and Physics, 12, 11975-11984, 2012(doi:10.5194/acp-12, 11975-2012).

[9] K.Arai, Monte Carlo ray tracing based sensitivity analysis of the atmospheric and oceanic parameters on the top of the atmosphere radiance, International Journal of Advanced Computer Science and Applications, 3, 12, 7-13, 2012.

[10] K.Arai Error analysis on estimation method for air temperature, atmospheric pressure, and relative humidity using absorption due to CO2, O2, and H2O which situated at around near infrared wavelength regions, International Journal of Advanced Computer Science and Applications, 3, 12, 192-196, 2012.

[11] Kohei Arai, Method for estimation of aerosol parameters based on ground based atmospheric polarization irradiance measurements, International Journal of Advanced Computer Science and Applications, 4, 2, 226-233, 2013

[12] Kohei Arai, Sensitivity analysis and validation of refractive index estimation method with ground based atmospheric polarized radiance measurement data, International Journal of Advanced Computer Science and Applications, 4, 3, 1-6, 2013.

[13] O.Uchino, T.Sakai, T.Nagai, I.Morino, T.Maki, M.Deushi, K.Shibata, M.Kajino, T.Kawasaki, T. Akaho, S.Takubo, H.Okumura, Kohei Arai, M.Nazato, T.Matsunaga, T.Yokota, Y.Sasano, DIAL measurement of

lower tropospheric ozone over Saga (33.24N, 130.29E) in Japan and comparison with a chemical climate model, Journal of Atmospheric Measurement Techniques, 7, 171-194, 2014.

[14] Kohei Arai, Comparative study among least square method, steepest descent method, and conjugate gradient method for atmospheric sounder data analysis, International Journal of Advanced Research in Artificial Intelligence, 2, 9, 30-37, 2013.

[15] Kohei Arai, Sensitivity analysis for aerosol refractive index and size distribution estimation methods based on polarized atmospheric irradiance measurements, International Journal of Advanced Research in Artificial Intelligence, 3, 1, 16-23, 2014.

[16] Kohei Arai, Aerosol refractive index retrievals with atmospheric polarization measuring data, Proceedings of the SPIE, 7461-06, 1-9, 2009.

[17] Kohei Arai, Reflectance based vicarious calibration of ASTER/VNIR with aerosol refractive index and size distribution estimation using measured atmospheric polarization irradiance, Proceedings of the SPIE, 7461-08, 1-9, 2009.

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission "A" of ICSU/COSPAR since 2008. He wrote 33 books and published 500 journal papers. He is now Editor-in-Chief of IJACSA and IJISA.

Defending Grey Attacks by Exploiting Wavelet Analysis in Collaborative Filtering Recommender Systems

Zhihai Yang, Zhongmin Cai* and Aghil Esmaeilikelishomi
Ministry of Education Key Lab for Intelligent Networks and Network Security,
Xi'an Jiaotong University, Xi'an, 710049, China

***Abstract*—"Shilling" attacks or "profile injection" attacks have always major challenges in collaborative filtering recommender systems (CFRSs). Many efforts have been devoted to improve collaborative filtering techniques which can eliminate the "shilling" attacks. However, most of them focused on detecting push attack or nuke attack which is rated with the highest score or lowest score on the target items. Few pay attention to grey attack when a target item is rated with a lower or higher scores than the average score, which shows a more hidden rating behavior than push or nuke attack. In this paper, we present a novel detection method to make recommender systems resistant to such attacks. To characterize grey ratings, we exploit rating deviation of item to discriminate between grey attack profiles and genuine profiles. In addition, we also employ novelty and popularity of item to construct rating series. Since it is difficult to discriminate between the rating series of attacker and genuine users, we incorporate into discrete wavelet transform (DWT) to amplify these differences based on the rating series of rating deviation, novelty and popularity, respectively. Finally, we respectively extract features from rating series of rating deviation-based, novelty-based and popularity-based by using amplitude domain analysis method and combine all clustered results as our detection results. We conduct a list of experiments on the Book-Crossing dataset in diverse attack models. Experimental results were included to validate the effectiveness of our approach in comparison with benchmarked methods.**

Keywords—recommender system; grey attack; discrete wavelet transform; shilling attack

I. INTRODUCTION

Collaborative filtering recommender systems (CFRSs) have become a popular and effective tool for information retrieval especially when users facing information overload. CFRSs also have played an important role in many popular web services such as Netflix, Amazon etc., which are designed to recommend items based on relevant information for the specific user [3], [5], [11], [14], [30], [33]. However, CFRSs are particularly vulnerable to "shilling" attacks or "profile injection" attacks in which an attacker signs up as a number of "puppet" users and rates fake scores in an attempt to promote or demote the recommendations of specific items by using knowledge of the recommender algorithms [2], [20], [21], [25], [26]. In such attacks, the attackers deliberately insert attack profiles into genuine profiles to change the prediction results which would reduce the trustworthiness of recommendation. The attack profiles indicate the attacker's intention that he wishes a particular item can be rated highest score (called push attack) or lowest score (called nuke attack) [4], [6], [7], [9], [10], [12], [16], [18], [19]. In addition, to avoid being detected easily by traditional detection techniques, the attackers may rate a higher score or lower score on the target items, which generates relatively hidden attack intents in comparison with push attacks or nuke attacks [24], we also call them grey attacks. Of course, they belong to the "shilling" attacks. Therefore, constructing an effective system to defend the attackers and remove them from the CFRSs is crucial.

Although existing work in this area have focused on detecting and preventing the "shilling" attacks or "profile injection" attacks, it has not reached a fully acceptable level of detection performance. We can briefly summarize that it is difficult to improve detection performance for detecting such attacks when filler size or attack size is small. Moreover, few pay attention to the grey attack detection. As an attacker demotes (nuke attack) the target items by rating lowest score or promotes (push attack) the target items by rating highest score, he also can demote or promote the target items by rating lower or higher scores. In fact, the rating behavior of an attacker is very similar to the behavior of a genuine user if the rating of target item is close to the actual rating. For the nuke attack, an attacker is simply shifting the rating given to the target item from the minimum rating to a rating one step higher, for the push attack, and vice versa [24]. Any profile that includes these ratings is likely to be less suspect. Although a minor change, this has a key effect. Thus, a challenging detection method should not only perform well when attack size or filler size is small, but also effectively defend the grey attack profiles.

In this paper, we propose an unsupervised attack detection method to make recommender systems resistant to such attacks, which combines discrete wavelet transform (DWT) and EM-based (Expectation-maximization based) clustering method. Since the attackers mimic some rating details of genuine users in shilling attacks, the rating behavior between attackers and genuine users will become more similar, especially for the grey attacks. Although existing features extracted from user profiles can characterize the shilling attacks to some extent, it's difficult to fully discriminate between attack profiles and genuine profiles. Moreover, the above challenges are also significant in grey attacks. Our basic assumption is that we can use DWT to amplify the differences between attack profiles and genuine profiles. In addition, to

characterize the features of grey ratings, we use rating deviation of item to address this crucial problem. To construct input series for DWT, we create a list of transformed rating series to address this problem, which exploits the novelty, popularity and rating deviation of item for each user profiles, respectively. Moreover, we employ the empirical model decomposition (EMD) method to extract intrinsic mode functions (IMFs) from the rating series [17]. These can be seen that there are some but not obvious difference between the attack profiles and genuine profiles (as shown in Figures 4-6). To amplify the difference, we further use DWT to transform these series. In essence, a rating series is a non-stationary random series. Therefore, it is very suitable to be processed by DWT which performs well for non-stationary data [17]. After DWT, the differences between attack profiles and genuine profiles become more obvious (as shown in Figures 7-9). Based on the output series of DWT, we extract a list of effective features by using amplitude domain analysis method. And then exploiting EM clustering method to discriminate jointly attackers and genuine users based on the extracted features. In addition, the effectiveness of our proposed approach is validated and benchmark methods are briefly discussed. Experimental results show that our approach performs well for detecting the grey attacks in comparison with the benchmarked methods.

The remaining parts of this paper are organized as follows: Section 2 reviews some related work. Section 3 describes the attack model and introduces the theory of discrete wavelet transform. Our proposed detection method is introduced in Section 4. Experimental results and analysis are presented and discussed in Section 5. Finally, we conclude the paper with a brief summary and directions for future work.

II. Related work

Although existing detection techniques have focused on detecting and preventing the "shilling" attacks or "profile injection" attacks, it has not reached a fully acceptable level of detection performance. To name only a few, Burke et al. [3] proposed and studied several attributes derived from user profiles for their utility in attack detection. They employed kNN as their classification approach. However, it was unsuccessful when detecting attacks with small filler size and also suffered from low classifier precision. Then, Williams et al. [15], [24], [28] used several trained classifiers to detect shilling attacks based on extracted features of user profiles. Although, [24] used the higher or lower ratings instead of the maximum or minimum ratings to the target item, discussion of detecting such attacks was limited. Moreover, the detection performance was limited when filler size is small. Mobasher et al. [29] employed signatures of attack profiles and were moderately accurate. But, the method suffered from low accuracy in detecting shilling attack. They just focused on individual users and mostly ignored the combined effect of such attackers. In addition, the detection performance was limited when the attack profiles are obfuscated. Zhang et al. [31] proposed an ensemble approach to improve the precision of detection by using meta-learning technique. Their proposed method performs better detection performance than the bench marked methods. He et al. [32] employed rough set theory to detect shilling attacks though taking features of user profiles as the condition attributes of the decision table. However, their method also suffered from low precision. F. Zhang et al. [17] proposed an online method to detect profile injection attacks based on HHT and SMV. Zhou et al. [1] proposed a detection technique for identifying group attack profiles, called DeR-TIA, which combines an improved metric based on Degree of Similarity with Top Neighbors (DegSim) and Rating Deviation from Mean Agreement (RDMA). Zhang et al. [19] proposed a spectral clustering method to make recommender systems resistant to the shilling attacks in the case that the attack profiles are highly correlated with each other. Their experimental results reported good performance in random, average and bandwagon attacks. However, it also performed poor precision and recall in AOP attack when attack size is small.

III. Preliminaries

In this section, we firstly describe the attack profiles and attack models. Then, we introduce the theory of discrete wavelet transform to facilitate discussions later.

A. Attack profiles and attack models

In the literature, "shilling" attacks are classified into two ways: nuke attack and push attack [3]. In order to nuke or push a target item, the attacker should be clearly known the form of an attack profile. The general form of an attack profile is shown in Table 1. The details of the four sets of items are described as follows:

I_S: The set of selected items with specified rating by the function $\sigma(i_k^S)$ [13];

I_F: A set of filler items, received items with randomly chosen by the function $\rho(i_l^F)$;

I_N: A set of items with no ratings;

I_T: A set of target items with singleton or multiple items, called single-target attack or multiple-targets attack. The rating is $\gamma(i_j^T)$, generally rated the maximum or minimum value in the entire profiles.

In this paper, we utilize 8 attack models to generate attack profiles. The involved attack profiles and corresponding explanations are listed in Table 2. The details of these attack models in our experiments are described as follows:

1) AOP attack: A simple and effective strategy to obfuscate the Average attack is to choose filler items with equal probability from the top x% of most popular items rather than from the entire collection of items [22].

2) Random attack: $I_S = \emptyset$ *and* $\rho(i) \sim N(\bar{r}, \bar{\sigma}^2)$ *[13];*

3) Average attack: $I_S = \emptyset$ *and* $\rho(i) \sim N(\bar{r}_i, \bar{\sigma}_i^2)$ *[13];*

4) Bandwagon (average): I_S *contains a set of popular items. Then, we use these items as* I_S *,* $\sigma(i) = r_{max}$ *or* r_{min} *or* r_{grey} *(push or nuke or grey) and* $\rho(i) \sim N(\bar{r}_i, \bar{\sigma}_i^2)$ *[13];*

5) Bandwagon (random): I_S *contains a set of popular items,*

TABLE I. GENERAL FORM OF ATTACK PROFILES

I_T			I_S			I_F			I_N		
i_1^T	…	i_j^T	i_1^S	…	i_k^S	i_1^F	…	i_l^F	i_1^N	…	i_v^N
$\gamma(i_1^T)$	…	$\gamma(i_j^T)$	$\sigma(i_1^S)$	…	$\sigma(i_k^S)$	$\rho(i_1^F)$	…	$\rho(i_l^F)$	null	…	null

TABLE II. ATTACK MODELS

Attack Model	I_S		I_F		I_N	I_T (push or nuke or grey)
	Items	*Rating*	*Items*	*Rating*		
AOP	null		x-% popular items, ratings set with normal dist around item mean.		null	r_{max} or r_{min} or r_{grey}
Random	null		randomly chosen	system mean	null	r_{max} or r_{min} or r_{grey}
Average	null		randomly chosen	item mean	null	r_{max} or r_{min} or r_{grey}
Bandwago (average)	popular items	r_{max} or r_{min}	randomly chosen	item mean	null	r_{max} or r_{min} or r_{grey}
Bandwagon (random)	popular items	r_{max} or r_{min}	randomly chosen	system mean	null	r_{max} or r_{min} or r_{grey}
Segment	segmented items	r_{max} or r_{min}	randomly chosen	r_{min} or r_{max}	null	r_{max} or r_{min} or r_{grey}
Reverse Bandwagon	unpopular items	r_{min} or r_{max}	randomly chosen	system mean	null	r_{max} or r_{min} or r_{grey}
Love/Hate	null	null	randomly chosen	r_{min} or r_{max}	null	r_{max} or r_{min} or r_{grey}

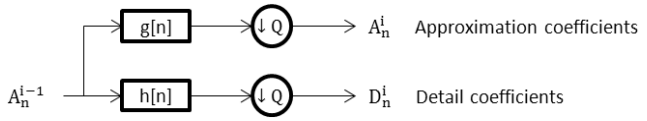


Fig. 1. Block diagram of filter analysis

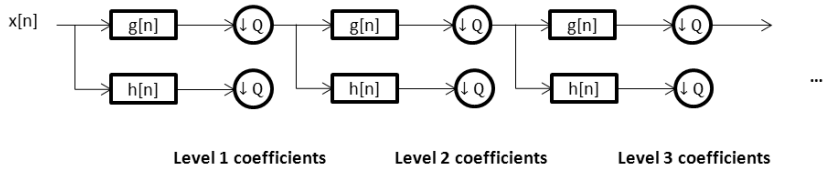


Fig. 2. K (k greater than or equal to 1) levels of filter bank

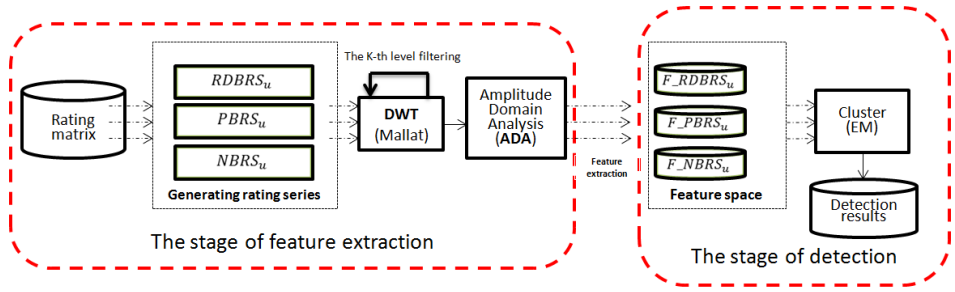


Fig. 3. The framework of our proposed method which consists of two stages: the stage of feature extraction and the stage of detection

$\sigma(i) = r_{max}$ or r_{min} or r_{grey} and $\rho(i) \sim N(\bar{r}, \bar{\sigma}^2)$ (nuke or grey) [13];

6) Segment attack: I_S contains a set of segmented items, $\sigma(i) = r_{max}$ or r_{min} or r_{grey} and $\rho(i) = r_{min}$ or r_{max} or r_{grey} (push or nuke or grey) [8];

7) Reverse Bandwagon attack: I_S contains a set of unpopular items, $\sigma(i) = r_{min}$ or r_{max} or r_{grey} (push or nuke or grey) and $\rho(i) \sim N(\bar{r}, \bar{\sigma}^2)$ [9];

8) Love/Hate attack: $I_S = \emptyset$ and $\rho(i) = r_{max}$ or r_{grey} (nuke or grey) [9].

B. Discrete wavelet transform

Discrete wavelet transform (DWT) has been recognized as a natural wavelet transform for discrete time signals. Both time and scale parameters are discrete. For a discrete-time sequence x[n], nϵZ, DWT is defined by discrete-time multi-resolution decomposition which could be computed by Mallat pyramidal decomposition algorithm (as shown in Equations (1)-(3)) [23]. However, since half the frequencies of the signal have now been removed, half the samples can be discarded according to Nyquist's rule. The filter outputs are then sub-

sampled by 2 (Mallat's and the common notation is the opposite, g- high pass and h- low pass):

$$A_n^0 = x[n],\ n\epsilon N \quad (1)$$

$$A_n^i = \sum_{k\epsilon Z} g(k-2n)A_k^{i-1},\ i = 1,2,\dots,L \quad (2)$$

$$D_n^i = \sum_{k\epsilon Z} h(k-2n)A_k^{i-1},\ i = 1,2,\dots, \quad (3)$$

where h and g are impulse responses of high-pass filter H and low-pass filter G, respectively. $\{A_n^i\}$ and $\{D_n^i\}$ are scale sequence and wavelet sequence of 2^{-i} scale. L is the maximum possible scale of the discrete signal x[n]. The signal is also decomposed simultaneously using a high-pass filter. The outputs give the detail coefficients (from the high-pass filter) and approximation coefficients (from the low-pass) as shown in Figure 1. It is important that the two filters are related to each other and they are known as a quartered mirror filter.

DWT of a signal is calculated by passing it through a series of filters. The decomposition is repeated to further increase the frequency resolution and the approximation coefficients decomposed with high and low pass filters and then down-sampled (see Figure 2). This is represented as a binary tree with nodes representing a sub-space with different time-frequency localization. And the tree is known as a filter bank.

IV. Our Proposed Approach

In this section, we firstly introduce the framework of our proposed approach. And then we give several definitions of rating series used in this paper. Finally, we briefly describe our detection method.

A. The framework

As shown in Figure 3, our proposed algorithm consists of two stages: the stage of feature extraction and the stage of detection. At the stage of feature extraction, the feature is extracted one by one from user profiles by using the proposed feature extraction method (see subsection 4.2). Inspired from previous studies (Zhang et al. [17]), we incorporate into two concepts: Empirical Mode Decomposition (EMD) and Intrinsic Mode function (IMF). EMD is an adaptive and highly efficient decomposition method and is also a necessary step to reduce any given data into a collection of intrinsic mode functions (IMF) where the DWT analysis can be applied. As we all know, DWT is a method for analyzing non-stationary data, since the rating series are non-stationary data. The IMF is defined as a function that satisfies the following requirements: (a) In the whole data set, the number of extreme and zero-crossings must either be equal or differ at most by one; (b) At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

With this method, rating series can be decomposed into a finite signal and regard the signal as the input of discrete wavelet transform [17], [27]. In our proposed approach, we decompose respectively each user profiles into novelty-based, popularity-based and rating deviation-based rating series as the input signals. And then, the input signals are passed through the series of filters (including low-pass and high-pass filter, as shown in Figure 3.) to generate corresponding output signals. In the process of DWT, we perform one level transformation to get the output signals. Then, by using amplitude domain analysis method to extract features from the output signal. At the stage of detection, based on the extracted features, we respectively use EM method to cluster two groups. Finally, combing the three parts of clustering results to return our detection result.

B. Feature extraction

Previous studies [17] have disclosed that using the novelty and popularity of items to construct rating series for user profiles implies useful information. Inspired from this research, we investigate using rating deviation of items to construct rating series in order to extract features from grey attack profiles. Novelty [1] in recommendation is focusing on recommending the log-tail items (i.e., less popular items) which is generally considered to be particularly valuable to users. Popularity of items usually reflects the genuine users' tastes or preferences in collaborative recommender system. By sorting the items according to their novelty, popularity and rating deviation, we can create respectively the rating deviation-based, novelty-based and popularity-based rating series for the user profiles. Firstly, two definitions of the rating deviation are described in the following:

Definition 1 (Rating Deviation of Items, RDoI).

The $RDoI_i$ (rating deviation of item i) is defined as follows:

$$RDoI_i = \begin{cases} |r_{ui} - \bar{r}_i|, & r_{ui} \neq \perp, u \in R_g \\ 0, & r_{ui} = \perp \end{cases}, \quad (4)$$

where r_{ui} denotes the rating of user u on item i. $\bar{r}_i$ is the mean rating of item i in the system. $r_{ui} \neq \perp$ denotes item i is rated by user u, $r_{ui} = \perp$ denotes item i is not rated by user u. R_g denotes the set of genuine users in dataset.

Definition 2 (Rating Deviation-based Rating Series, RDBRS).

Let $RDoI_i$ denotes the rating deviation of item i. Sort all items in set *I* (a set of the entire items in the recommender system.) according to $RDoI_i$ in descending order and let $i = 1,2,\dots,|I|$ denotes the order of items after sorting, where |I| denotes total number of items in the recommender system. The $RDBRS_u(i)$ [2] is defined as follows:

$$RDBRS_u(i) = \begin{cases} 1, & r_{u,i} \neq \perp \text{ and } (i = 1 \text{ or } RDNRS_u(i-1) \neq 1), \\ -1, & r_{u,i} = \perp \text{ and } (i = 1 \text{ or } RDNRS_u(i-1) \neq -1), \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

where zero value is used to meet the requirements of extreme for DWT. $r_{u,i} \neq \perp$ denotes item i is rated by user u. $r_{u,i} = \perp$ denotes item i is not rated by user u.

Novelty of Items, NoI

The NoI_i(novelty of item i) is defined as follows:

$$NoI_i = \frac{1}{|R_g|}\sum_{u\in R_g, r_{u,i}\neq\perp} NoI_{u,i}, \quad (6)$$

[1] The novelty of an item refers to the degree to which it is unusual with respect to the user's normal tastes.

[2] The rating deviation-based rating series of user u.

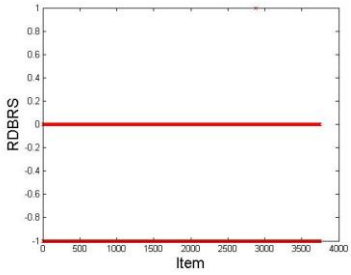
where $NoI_{u,i}$ denotes the novelty of item i for user u [17].

$$NoI_{u,i} = \frac{1}{|N_j|}\sum_{u\in R_g, r_{u,j}\neq\perp}(1 - simi(i,j)) \quad (7)$$

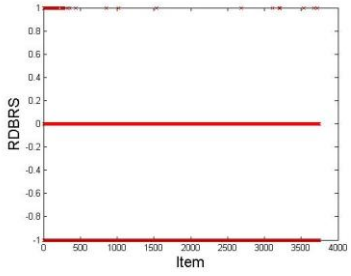
where N_j denotes the number of users who rate on item j. R_g denotes the set of genuine users in dataset. $simi(i,j)$ (Jaccard coefficient) denotes the similarity between item i and item j, which can be calculated as follows:

$$simi(i,j) = \frac{|V_i \cap V_j|}{|V_i \cup V_j|} \quad (8)$$

Where V_i is set of users rated by item i, V_j is the set of users rated by item j. If both V_i and V_j are empty, we define $simi(i,j) = 0$. Clearly, $0 \leq simi(i,j) \leq 1$.

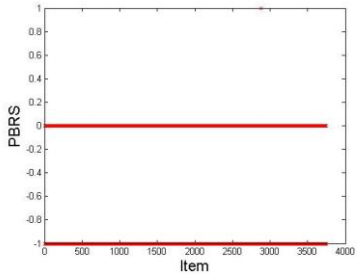


(a) Genuine profile

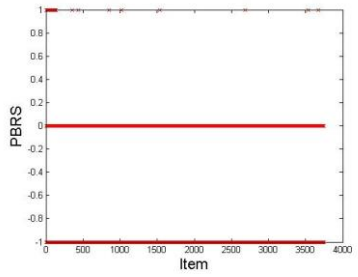


(b) Average attack profile

Fig. 4. Rating Deviation-based rating series. (a) The signal of a genuine profile before DWT; (b) The signal of an average attack profile before DWT

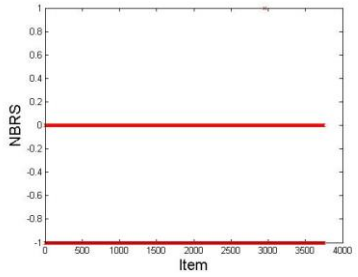


(a) Genuine profile

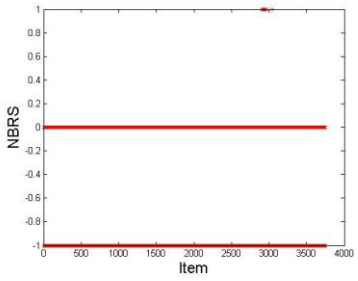


(b) Average attack profile

Fig. 5. Popularity-based rating series. (a) The signal of a genuine profile before DWT; (b) The signal of an average attack profile before DWT



(a) Genuine profile



(b) Average attack profile

Fig. 6. Novelty-based rating series. (a) The signal of a genuine profile before DWT; (b) The signal of an average attack profile before DWT

Novelty-based Rating Series, NBRS

Let NoI_i denotes the novelty of item i. Sort all items in set I according to NoI_i in descending order and let $i = 1,2,\ldots,|I|$ denotes the order of items after sorting. The novelty-based rating series of user u, $NBRS_u(i)$ is defined as follows:

$$NBRS_u(i) = \begin{cases} 1, & r_{u,i} \neq \perp \text{ and } (i = 1 \text{ or } NBRS_u(i-1) \neq 1), \\ -1, & r_{u,i} = \perp \text{ and } (i = 1 \text{ or } NBRS_u(i-1) \neq -1), \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

where zero value is used to meet the requirements of extreme for DWT [17].

Popularity of Items, PoI

The popularity of item i, PoI_i, is defined as the number of ratings given to item i by genuine users in data set [17].

Popularity-based Rating Series, PBRS

Let PoI_i denotes the popularity of item i. Sort all items in set I according to PoI_i in descending order and let $i = 1,2,\ldots,|I|$ denotes the order of items after sorting. The popularity-based rating series of user u, $PBRS_u(i)$, is defined as follows:

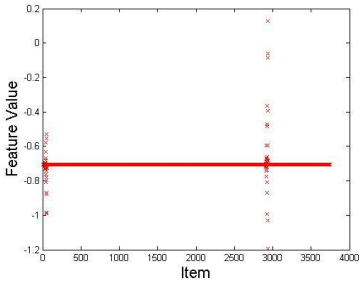
$$PBRS_u(i) = \begin{cases} 1, & r_{u,i} \neq \perp \text{ and } (i = 1 \text{ or } PBRS_u(i-1) \neq 1), \\ -1, & r_{u,i} = \perp \text{ and } (i = 1 \text{ or } PBRS_u(i-1) \neq -1), \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

where zero value is used to meet the requirements of extreme for DWT [17].

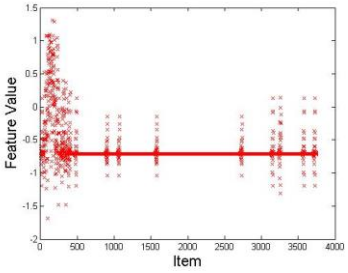
To show the difference between genuine and attack profiles in rating series, we give examples of the novelty-based, popularity-based and rating deviation-based rating series in Figures 4-6. These rating series are constructed by the genuine profiles and the average attack profiles (take average attack for example). The genuine profiles are selected from the Book-Crossing dataset. As shown in Figures 4-6, there are very little difference between the genuine and average attack profiles in rating series. We can observe that the RDBRS for the genuine profile barely changed from starting position to ending position in compared to the RDBRS of the average attack profile decreased gradually for the rating deviation-based rating series. For the popularity-based rating series, the PBRS for the genuine profile barely changed with the item increased while the PBRS of the average attack profile decreased gradually. And for the novelty-based rating series, the NBRS for genuine profile also almost remain unchanged with the item increased, while the NBRS of the average attack profile show characteristics of more concentrated. As mentioned above, it is

difficult to discriminate between genuine profiles and attack profiles regardless of using Rating Deviation-based, Popularity-based and Novelty-based rating series. To amplify the difference between genuine profiles and attack profiles, we use DWT to transform the rating series in order to extract features from output signal by using amplitude domain analysis method.

After K (k greater than or equal to 1) level discrete wavelet transform (as shown in Figure 2), we can get the local properties, which passes a series low-pass filters to obtain an approximation coefficients. As shown in Figures 7-9, we can observe that there is a more significant difference between genuine profiles and average attack profiles on rating series than before using DWT. In Figure 7, the strength of oscillations of genuine profiles show characteristics of more concentrated with the item increased while the strength of

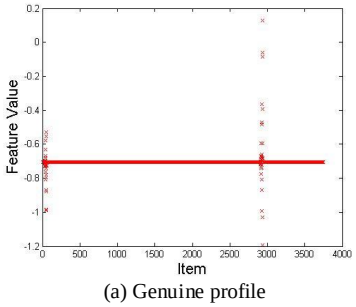


(a) Genuine profile

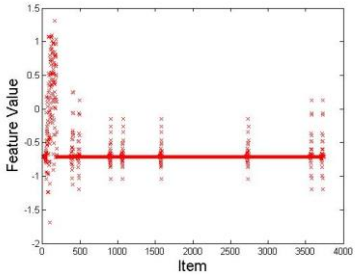


(b) Average attack profile

Fig. 7. The first low-pass output of the rating deviation-based rating series. (a) The signal of a genuine profile after DWT; (b) The signal of an average attack profile after DWT.

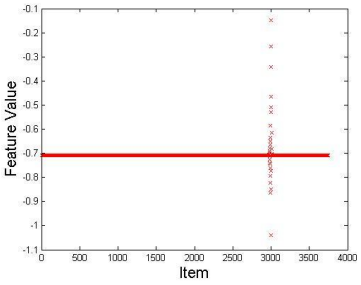


(a) Genuine profile

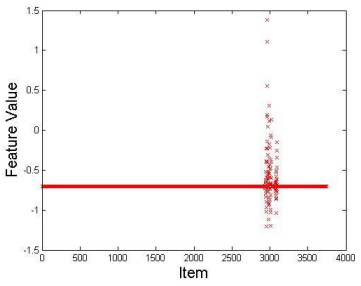


(b) Average attack profile

Fig. 8. The first low-pass output of the popularity-based rating series. (a) The signal of a genuine profile after DWT; (b) The signal of an average attack profile after DWT

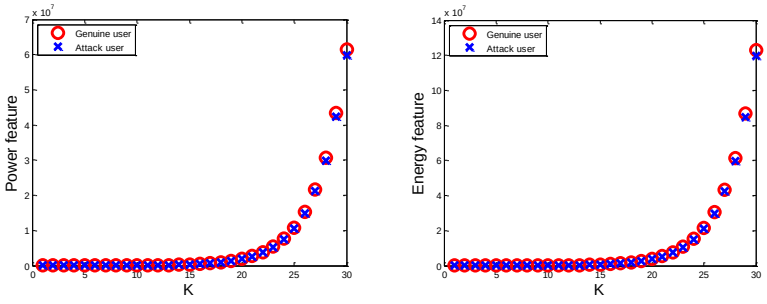


(a) Genuine profiles



(b) Average attack profiles

Fig. 9. The first low-pass output of the novelty-based rating series. (a) The signal of a genuine profile after DWT; (b) The signal of an average attack profile after DWT



(a) (b)

Fig. 10. The power feature and the energy feature in different K levels output of discrete wavelet transforms for a genuine user and an attacker. (a) Power features; (b) Energy features

oscillations of average attack profile decreased gradually from starting position to ending position. For the popularity-based rating series, the same observations are also clear in Figure 8. And for the novelty-based rating series, we can observe that there is a little difference between the genuine profiles and average attack profiles, although they show characteristics of more concentrated similarly as illustrated in Figure 9.

Let F_RDBRS_u, F_PBRS_u and F_NBRS_u denotes the feature vector of user u on the rating deviation-based, novelty-based and popularity-based after DWT, respectively. The proposed feature extraction algorithm is described in algorithm 1. In algorithm 1, from step 1 to step 3 create the rating deviation-based, novelty-based and popularity-based rating series for user u respectively. Step 4 is the process of DWT. Step 5 extract features from approximation parts of rating deviation, popularity and novelty rating series, termed A_RD_k, A_P_k and A_N_k by using amplitude domain analysis method. The last step generates a feature space for the stage of detection.

Algorithm 1: Feature extraction algorithm for user profiles

Input: Rating Matrix;
Output: F_RDBRS_u, F_PBRS_u and F_NBRS_u;
Step 1: Create rating series $\mathrm{RDBRS}_u(i)$ of u by using rating matrix and Equations (4)-(5);
Step 2: Create rating series $\mathrm{NBRS}_u(i)$ of u by using rating matrix and Equations (6)-(9);
Step 3: Create rating series $\mathrm{PBRS}_u(i)$ of u by using rating matrix and Equation (10);
Step 4: Generate approximation parts A and detail parts D by exploiting Mallat (discrete wavelet transform) algorithm on the rating series of $\mathrm{RDBRS}_u(i)$, $\mathrm{PBRS}_u(i)$ and $\mathrm{NBRS}_u(i)$ by using Equations (1)-(3), respectively;
Step 5: Take the K level approximation parts A_RD_k, A_N_k and A_P_k from Step 4's output, respectively. And extract features from the approximation parts by using amplitude domain analysis method on A_RD_k, A_N_k and A_P_k respectively;
Step 6: Generate and return the feature space F_RDBRS_u, F_PBRS_u and F_NBRS_u respectively.

TABLE III. THE FEATURES OF THE SIGNAL AMPLITUDE DOMAIN AND THEIR DESCRIPTION

Features	Equations	Descriptions
Minimum value	$\mathrm{x_{min}} = \min(\mathrm{X})$	The minimum value of the amplitude of the signal.
Maximum value	$\mathrm{x_{max}} = \max(\mathrm{X})$	The maximum value of the amplitude of the signal.
Mean value	$\overline{\mathrm{X}} = \mathrm{mean}(\mathrm{X})$	The average value of the amplitude of the signal.
Peak value	$\mathrm{x_p} = \max(\mathrm{abs}(\mathrm{X}))$	The maximum of the absolute value of the amplitude of the signal.
Root mean square value	$\mathrm{X_{rms}} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{x_i^2}}$	The root mean square value of the amplitude of the signal.
Root mean square amplitude value	$\mathrm{X_r} = \left(\frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \sqrt{\lvert \mathrm{x_i} \rvert}\right)^2$	Represent the energy size of the signal.
Absolute mean	$\lvert\overline{\mathrm{X}}\rvert = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \lvert \mathrm{x_i} \rvert$	Absolute mean value of the amplitude of the signal.
Variance	$\sigma_\mathrm{x}^2 = \mathrm{X_{rms}^2} - \overline{\mathrm{X}}^2$	Represent the degree of dispersion of the signal.
Skewness	$\alpha = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{x_i^3}$	Represent the asymmetry of amplitude probability density function on the vertical axis.
Kurtosis	$\beta = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{x_i^4}$	Represent the steep degree of the signal curve.
Shape factor	$\mathrm{S_f} = \mathrm{X_{rms}}/\lvert\overline{\mathrm{X}}\rvert$	A shape factor refers to a value that is affected by an object's shape but is independent of its dimensions
Crest factor	$\mathrm{C_f} = \mathrm{X_{max}}/\mathrm{X_{rms}}$	Crest factor is a measure of a waveform, showing the ratio of peak values to the average value.
Impulse factor	$\mathrm{I_f} = \mathrm{X_{max}}/\lvert\overline{\mathrm{X}}\rvert$	Non-dimensional parameter in amplitude domain.
Clearance factor	$\mathrm{CL_f} = \mathrm{X_{max}}/\mathrm{X_r}$	Non-dimensional parameter in amplitude domain.
Kurtosis value	$\mathrm{K_v} = \beta/\mathrm{X_{rms}^4}$	Non-dimensional parameter in amplitude domain.

For different types of signal, there are different analysis methods such as time domain analysis, frequency domain analysis and amplitude domain analysis. As shown in Figure 10, we can observe that these are no significant difference between genuine user and attacker with the K (the K level output of DWT) increased, regardless of using the power features or energy features. In this paper, we use amplitude domain analysis to extract features from signals. The details of signal features in amplitude domain are showed in Table 3. We have 15 features to characterize the signal which extracts from the K level (we set K equal to 1 in our work) output of DWT.

C. Detection algorithm

In order to get better detection performance as far as possible, we combine the rating deviation-based, novelty-based and popularity-based methods to distinguish between genuine profiles and attack profiles. And then, we utilize EM (Expectation-maximization) clustering method (Clustering results and EM clustering method were created using Weka [3]) to separate attackers from genuine users as far as possible. Let D denotes the set of detection result. The proposed method for detecting grey attacks is described in algorithm 2. In algorithm 2, from step 1 to 3 perform EM algorithm on feature vector $\mathrm{F\_RDBRS_u}$, $\mathrm{F\_PBRS_u}$ and $\mathrm{F\_NBRS_u}$, respectively. Step 4 obtains the set of attackers decided by using the smaller cluster, since the number of attackers less than the number of genuine users in the recommender system. In step 5, we exploit the intersection of the set D_RD, D_P and D_N, and then the detection result D was generated.

[3] http://www.cs.waikato.ac.nz/ml/weka/

Algorithm 2: Detection algorithm

Input: The set of users' feature space F_RDBRS_u, F_PBRS_u and F_NBRS_u; The number of clusters k;
Output: The detected result D;
Step 1: $\{C\_\mathrm{RD}_1, C\_\mathrm{RD}_2\} \leftarrow \mathrm{EM}(F\_\mathrm{RDBRS}_u)$;
Step 2: $\{C\_\mathrm{P}_1, C\_\mathrm{P}_2\} \leftarrow \mathrm{EM}(F\_\mathrm{PBRS}_u)$;
Step 3: $\{C\_\mathrm{N}_1, C\_\mathrm{N}_2\} \leftarrow \mathrm{EM}(F\_\mathrm{NBRS}_u)$;
Step 4: $D\_\mathrm{ARD} = \min(C\_\mathrm{RD}_1, C\_\mathrm{RD}_2)$, $D\_\mathrm{P} = \min(C\_\mathrm{P}_1, C\_\mathrm{P}_2)$, $D\_\mathrm{N} = \min(C\_\mathrm{N}_1, C\_\mathrm{N}_2)$;
Step 5: $D \leftarrow \{D | D\_\mathrm{RD} \cap D\_\mathrm{P} \cap D\_\mathrm{N}\}$;
Return D.

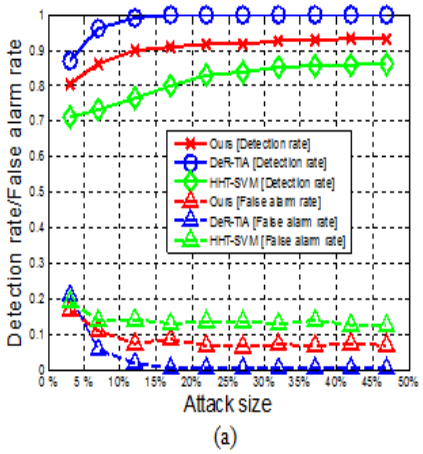


Fig. 11. The comparison of detection rate and false alarm rate in different attack sizes. (a) Grey rating is 1, filler size is 5%, single-target bandwagon (random) attack; (b) Grey rating is 3, filler size is 5%, single-target bandwagon (random) attack

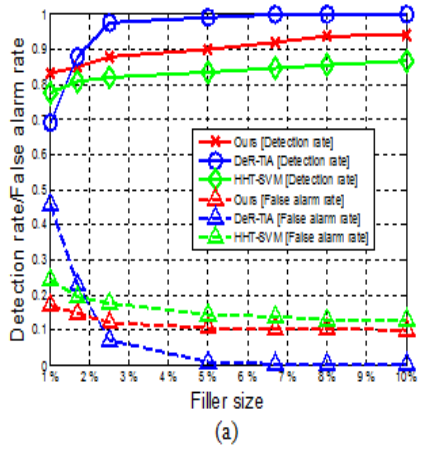


Fig. 12. The comparison of detection rate and false alarm rate in different filler sizes. (a) Grey rating is 1, attack size is 17%, single-target bandwagon (random) attack; (b) Grey rating is 3, attack size is 17%, single-target bandwagon (random) attack

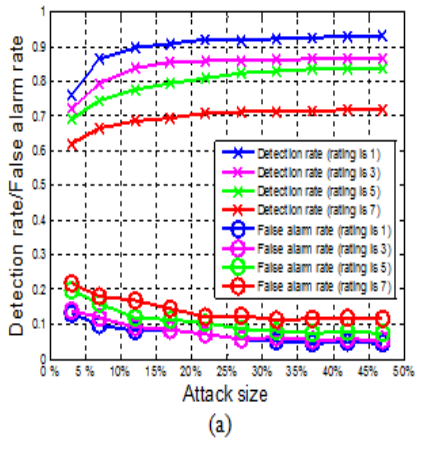


Fig. 13. The comparison of detection rate and false alarm rate with different grey ratings in single-target attack. (a) Filler size is 5%, attack size varies in bandwagon (average) attack. (b) Attack size is 17%, filler size varies in bandwagon (average) attack

TABLE IV. COMPARISON OF THE DETECTION PERFORMANCE OF OUR METHOD WITH TWO BENCHMARKED METHODS

Attack models	Methods	Rating							
		1		*3*		*5*		*7*	
		DR	***FAR***	***DR***	***FAR***	***DR***	***FAR***	***DR***	***FAR***
AOP	HHT-SVM	0.845	0.095	0.819	0.15	0.79	0.177	0.673	0.21
	DeR-TIA	1.0	0.005	0.715	0.185	0.734	0.225	0.707	0.275
	Ours	0.911	0.0785	**0.835**	**0.093**	**0.813**	**0.102**	**0.702**	**0.11**
Random	HHT-SVM	0.819	0.12	0.765	0.15	0.7345	0.14	0.68	0.21
	DeR-TIA	1.0	0.0025	0.735	0.175	0.727	0.195	0.731	0.265
	Ours	0.904	0.081	**0.834**	**0.086**	**0.801**	**0.093**	**0.707**	**0.11**
Average	HHT-SVM	0.873	0.1091	0.782	0.13	0.759	0.158	0.665	0.182
	DeR-TIA	1.0	0.0025	0.763	0.165	0.750	0.205	0.752	0.195
	Ours	0.907	0.085	**0.837**	**0.090**	**0.805**	**0.079**	**0.703**	**0.125**
Bandwagon (average)	HHT-SVM	0.906	0.09	0.8279	0.14	0.7869	0.16	0.675	0.19
	DeR-TIA	1.0	0.005	0.755	0.18	0.734	0.25	0.752	0.285
	Ours	0.935	0.0615	**0.852**	**0.0713**	**0.823**	**0.0682**	**0.705**	**0.115**
Bandwagon (random)	HHT-SVM	0.910	0.095	0.8179	0.13	0.8069	0.18	0.67	0.21
	DeR-TIA	1.0	0.005	0.747	0.165	0.735	0.205	0.750	0.27
	Ours	0.934	0.055	**0.868**	**0.075**	**0.83**	**0.069**	**0.718**	**0.115**
Segment	HHT-SVM	0.897	0.0891	0.819	0.13	0.7869	0.167	0.667	0.193
	DeR-TIA	1.0	0.0055	0.752	0.15	0.730	0.185	0.731	0.25
	Ours	0.915	0.075	**0.846**	**0.08**	**0.815**	**0.086**	**0.70**	**0.11**
Reveres bandwagon	HHT-SVM	0.895	0.087	0.8179	0.125	0.796	0.145	0.66	0.195
	DeR-TIA	1.0	0.005	0.739	0.175	0.754	0.185	0.727	0.26
	Ours	0.933	0.065	**0.868**	**0.075**	**0.815**	**0.0775**	**0.705**	**0.125**
Love/Hate	HHT-SVM	0.849	0.105	0.807	0.135	0.7569	0.175	0.67	0.205
	DeR-TIA	1.0	0.0025	0.752	0.16	0.727	0.195	0.750	0.24
	Ours	0.917	0.075	**0.845**	**0.065**	**0.81**	**0.0785**	**0.717**	**0.135**

V. EXPERIMENTS AND ANALYSIS

In this section, we firstly show the experimental data and settings on a real-world dataset. Then, we discuss our experimental results.

A. Experimental data and settings

In our experiments, we use the Book-Crossing [4] dataset. It contains 278,858 users providing 1,149,780 ratings (explicit or implicit) about 271,379books and each rater had to rate at least 1 books. All ratings are in the form of integral values between minimum value 1 and maximum value 10. The minimum score means the rater dislikes the book, while the maximum score means the rater enjoyed the book. We randomly select 800 genuine profiles from the dataset as the samples of genuine profiles. For the attack profiles, we just focus on nuke attacks and their grey attacks, push attacks can be detected in the analogous manner. For each attack model (as shown in Table 2), we respectively generate nuke and grey attack profiles according to the corresponding attack models with diverse attack sizes [5] {3%, 7%, 12%, 17%, 22%, 27%, 32%, 37%, 42%, 47%} and filler sizes [6] {1%, 1.7%, 2.5%, 5%, 6.7%, 8%, 10%}. In addition, to ensure the rationality of the results, the target item is randomly selected for these attack profiles. Especially in Table 2, the r_{grey} is the grey rating on target items rated by lower scores such as 1, 3, 5 and 7.

The generated attack profiles are respectively inserted into the sampled genuine profiles to construct our test datasets. Therefore, we have 560 ($8 \times 10 \times 7$) test datasets including 8 diverse attack models, 10 different attack sizes and 7 different filler sizes. Notice that, these process is repeated 10 times and the average value of detection results are reported for the experiments. All numerical studies are implemented using MATLAB R2012a on a personal computer with Intel(R) Core(TM) i7-4790 3.60GHz CPU, 16G memory and Microsoft windows 7 operating system.

To measure detection performance of the proposed methods, we use detection rate and false alarm rate in our experiments.

$$\text{detection rate} = \frac{|D \cap A|}{|A|} \tag{11}$$

$$\text{false alarm rate} = \frac{|D \cap G|}{|G|} \tag{12}$$

where D is the set of the detected user profiles, A is the set of attacker profiles, and G is the set of genuine user profiles [11].

B. Experimental results and analysis

To validate the detection performance of our proposed method, we employ two benchmarked methods including HHT-SVM [17] and DeR-TIA [1] to demonstrate the outperformance of our method. Take bandwagon (random) attack for example, Figures 11 and 12 demonstrate how each method performs under varying attack sizes and filler sizes, respectively. In the bandwagon (random) attack, a group isolated attackers always provide maximal or minimal or grey rating on a set of items when they are selected as the selected items or the filler items. As shown in Figures 11(a) and 12(a), the detection rate increased gradually and false alarm rate decreased gradually when the attack size increased and the filler size is fixed with 5% (in Figure 11 (a)) and filler size increased and attack size is 17% (in Figure 12 (a)). In addition, we can observe that our method shows significantly better

[4] http://www.informatik.uni-freiburg.de/~cziegler/BX/

[5] The ratio between the number of attackers and genuine users.

[6] The ratio between the number of items rated by user u and the number of entire items in the recommender systems.

detection performance than HHT-SVM with the attack size increased. This might be attributed to the combination of novelty-based, popularity-based and rating deviation-based rating series adopted by our proposed algorithm. The rating deviation-based strategy calculates a rating offset on a target item which can identify between the genuine profiles and attack profiles. The second observation is that DeR-TIA shows the best performance among the three algorithms. With the attack size increasing, the detection rate almost keeps maximum 100% and the false alarm rate almost keeps minimum 0, except for the early stages (attack size < 17%) as illustrated in Figure 11 (a). The same observations are also clear in Figure 12(a). However, for grey rating, as shown in Figures 11 (b) and 12 (b), we set a grey rating equal to 3 (integer rating from 1-10 in the datasets). Our method shows the best detection performance among the three methods, although the detection rate of our method shows lower than DeR-TIA in the early stage (attack size < 12%) as illustrated in Figure 11 (b). To compare with our proposed method and HHT-SVM, DeR-TIA shows higher false alarm rate than the others. Moreover, the detection rate of DeR-TIA almost remained unchanged with the attack size increased, and similar results can be observed in Figure 12 (b). The results might be attributed to grey rating. The first phase of DeR-TIA can filter out a part of genuine users by using similarity threshold, but it is difficult to capture the suspected profiles which rate grey ratings in their second phase. They defend and remove the suspected users almost depend on the similarity threshold, so they perform lower detection performance. For our proposed method, we pay more attention to the details of the all ratings rated by a user and explore the top-N items which has sorted by the rating deviation of item in order to characterize the grey ratings.

To examine the detection performance of our method in bandwagon (random) attack with different grey ratings (take bandwagon (random) attack for example), we conduct a list of experiments with diverse attack sizes and filler sizes. As shown in Figure 13, we perform 4 different grey ratings including 1, 3, 5 and 7 on the target items. One observation is that the detection rate gradually increased and false alarm rate gradually decreased with the attack size increasing (in Figure 13 (a)) or filler size increasing (in Figure 13 (b)). The other observation is that the detection performance gradually performs poor when the grey rating increased from 1 to 7, regardless of different attack sizes and filler sizes. The results may indicate that the grey ratings are close to average rating in the entire system with the grey rating on the target items increasing. The attackers rate a mean rating may show a rating behavior like genuine users, which is difficult to discriminate between attackers and genuine users and shows higher false alarm rate.

To further illustrate the detection performance of our proposed method under different attack models with different grey ratings, we conduct a list of experiments in 8 attack models for comparing the performance of our proposed method with HHT-SVM and DeR-TIA. We use 4 different ratings including 1, 3, 5 and 7 score when filler size is 5% and attack size is 17%. As shown in Table 4, we can observe that the detection rate (DR) of our method reports higher than other two benchmarked methods when the grey rating increasing, except for the grey rating is 1. Similarly, the false alarm rate (FAR) of our method reports lower than others. In addition, the second observation is that the proposed method reports better detection performance under bandwagon (both random and average) and reverse bandwagon attacks in comparison with the other attack models, especially for grey ratings (such as 3, 5 and 7 score). These results may indicate that we combine the rating deviation-based, novelty-based and popularity-based rating series in our method is useful to discriminate difference between grey attack profiles and genuine profiles. The rating deviation-based rating series may easily characterize the grey attacks in comparison with the other two methods.

VI. Conclusions and Future Work

In this paper, we highlighted the challenges faced by the grey attacks, and then we develop an unsupervised detection approach based on discrete wavelet transform by combing the rating deviation-based, novelty-based and popularity-based rating series. Extensive experiments on the Book-Crossing dataset have demonstrated the effectiveness of the proposed approach. One of the limitations of our proposed method directly comes from the time consumption, which constructs the signals of rating series. In our future work, we intend to extend and improve grey attack detection in the following directions: 1) Considering more attack models such as Power users attack or Power items attack, etc.; 2) We will explore specific and simple method to detect grey attacks and develop better approach to construct the rating series. 3) Extracting more simpler and effective features to characterize grey attack profiles is still an open issue.

Acknowledgment

The research is supported by NFSC (61175039, 61221063), 863 High Tech Development Plan (2012AA011003), Research Fund for Doctoral Program of Higher Education of China (20090201120032), International Research Collaboration Project of Shaanxi Province (2013KW11) and Fundamental Research Funds for Central Universities (2012jdhz08).

References

[1] W Zhou, Y. S. Koh, J. H. Wen, S Burki and G Dobbie. Detection of abnormal profiles on group attacks in recommender systems. Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, Pages 955-958, 2014.

[2] D. Jia, F. Zhang and S. Liu. A robust collaborative filtering recommendation algorithm based on multidimensional trust model. Journal of Software, vol. 8, no. 1, 2013.

[3] R. Burke, B. Mobasher and C. Williams. Classification features for attack detection in collaborative recommender systems. In Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining, pages 17–20, 2006.

[4] B. Mobasher, R. Burke and J. Sandvig. Model-based collaborative filtering as a defense against profile injection attacks. AAAI. 1388, 2006.

[5] K. Bryan, M. O'Mahony and P. Cunningham. Unsupervised retrieval of attack profiles in collaborative recommender systems. In RecSys'08: Proceedings of the 2008 ACM conference on Recommender systems, pages 155–162 , 2008.

[6] H. Hurley, Z. Cheng and M. Zhang. Statistical attack detection. In: Proceedings of the Third ACM Conference on Recommender Systems (RecSys'09), pages 149–156 , 2009.

[7] B. Mehta. Unsupervised shilling detection for collaborative filtering. AAAI, 1402-1407, 2007.

[8] C Li and Z Luo. Detection of shilling attacks in collaborative filtering recommender systems. In: Proceedings of the international conference of soft computing and pattern recognition, Dalian, China, pages 190–193, 2011.

[9] I Gunes, C Kaleli, A Bilge and H Polat. Shilling attacks against recommender systems: A comprehensive survey. Artificial Intelligence Review, pages 1-33, 2012.

[10] N Giseop, Y. Kang and C. Kim. Ecsy-Recsy: Considering Sybil attack with time dynamics and economics in recommender system. International Conference on Information Networking (ICOIN), pages 566 - 571, 2013.

[11] C. Chung, P. Hsu and S. Huang. βP: A novel approach to filter out malicious rating profiles from recommender systems. Journal of Decision Support Systems, pages 314–325, April 2013.

[12] X. Zhang, T. Lee and G Pitsilis. Securing recommender systems against shilling attacks using social-based clustering. Journal of Computer Science and Technology (JCST), pages 616-624, July 2013.

[13] Z Zhang and S. Kulkarni. Graph-based detection of shilling attacks in recommender systems. IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pages 1-6, 2013.

[14] B. Mehta, T. Hofmann and P. Fankhauser. Lies and propaganda: detecting spam users in collaborative filtering. In: IUI '07: Proceedings of the 12th International Conference on Intelligent User Interfaces, pages 14–21, 2007.

[15] M Morid and M Shajari. Defending recommender systems by influence analysis. Information Retrieval, pages 137-152, April 2014.

[16] Z. Wu, J Cao, B Mao and Y. Zhang. Semi-SAD: Applying semi-supervised learning to shilling attack detection. Proceedings of the 5th International Conference on Recommender Systems. New York: ACM, pages 289–292, 2011.

[17] F. Zhang and Q. Zhou. HHT–SVM: An online method for detecting profile injection attacks in collaborative recommender systems, Knowl. Based Syst. 2014.

[18] J Zou and F Fekri. A belief propagation approach for detecting shilling attacks in collaborative filtering. Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (CIKM), pages 1837-1840, 2013.

[19] Z Zhang and SR Kulkarni, Detection of Shilling Attacks in Recommender Systems via Spectral Clustering. 2014 17th International Conference on Information Fusion (FUSION). Page(s):1-8, 7-10 July 2014.

[20] Fidel Cacheda, Victor Carneiro, Diego Fernandez and vreixo Formoso. Comparison of Collaborative Filtering Algorithms: Limitations of Current Techniques and Proposals for Scalable, High-Performance Recommender Systems. ACM Transactions on the Web (TWEB), Volume 5, Issue 1, February 2011.

[21] B. Mobasher, R. Burke, B. Bhaumil and C. Williams. Towards trustworthy recommender systems: an analysis of attack models and algorithm robustness. ACM Transactions on Internet Technology, 7 (4), pages 23–38, 2007.

[22] C. E. Seminario and D. C. Wilson. Attacking item-based recommender systems with power items. RecSys'14, October 6-10, 2014.

[23] M. J. Shensa, Wedding the a trous and Mallat algorithms, IEEE Trans. Signal Process. 40 (1992), 24642482.

[24] Williams, C., Mobasher, B., Burke, R., Sandvig, J., Bhaumik, R. Detection of obfuscated attacks in collaborative recommender systems. In: Workshop on Recommender Systems, ECAI, 2006.

[25] J. S. Lee, D. Zhu, Shilling attack detection: a new approach for a trustworthy recommender system, JNFORMS J. Comput. 24 (1) , pages 117–131, 2011.

[26] B. Mehta, W. Nejdl, Unsupervised strategies for shilling detection and robust collaborative filtering, User Model. User-Adap. Inter. 19 (1–2), pages 65–79, 2009.

[27] Mohamed Hamdi, Noureddine Boudriga. Detecting denial-of-service attacks using the wavelet transform. Computer Communications, 30 (16) (2007), pp. 3203–3213.

[28] C.A. Williams, B. Mobasher, R. Burke, R. Bhaumik, Detecting profile injection attacks in collaborative filtering: a classification-based approach, in: Proceedings of the 8th Knowledge Discovery on the Web International Conference on Advances in Web Mining and Web Usage Analysis (Lecture Notes in Computer Science), Springer-Verlag, 2007, pp. 167–186.

[29] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams, "Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness," ACM Transactions on Internet Technology (TOIT), Volume 7 , Issue 4 (October 2007), 2007.

[30] Z. A. Wu, J. J. Wu, J. Cao, D. C. Tao, HySAD: a semi-supervised hybrid shilling attack detector for trustworthy product recommendation, in: 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Beijing, China, August, 2012, pp. 985–993.

[31] F. Zhang, Q. Zhou, A meta-learning-based approach for detecting profile injection attacks in collaborative recommender systems, J. Comput. 7 (1) (2012) 226-234.

[32] F. He, X.Wang, B. Liu, Attack detection by rough set theory in recommendation system, in: Proceedings of 2010 IEEE International Conference on Granular Computing, 2010, pp. 692-695.

[33] W. Zhou, J. Wen, Y. S. Koh, Q. Xiong, M. Gao, G. Dobbie, and S. Alam. Shilling attacks detection in recommender systems based on target item analysis. PloS one, 2015.

Implementation of Computer Assisted CIPP Model for Evaluation Program of HIV/AIDS Countermeasures in Bali

I Made Sundayana
Director/Lecture of Health Education
Buleleng School of Health
Bali, Indonesia

***Abstract*—One of the fact within economical development of tourism in Bali is indicated by established tourism facilities in order to support Bali tourism industry. Consquently, It has brought up effect that large numbers of new citizen search for occupation to Bali.Those people who came and settle in Bali temporaly or permanently, consequently Bali become heterogeneous.Thus, Bali become over populated. Since, over populated in Bali has risen up the economic sector and it has been spreading HIV /AIDS rapidly. As anticipation and prevention for contagious, developed and spreading HIV of Bali Provinse has regulated (regional act) Number 3 2006 concerning of prevention act fr HIV/AIDS. As the matter of fact, regional act is not properly conducted yet as, therefore it is evaluation required f0r the rule and program that have been conducted by the government.One of the technical evaluation can be applied is CIPP model. However, CIPP model is still applied in conventional way and it has not yet contributed accurate evaluational count in processing the data, therefore by using CIPP model of computer assistance. This can be proved by ending up the result of the total program percentage of HIV /AID prevention by conventional counted result as much 88.000%, meanwhile the count with computer assistance end up with 88.400% in result. It shows high category.**

Keywords—Evaluation; Computer Assisted CIPP Model

I. INTRODUCTION

Tourism sector is a significant sector in order to achieve regional revenue goverment income.Tourim shall be perceived from several point of view as its complexity embedded within tourism activity. Among of those activity are tourism as resource, tourism as business, and toursm as industry. Those things indicate that tourism has potential in order to support economic sector.

One of the reality in economy base on develoving of Bali's tourism is facilities that had been established as an effort to support Bali Tourism. By Having established various of business, consequently, new comers have come to seek occupations. Pople who came to settle permanently or temporally having social interaction with local that creating heterogen society. Heterogenity is causing over populated in Bali, however it's rising up economy sector as well as spreading infected desease HIV/AIDS.

Masiive spreading of HIV/AID indicates high rate of infection. In Bali particularly HIV/AIDS infected not only in urban but also in rural. Large numbers of HIV/AIDS cases rose in rural . Until nowdays, the process of preventing HIV/AIDS structurally involves formal institutions, and traditional instutions yet socialized in rural based on geographical reason and daily activity of the traditional society.

On other side, effort to prevent HIV/AIDS consider the government policy voint of view, whereas the HIV/AIDs's subject and object is it's own. Various action of anticipation or prevention of spreading and contagious HIV/AIDs, Bali Province has Local Act Number 3 2006 regarding HIV/AIDS, however the provision is unble to well manage, therefore it is necessary to evaluate the act program which is conducted by the government.

One of the technical evaluation applied is CIPP model, However, CIPP model that has been applied conventionally yet shows accurate counted evaluation in processing its data.

It is appropriate on the results of research conducted by Dewa Gede Hendra Divayana about Program Evaluation of Management E-learning shows the model is done in the conventional CIPP still not provide an accurate evaluation calculation of the data processing[1]. From the results of these studies, the authors are interested in continuing the development of conventional CIPP model evaluation toward a computer assisted CIPP model.

II. LITERATURE REVIEW

A. Evaluation

In [2], Evaluation is a mean for understanding how things going.

In [3], Evaluation can be defined as the determination of conformity between the results achieved and the objectives to be achieved.

In [4], Evaluation can be defined as an activity or process to provide or specify a value above a certain object, things, institutions, and programs.

In [5], evaluation is a systematic and ongoing process to collect, describe, interpret and present information about a program to be used as a basis for making decisions.

From the opinions of the above can be concluded in general that the evaluation is an activity in collecting, analysing, and presenting information about an object of research and the results can be used to take a decision.

B. CIPP Model

In [6], the core concept of this model denoted by the CIPP acronym, which stands for the evaluation context, input, process, and product.

In [7], the CIPP evaluation there are four components that must be passed is the evaluation of the component context, the evaluation of input component, the evaluation of process components, and the evaluation of product components.

In [8], the CIPP model evaluation consists of four types, namely: context evaluation, input evaluation, process evaluation and product evaluation.

In the evaluation context is carried out to identification and assessment of the needs that underlie the program formulation. The input evaluation carried out to choose among several existing planning. In the process evaluation is carried out to access the implementation of the plan has been set. And the product evaluation conducted to identify and access the outputs and benefits of a program.

In [9], basically the CIPP evaluation model requires that a series of questions will be asked about four different elements of the model on the context, input, process, and product.

From the above opinions can be concluded in general that the CIPP model is a model in its activities through four stages of evaluation are: evaluation of the component context, input, process and product.

III. Methodology

A. Object dan Research Site

1) Research Object is HIV/AIDS countermeasures program.

2) Research Site at Health Department of Bali Province.

B. Data Type

In this research, the authors use primary data, secondary data, quantitative and qualitative data.

C. Data Collection Techniques

In this research, the authors use data collection techniques such as interviews, observation, and documentation.

D. Analysis Techniques

Analysis techniques used in this research is descriptive statistical.

E. Aspect of Evaluation

The aspects evaluated in HIV/AIDS countermeasures program can be seen in Table I bellow.

TABLE I. Evaluation Criteria

No	Component	Aspects
1.	Context	Local regulations of HIV/AIDS
		The mission and purpose of program
		Readiness from Head of Health Department in implementing the regulations of HIV/AIDS
2.	Input	Guide of the program implementation
		Human resources
		Facilities and infrastructure
3.	Process	Program planning of HIV/AIDS countermeasures
		Program implementation of HIV/AIDS countermeasures
4.	Product	The impact of implementation of HIV/AIDS countermeasures program
		The expected outcome form implementation of HIV/AIDS countermeasures program

IV. Result and Discussion

A. Result

The research results can be seen in Table II below.

TABLE II. Evaluation Results of Hiv/Aids Countermeasures Program with Cipp Model in Conventional

No	Dimension	Aspects	Respondents Score					X	%
			R1	R2	R3	R4	R 5		
1.	Context	C1	5	4	5	4	4	4.4	88
		C2	5	4	4	5	5	4.6	92
		C3	5	4	4	4	5	4.4	88
Percentage of Effectiveness on Context Dimension									89
2.	Input	I1	5	5	4	5	5	4.8	96
		I2	4	5	5	4	4	4.4	88
		I3	5	4	4	5	4	4.4	88
Percentage of Effectiveness on Input Dimension									91
3.	Process	P1	4	4	4	4	5	4.2	84
		P2	4	5	5	4	4	4.4	88
Percentage of Effectiveness on Process Dimension									86
4.	Product	O1	5	4	4	5	4	4.4	88
		O2	4	5	4	4	4	4.2	84
Percentage of Effectiveness on Product Dimension									86
Total Percentage of Effectiveness									88
Category									High

Explanation :

C1 : Local regulations of HIV/AIDS
C2 : The mission and purpose of program
C3 : Readiness from Head of Health Department in implementing the regulations of HIV/AIDS
I1 : Guide of the program implementation
I2 : Human resources

I3 : Facilities and infrastructure
P1 : Program planning of HIV/AIDS countermeasures
P2 : Program implementation of HIV/AIDS countermeasures
O1 : The impact of implementation of HIV/AIDS countermeasures program
O2 : The expected outcome form implementation of HIV/AIDS countermeasures program
X : Average
% : Percentage

Category of scale effectiveness:
Highest : 90%-100%
High : 80%-89%
Sufficient : 70%-79%
Low : $\leq$ 69%

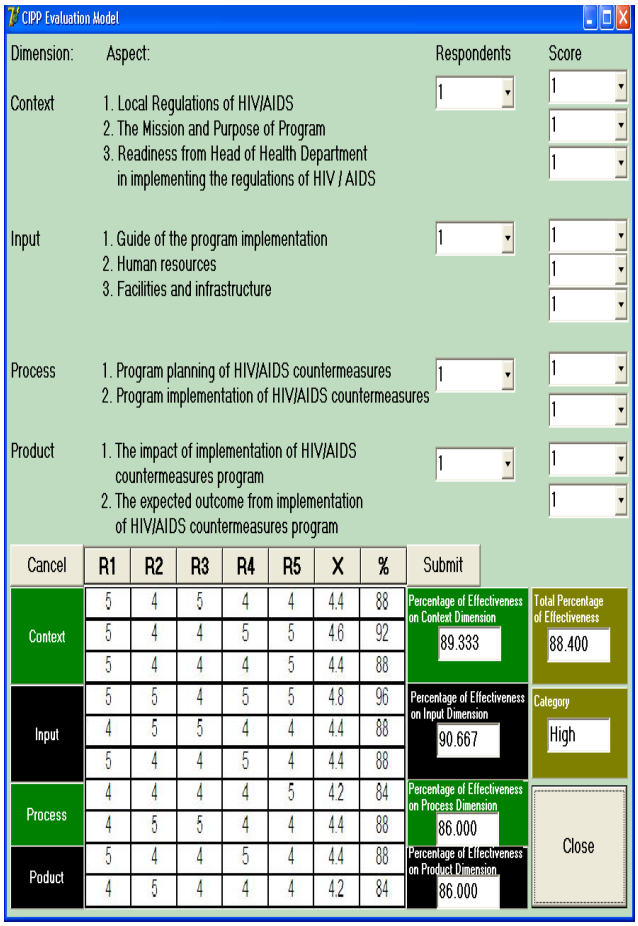


Fig. 1. Evaluation Results of HIV/AIDS Countermeasures Program With Computer Assisted CIPP Model

From the above results can be seen clearly that the results of the evaluation by Computer Assisted CIPP model calculation shows the results more accurate than using the conventional calculation method. It is seen from the results the percentage of effectiveness on context dimension with the conventional calculation shows result of 89.000%, while the computer aided calculation shows result of 89.333%. The percentage of effectiveness on input dimension with the conventional calculation shows result of 91.000%, while the computer-aided calculation shows result of 90.667%. The percentage of effectiveness on process dimension with the conventional calculation shows result of 86.000%, while the computer-aided calculation also obtained the same result of 86.000%. The percentage of effectiveness on product dimension with the conventional calculation shows result of 86.000%, while the computer-aided calculation also obtained the same result by 86.000%. The Total Percentage of Effectiveness of HIV/AIDS countermeasures program with the conventional calculation shows result of 88.000%, while the computer-aided calculation shows result of 88.400% with the higher category.

V. Conclusions

There is concusion to be drawn from this research that by applying CIPP evaluasion model based on computer assistance shall achieve more accurate and rapid counting compare to conventional way of counting. There after, decicion maker shall be quicker in order to make recomandation within decicion making whether the program shal be terminated or be proceded.

Acknowledgments

The author express their gratefulness to staff at Buleleng School of Health for all support and motivation. The author also generously thank to Dewa Gede Hendra Divayana, Ph.D. as Lecture at IT Education, Informatics Department, Ganesha University of Education.

References

[1] D.G.H. Divayana. Program Evaluation of Management E-learning (Thesis). Surabaya: YAPAN Surabaya School of Economic, 2014.

[2] L.H. Kuo, *et.al*., "An Evaluation Model of Integrating Emerging Technology into Formal Curriculum," in The International Journal of Education and Information Technologies Vol.6, No.3, 2012,pp.250-259.

[3] D. Mardapi, Measurement, Assessment, and Evaluation of Education (1st Edition). Yogyakarta: Nuha Medika, 2012.

[4] S. Rutoto, "Observing the Guidance and Counseling Program Evaluation at School Present and Future," in Mawas Vol.1,No.1, 2010,pp.1- 15.

[5] H. Sundoyo, T. Sumaryanto, Dwijanto, "Program Evaluation of Dual System Education Based of Countenance Stake Model," in Innovative Journal of Curriculum Vol.1,No.2,2012,pp.69-73.

[6] D.L. Stufflebeam, C.L.S. Coryn, "Evaluation Theory, Models, and Applications (2nd Edition). San Francisco: Jossey-Bass,2014.

[7] Wirawan. Evaluation Theory, Model, Standards, Applications, and Profession (1st Edition). Jakarta: Rajawali Pers,2011.

[8] M. Tiantong, P. Tongchin, "A Multiple Intelligences Supported Web-based Collaborative Learning Model Using Stufflebeam's CIPP Evaluation Model," in International Journal of Humanities and Social Science Vol.3, No.7,2013,pp.157-165.

Authors Biography

Ns. I Made Sundayana, S.Kep., M.Si. was born in Cipanas, West Java, in 1969. He worked as Director and also Lecture at Buleleng School of Health, Buleleng, Bali. Now, he is completing doctoral program studies at Jakarta State University.

Prediction of New Student Numbers using Least Square Method

Dwi Mulyani
College of Informatics And Computer Management (STMIK) Banjarbaru
Banjarbaru Kalsel, Indonesia

***Abstract*—STMIK BANJARBARU has acquired less number of new students for the last three years compared to the previous years. The numbers of new student acquisition are not always the same every year. The unstable number of new student acquisition made the difficulty in designing classes, lecturers, and other charges. Knowing the prediction number of new student acquisition for the coming period is very important as a basis for further decision making. Least Square method as the method of calculation to determine the scores prediction is often used to have a prediction, because the calculation is more accurate then moving average.**

The study was aimed to help the private colleges or universities, especially STMIK BANJARBARU, in predicting the number of new students who are accepted, so it will be easier to make decisions in determining the next steps and estimating the financial matters.

The prediction of the number of new student acquisition will facilitates STMIK BANJARBARU to determine the number of classes, scheduling, etc.

From the results of the study, it can be concluded that prediction analysis by using Least Square Method can be used to predict the number of new students acquisition for the coming period based on the student data in the previous years, because it produces valid results or closer to the truth. From the test results in the last 3 years, the validity shows 97.8%, so it can be said valid.

Keywords—Prediction of New Students; Least Square method

I. INTRODUCTION

STMIK BANJARBARU as one of the colleges in the field of computer becomes one destination for new student candidates to continue their education. In the first year, the number of student candidates could be predicted since STMIK BANJARBARU was the only one computer college in South Kalimantan. However, in the last few years there are many other universities in South Kalimantan which provide department of computer science so it is assumed that the number of student candidates are divided into some computer universities in South Kalimantan. This causes the regression of the number of new students in the last three years. Based on the student data of STMIK BANJARBARU, the numbers of New Students accepted were 407 in 2009, 516 in 2010, 528 in 2011, 374 in 2012, 375 in 2013 and 386 in 2014. In 2012 to 2014, the number of new students accepted in STMIK BANJARBARU has decreased to 29.16%.

The problem faced by STMIK BANJARBARU is estimating the number of new students due to the regression of the number of new student acquisition in the last three years.

The study was aimed to help the private universities, especially STMIK BANJARBARU, in predicting the number of new students who are accepted, so it will be easier to make decisions in determining the next steps and estimating the financial matters. The prediction of the number of new student acquisition will facilitates STMIK BANJARBARU in determining the number of classes, setting schedules and others.

II. THEORITICAL BASIS

A. Prediction or forecasting

Prediction or forecasting is an important tool in an effective and efficient plan, especially in the economic field. In modern organizations, knowing the coming state is very important to look at the good or bad and aimed to prepare for the next activities (Rambe,2012).

According Heizer and Render (2009: 162), forecasting is the art and science to predict future events. This can be done by involving the retrieval of historical data and projected into the future with a form of mathematical models or predictions are subjective intuition, or using a combination of mathematical models that are tailored to the good judgment of a manager.

According Prasetya and Lukiastuti (2009: 43), forecasting is an attempt to predict the future state through state testing in the past. Forecasting relates to attempt to predict what happens in the future, based on the scientific method (science and technology) and carried out mathematically. However, forecasting activities are not solely based on scientific procedures or organized, because there is activity forecasting that uses intuition (feeling) or through informal discussions in a group.

According to Yamit, Forecasting is a prediction, projection or estimate of the level of an uncertain event in the future (Rambe 2012). According Makridakis, Forecasting is predictive values of a variable based on the known value of the variable or variables related (Rambe 2012).

According Pangestu Subagyo (1991: 1), forecasting is an activity / business to know (event) will happen in the future

regarding a particular object by using experience / historical data. According to T Hani Handoko (1994; 260) Forecasting is an attempt to predict the future state through testing in the past.

From some of these explanations, it can be concluded that forecasting is a process or method of predicting an event that will occur in the future by basing it self on certain variables.

B. Least Square method

Least Square is a method for determining the approach polynomial function y = f (x) closest to the data (x1, y1) to (xn, yn). (Basuki 2014)

In the Collins English Dictionary says that the Least Square method is the best method to determine the value of an unknown quantity related to one or more sets of observations or measurements. (Harper Collins, 1991, 1994, 1998, 2000, 2003).

According to Dr. Setijo in his module entitled "Linear Regression with Least Squares Method" said that Least Square method is an approach method which is widely used for:

1) *Regression modeling based on the equation of the discrete data points*

2) *Analysis of measurement error (model validation) (Kristalina, 2015).*

This method is most often used to predict (Y), because the calculation is more accurate. The equation of the trend line to be searched is

$$Y = a0 + bx \quad a = (\Sigma Y) / n \qquad (1)$$

$$b = (\Sigma XY) / \Sigma x2 \qquad (2)$$

with:

Y = periodic data (time series) = estimated trend value.

a0 = trend value in the base year.

b = the annual average growth of value trend.

x = time variable (days, weeks, months or years).

To perform a calculation, it will require a specific value on the time variable (x), so that the amount of the time variable value is zero or $\Sigma x = 0$.

In this case,it will be devoted to discuss the analysis of time series with Least Square method which is divided into two cases, namely the even data case and the odd data case.

For odd n, then:

1) *The distance between the two time is rated one unit.*
2) *Above 0 is marked negative*
3) *Under 0is marked positive.*

For even n, then:

1) *The distance between the two time is rated two units.*
2) *Above 0 is marked negative*
3) *Under 0 is marked positive.*

In the data processing of odd data, the registration data of new student candidates for the past five years is required, which are the registration data in 2010 until the registration data in 2014. In the data processing of even data, the registration data of new student candidates for the past five years is required, which are the registration data in 2009 until the registration data in 2014.

In general, linear line equation of time series analysis is:

$$Y = a + b\,X. \qquad (3)$$

Information:

Y is the variable whichits trendis sought

X is the variable of time (years).

While, to find constant value (a) and parameter (b) is:

$$a = \Sigma Y / N \qquad (4)$$

and

$$b = \Sigma XY / \Sigma X2 \qquad (5)$$

III. System Analysis and Design

A. Literature Review

The previous study, conducted by Muhammad Ihsan Fauzi Rambe in 2012, examined the prediction of medicine supply using least square method which took the case study at Mutiara Hati Pharmacy, Medan. The study found that Least Square Method can be used to predict the medicine sales in the coming period based on the sales data in the previous year. Further, the analysis application can yield predictions and has minimized the forecast errors of the level of medicine sale in Pharmacies.

B. Data Requirements

The data required in the study is the data of Students accepted in STMIK BANJARBARU in 2009 to 2014.

TABLE I. DATA OF NEW STUDENTS OF STMIK BANJARBARU

No.	Year	Total
1	2009	407
2	2010	513
3	2011	528
4	2012	374
5	2013	375
6	2014	385

(Source: PMB STMIK BANJARBARU)

IV. Results and Discussion

A. Odd Data Case

Before calculating the prediction of new student acquisition in 2015, some trialswere conducted in calculating the number of new student acquisition in 2012, 2013 and 2014 to determine the validity of the Least Square Method formula.

In calculating the prediction result of the number of students in 2012, the researcher used the student data in 2007 to 2011.

TABLE II. DATA OF NEW STUDENTS IN 2007 TO 2011

No.	Year	Total
1	2007	350
2	2008	512
3	2009	407
4	2010	513
5	2011	528

(Source: PMB STMIK BANJARBARU)

The next step is determining the values of variable X, XY and X^2.

TABLE III. DATA OF NEW STUDENTS IN 2007 TO 2011

No.	Year	Amount(Y)	X	XY	X^2
1	2007	350	-2	-700	4
2	2008	512	-1	-512	1
3	2009	407	0	0	0
4	2010	513	1	513	1
5	2011	528	2	1056	4
Total		2310		357	10

(Source: PMB STMIK BANJARBARU)

It is known that:

$\Sigma Y=2310$

$N=5$

$\Sigma XY=357$

$\Sigma X2=10$

Then, to find the value of a:

$a =\Sigma Y/N$

$a =2310/5$

$a =462$

And to find the value of b:

$b =\Sigma XY/\Sigma X2$

$b=357/10$

$b=35.7$

After the values of a and b are obtained, then the linear line equation is:

$Y= a+bX$

$Y=462+(35.7) X$

With the calculated equation of linear line, the number of new student in 2012 can be predicted:

$Y=462+(35.7) X$(For the year of 2012, the value of X is 3)

so that:

$Y=462+(35.7 x3)$

$Y=462+107.1$

$Y=569.1$ (6)

It means that the number of new student candidates who registered in 2012 was 569 people.

The next was calculating the result of the number of new student acquisition in 2013. The data used was the data of new student in 2008 to 2012.

TABLE IV. NEW STUDENT DATA IN 2008 TO 2012

No.	Year	Total
1	2008	512
2	2009	407
3	2010	513
4	2011	528
5	2012	374

(Source:PMB STMIK BANJARBARU)

Next was determining the values of X, XY and X^2.

TABLE V. NEW STUDENT DATA IN 2008 TO 2012

No.	Year	Amount	X	XY	X^2
1	2008	512	-2	-1024	4
2	2009	407	-1	-407	1
3	2010	513	0	0	0
4	2011	528	1	528	1
5	2012	374	2	748	4
Total		2334		-155	10

(Source: PMB STMIK BANJARBARU)

It is known that:

$\Sigma Y = 2334$

$N = 5$

$\Sigma XY = -155$

$\Sigma X^2 = 10$

Then, to find the value of a:

$a = \Sigma Y / N$

$a= 2334/5$

$a= 466.8$

And to find the value of b:

$b= \Sigma XY/ \Sigma X^2$

$b= -155/10$

$b= -15.5$

After the values of a and b are obtained, then the equation of linear line is:

$Y = a + b X$

$Y = 466.8 + (-15.5)X$

With the calculated linear line, it can be predicted that the number of new students in 2013 is:

$Y = 466.8 + (-15.5) X$ (For the year of 2013, the value of X is 3)

Thus:

Y = 466.8– (15.5 x 3)

Y = 466.8 – 46.5

Y = 420 (7)

It means the number of candidates who registered in 2013was420. Next was calculating the number of new student acquisition in 2014. The data used was the student data in 2009 to 2013.

TABLE VI. DATA OF NEW STUDENT IN 2009 TO 2013

Year	Amount
2009	407
2010	513
2011	528
2012	374
2013	375
Total	2197

(Source:PMB STMIK BANJARBARU)

The next step is determining the variable values of X, XY and X^2.

TABLE VII. DATA OF NEW STUDENTS IN 2009 TO 2013

Year	Amount	X	XY	X^2
2009	407	-2	-814	4
2010	513	-1	-513	1
2011	528	0	0	0
2012	374	1	374	1
2013	375	2	750	4
Total	2197		-203	10

(Source: PMB STMIK BANJARBARU)

It is known that:

$\Sigma Y = 2197$

$N = 5$

$\Sigma XY = -203$

$\Sigma X^2 = 10$

Then, to find the value of a:

$a = \Sigma Y / N$

a= 2197/5

a= 439.4

And to find the value of b:

$b= \Sigma XY/ \Sigma X^2$

b= -203/10

b= -20.3

After the values of a and b are obtained, then the equation of linear line is:

Y = a + b X

Y = 439.4 + (-20.3)

With the calculated linear line, it can be predicted that the number of new students in 2014 is:

Y = 439.4+ (-20.3) X (For the year of 2014, the value of X is 3)

Thus:

Y = 439.4 – (20.3 x 3)

Y = 439.4 – 60.9

Y = 378.5 (8)

It means that the number of new student candidates who registered in 2014 was 378.

From the results of calculations in predicting the acquisition of the number of new students (6) (7) (8), it was found that in 2012 there was 569, in 2013 there was 420 and in 2014 there was 378. It is determined that if the deviation between the fact and the calculation with Least Square Method is >50 people, then the result is invalid. Compared to the tangible result obtained in 2012, the deviation is 34.2% (195 people), meaning that the result is invalid. In 2013, the deviation is 10% (45 people), meaning that the result is valid. In 2014, the deviation is 2.07% (8 people), meaning that the result is valid. From the three comparisons, it is found that two results are valid and one result is invalid. This means that the formula of Least Square Method is valid or closer to the truth. Next, the calculation would be performed to predict the number of new student acquisition in 2015. In the data processing of Odd Data case, the data of new students needed is the data from the last 5 years, from 2010 to 2014.

TABLE VIII. DATA OF NEW STUDENT REGISTRATION IN 2010 TO 2014

Year	Amount
2010	513
2011	528
2012	374
2013	375
2014	386
Total	2176

(Source: PMB STMIK BANJARBARU)

Then, determining the variable values of X, XY dan X^2.

TABLE IX. DATA OF NEW STUDENT REGISTRATION IN 2010 TO 2014

Year	Amount(Y)	X	XY	X^2
2010	513	-2	-1026	4
2011	528	-1	-528	1
2012	374	0	0	0
2013	375	1	375	1
2014	386	2	772	4
Total	2176		-407	10

(Source: PMB STMIK BANJARBARU)

Based on Table 3, the values of a and b will be discovered. To find the values of a and b:

It is known, that:

$\Sigma Y = 2176$

$N = 5$

$\Sigma XY = -407$

$\Sigma X^2 = 10$

Then, to find the value of a:

$a = \Sigma Y / N$

a= 2176/5

a= 435,2

And to find the value of b:

$b= \Sigma XY/ \Sigma X^2$

b= -407/10

b= -40.7

After the values of a and b are obtained, then to find the equation of linear line:

$Y = a + b X$

$Y = 435.2 + (-40.7)$

With the calculated equation of linear line, the number of new student in 2015 can be calculated:

Y = 435.2 + (-40.7) X (For the year of 2015, the value of X is 3)

So that:

$Y = 435.2 – (40.7 \times 3)$

$Y = 435.2 – 122.1$

$Y = 313.1$ (9)

It means that the numbers of new student candidates who register are 313 people.

B. Even Data Case

TABLE X. DATA OF NEW STUDENT REGISTRATION IN 2009 TO 2014

Year	Amount
2009	407
2010	513
2011	528
2012	374
2013	375
2014	386
Total	2583

(Source: PMB STMIK BANJARBARU)

The next step is determining the variable values of X, XY and X^2.

TABLE XI. DATA OF NEW STUDENT REGISTRATION IN 2009 TO 2014

Year	Amount (Y)	X	XY	X^2
2009	407	-5	-2035	25
2010	513	-3	-1539	9
2011	528	-1	-528	1
2012	374	1	374	1
2013	375	3	1125	9
2014	386	5	1930	25
Total	2583		-673	70

(Source: PMB STMIK BANJARBARU)

Based on Table 4, the values of a and b will be discovered. To find the values of a and b:

It is known, that:

$\Sigma Y = 2583$

$N = 6$

$\Sigma XY = -673$

$\Sigma X^2 = 70$

Then, to find the value of a:

$a = \Sigma Y / N$

a= 2583 / 6

a= 430.5

And to find the value of b:

$b= \Sigma XY/ \Sigma X^2$

b= -673/70

b= -9.6

After the values of a and b are obtained, then to find the equation of linear line:

$Y = a + b X$

$Y = 430.5 + (-9.6)$

With the calculated equation of linear line, the number of new student in 2015 can be calculated:

Y = 430.5 + (-9.6) X (For the year of 2015, the value of X is 7)

Thus:

$Y = 430.5 – (9.6 \times 7)$

$Y = 430.5 – 67.2$

$Y = 363$ (10)

It means that the numbers of new student candidates who register are 363 people.

From the calculation using Least Square formula (9) (10), the results showed that the prediction of the number of new student acquisition in 2015 for Odd Data is 313 people and for Even Data is 363 people. But the calculation result of the numbers of new student prediction can be damaged or fell due to several reasons, for example because of changes in government regulations, regression of high school graduates,or other reasons.

V. CONCLUSIONS

Based on the results of the study, it can be concluded that:

1) Prediction or forecasting analysis using Least Square method can be used to predict the number of new students acquisition for the coming period based on the data of the previous years, because the results are valid or closer to the truth.

2) From the results of calculations in predicting the acquisition of the number of new students, it was found that in 2012 there was 569, in 2013 there was 420 and in 2014 there was 378. It is determined that if the deviation between the fact and the calculation with Least Square Method is >50 people, then the result is invalid. Compared to the tangible result obtained in 2012, the deviation is 34.2% (195 people), meaning that the result is invalid. In 2013, the deviation is 10% (45 people), meaning that the result is valid. In 2014, the deviation is 2.07% (8 people), meaning that the result is valid. From the 3 comparisons, it is found that 2 results are valid and 1 result is invalid. This means that the formula of Least Square Method is valid or closer to the truth.

REFERENCES

[1] Basuki, A. (2014). Metode Least Square. taken from i http://basuki.lecturer.pens.ac.id/lecture/numerik5.pdf

[2] Beny Mulyandi, Y. I. (2010). Sales Forecasting Analysis of Fuel type Premium at the pump Heroes Bandung Asri. National Conference: Design And Application Of Technology .

[3] Cahyo Adi Basuki, I. A. (2008). Analysis of Fuel Consumption In Steam Power Plant.

[4] Geer, S. A. (2005). Least Squares Estimation. Encyclopedia of Statistics in Behavioral Science, Volume 2, pp. 1041–1045 .

[5] Handoko, T. H. (1994). Dasar – Dasar Manajemen Produksi dan Operasi. Yogyakarta: BPFE.

[6] HarperCollins. (1991, 1994, 1998, 2000, 2003). Dictionary / Thesaurus. Taken from The Free Dictionary By Farlex: http://www.thefreedictionary.com/Least-squares+method

[7] Heizer, J. d. (2009). Manajemen Operasi, Edisi 9.

[8] Kosasih, S. (2009). Manajemen Operasi - Bagian Pertama. Edisi 1. 74.

[9] Kristalina, P. (2015). Metode Least Square.

[10] Mia Savira, Nadya N.K. Moeliono, S.SOS, MBA. (2014). Sales Forecasting Analysis of generic drugs bearing (OGB) At PT. Indonesia Farma.

[11] Rambe, M. I. (2012). Perancangan Aplikasi Peramalan Persediaan Obat Obatan Menggunakan Metode Least Square. Pelita Informatika Budi Darma, Volume : VI, Nomor: 1, Maret 2014

[12] Subagyo, P. (1999). Forecasting (Konsep dan Aplikasi). Yogyakarta.

[13] Sahara, Afni. (2013). Sistem Peramalan Persediaan Unit Mobil Mitsubishi Pada PT. Sardana Indah Berlian Motor Dengan Menggunakan Metode xponential Smoothing. Informasi dan Teknologi Ilmiah(INTI), Volume : I, Nomor : 1, Oktober 2013

[14] Tanojo, E. (2007.). DeriVation Of Moving Least Squares Approximation Shape Functions And ITS Derivatives Using The Exponential Weight Function. Civil Engineering Dimension, Vol 9, No 1, 19-24, March 2007 .

[15] Widodo, J. (2008). Ramalan Penjualan Sepeda Motor Honda pada pada CV. Mitra Roda Lestari. (xii + 32 + Lampiran).

Compressed Sensing Based Encryption Approach for Tax Forms Data

Adrian Brezulianu
"Gheorghe Asachi" Technical University of Iasi
Iasi, Romania

Monica Fira
Romanian Academy
Institute of Computer Science
Iasi, Romania

Marius Daniel Peştină
"Gheorghe Asachi" Technical University of Iasi
Iasi, Romania

***Abstract*—In this work we investigate the possibility to use the measurement matrices from compressed sensing as secret key to encrypt / decrypt signals. Practical results and a comparison between BP (basis pursuit) and OMP (orthogonal matching pursuit) decryption algorithms are presented. To test our method, we used 10 text messages (10 different tax forms) and we generated 10 random matrices and for distortion validate we used the PRD (the percentage root-mean-square difference), its normalized version (PRDN) measures and NMSE (normalized mean square error). From the practical results we found that the time for BP algorithm is much higher than for OMP algorithm and the errors are smaller and should be noted that the OMP does not guarantee the convergence of the algorithm. We found that it is more advantageous, for tax forms (or other templates that show no interest for encryption) to encrypt only the recorded data. The time required for decoding is significantly lower than the decryption for the entire form**

Keywords—compressed sensing; encryption; security; greedy algorithms

I. INTRODUCTION

The theory of compressed sensing, perfected in the past few years by prestigious researchers such as D. Donoho [1], E. Candès [2], M. Elad [3], demonstrates the feasibility of recovering sparse signals from a number of linear measurements, dependent with the signal sparsity. Compressed sensing (CS) is a new method which draws the attention of many researchers and it is considered to have an enormous potential, with multiple implications and applications, in all fields of exact sciences [1-4]. Specifically, CS is a new technique for finding sparse solutions to underdetermined linear systems. In the signal processing domain, the compressed sensing technic is the process of acquiring and reconstructing a signal that is supposed to be sparse or compressible.

The perfect secrecy together with the secret communication is a well-defined field of research, being a difficult problem in the domain of information theory. One of the requirements for the information theoretic secrecy is to assure that a spy who listens a transmission containing messages will collect only small number of information bits from message. Additionally, it should provide protection against of an computationally unlimited adversary based on the statistical properties of a system. Shannon introduced the idea of perfect secrecy, in his fundamental paper [5].

An encryption idea by utilizing CS has been mentioned for the first time in [7], but not been addressed in detail [6]. In paper [8], the secrecy of CS is researched, and whose result is that CS can provide a computational guarantee of secrecy. In [9] examine the security and robustness of the CS-based encryption method. In paper [10], the authors describe a new coding scheme for secure image using the principles of compressed sensing (CS) and they analyze the secrecy of the scheme.

II. BACKGROUND

A. Compressed Sensing

Compressed sensing studies the possibility of reconstructing a signal x from a few linear projections, also called measurements, given the a priori information that the signal is sparse or compressible in some known basis Ψ.

To define sparsity precisely, we introduce the following notation: for Ψ - a matrix whose columns form an orthonormal basis, we define a K-sparse vector $x \in R^n$ as $x = \Psi\theta$, where $\theta \in R^N$ has K non-zero entries (i.e., is K-sparse) and Ω_K as the set of K indices over which the vector θ is non-zero.

The vectors on which x is projected onto are arranged as the rows of a nxN projection matrix Φ, n < N, where N is the size of x and n is the number of measurements. Denoting the measurement vector as y, the acquisition process can be described as:

$$y = \Phi x = \Phi\Psi\gamma \quad (1)$$

$$\hat{\gamma} = \arg\min_{\gamma} \|\gamma\|_{l_0} \quad subject\ to \quad y = \Phi\Psi\gamma \quad (2)$$

$$\hat{x} = \Psi\hat{\gamma} \quad (3)$$

The equations system (1) is obviously undetermined. Under certain assumptions on Φ and Ψ, however, the original expansion vector γ can be reconstructed as the unique solution to the optimization problem (2); the signal is then reconstructed with (3). Note that (2) amounts to finding the sparsest decomposition of the measurement vector y in the dictionary $\Phi\Psi$. Unfortunately, (2) is combinatorial and unstable when considering noise or approximately sparse signals.

For a K-sparse signal, only "K+1 projections of the signal onto the incoherent basis are required to reconstruct the signal with high probability"[5]. In this case, is necessary to use combinatorial search with huge complexity. In [1] and [2] is proposed tractable recovery procedures based on linear programming. In these papers is demonstrated that the tractable recovery procedures obtain the same results toward combinatorial search when for signal reconstruction are used aprox. 3 or 4 cK projections.

Two directions have emerged to circumvent these problems:

- Pursuit and thresholding algorithms seek a sub-optimal solution of (2)
- The Basis Pursuit algorithm [1] relaxes the l_0 minimization to, solving the convex optimization problem (4) instead of the original.

$$\hat{\gamma} = \arg\min_{\gamma} \|\gamma\|_{l_1} \quad subject\ to \quad y = \Phi\Psi\gamma \qquad (4)$$

The matrix Φ satisfies a restricted isometry property of order K whether there is a constant $\delta_K \in (0,1)$ such that the inequation (5),

$$(1-\delta_K)\|x\|_2^2 \le \|\Phi x\|_2^2 \le (1+\delta_K)\|x\|_2^2 \qquad (5)$$

holds for all x with sparsity K.

A. Notions of secrecy and Model

In cryptography, "a secret key system is an encryption system where both sender and receiver use the same key to encrypt and respectively, decrypt the message" [11-12].

A conventional encryption scheme consists of five elements [13-14]:

• *Plain text:* This is the original message or input information for the encryption algorithm.

• *Encryption algorithm*: This algorithm performs various substitutions and modifications to the clear text.

• *Secret Key:* This key is an input to the encryption algorithm.

• *Ciphertext:* The text resulting from encryption algorithm and it is depends on the plaintext and the secret key. Thus, for a given message, two different secret keys produce two different ciphertexts.

• *Decryption algorithm:* This algorithm is the inverse of the encryption algorithm. The decryption algorithm is applied with the same secret key to the ciphertext in order to get the original clear text.

Following two elements must be taken into account in order to achieve a secure encryption [15]:

1) The encryption algorithm should be very strong. If an attacker knows the encryption algorithm (encryption) and has access to one or more ciphertext, he cannot decrypt the ciphertext or find the secret key.

2) Both the transmitter and the receiver must obtain the secret key in a safe manner (on a secure communication channel) and to keep it secret.

Based on previous remarks, in Figure 1 (in the upper half) is shown the basic model for CS and it includes two major aspects: measurements taking and signal recovery. The measurements taking involve an encryption algorithm and signal recovery is associated with a decryption algorithm from the perspective of symmetric-key cipher. The relationship between CS and symmetric cryptography indicates that some possible cryptographic features can be embedded in CS.

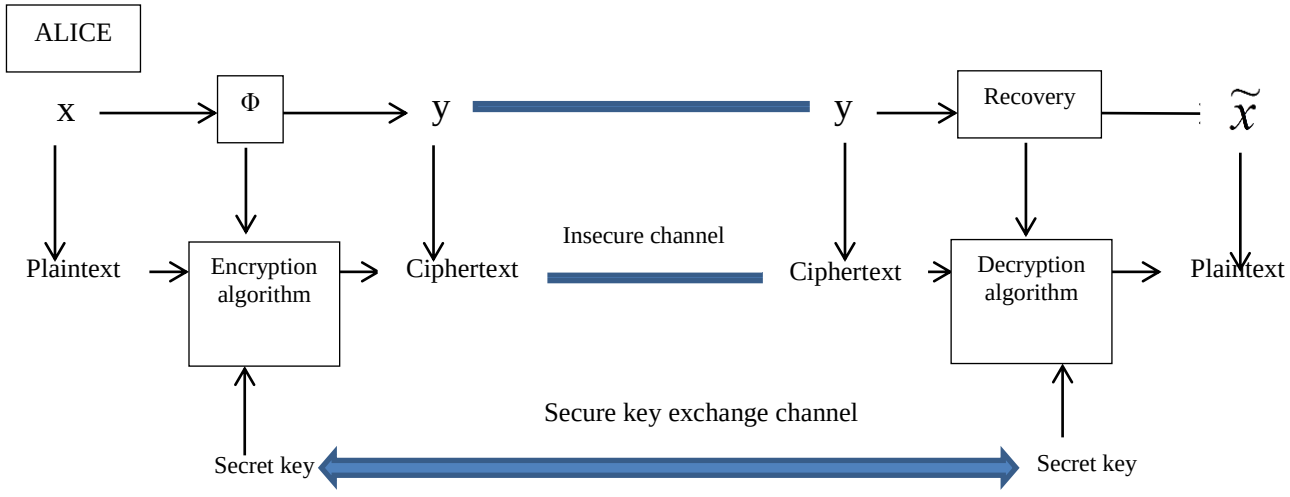


Fig. 1. The relationship between CS and symmetric-key cipher

The classical example of communication of a secret message from Alice to Bob assumes that Alice must use key from the set of keys. In this paper, let be i a key chosen by Alice with equal probability, and used to encrypt the message x with help of Φ_i matrix (via matrix multiplication operation). The result of multiplication is the cryptogram y which is

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

transmitted to Bob. The recipient knows the key used for encryption of the message. Knowing Φ_i and y, the compressed sensing literature provides conditions for x and Φ_i to allow the recovery of the original message x. The classical example of secret message communication assumes that the Alice's encrypted message y is being intercepted by an eavesdropper named Eve. For the third person, the used key the message encryption is unknown.

In our case, the measurement matrix Φ can be selected from a set of keys that is known for the transmitter (Alice) and the permitted receiver (Bob). Each random measurement matrix Φ is generated with a seed which can be exchanged through a secure approach between two desired sides [16-17].

A computational encryption scheme is secure if the ciphertext has one or two properties:

- The cost of breaking ciphertext is much higher than the encrypted information.
- The time needed for breaking ciphertext is longer than the lifetime of the information.

A brute force attack on the compressive sampling based encryption scheme would be guessing the linear measurement matrix Φ_i. Thus, an eavesdropper, e.g. Eve, could directly try to do this by performing an exhaustive search over a "grid" of values for Φ_i. But, the step size of this grid is critical because a too large step size may cause the search to miss the correct value and a too small grid size will increase the computational task unnecessarily.

The computational cost of signal reconstruction is high. For the best optimization algorithm (BP), the computational cost is in the order of $O(N^3)$ and a random search will make the search too expensive.

III. Simulations and Discussions

To test our method, we used 10 text messages (10 different tax forms) and we generated 10 random matrices.

To validate the decoding results, we evaluate the distortion between the original plaintext and the reconstructed plaintext by means of the PRD (the percentage root-mean-square difference), its normalized version (PRDN) measures and NMSE (normalized mean square error).

The percentage root-mean-square difference (PRD) measure defined as (6):

$$PRD\% = 100\sqrt{\frac{\sum_{n=1}^{N}(x(n)-\tilde{x}(n))^2}{\sum_{n=1}^{N}x^2(n)}} \quad (6)$$

is employed, where $x(n)$ is the original signal, $\tilde{x}(n)$ is the reconstructed signal, and N is the length of the window over which the PRD is calculated. The normalized version of PRD, PRDN, which does not depend on the signal mean value, $\bar{x}$, is defined as (7):

$$PRDN\% = 100\sqrt{\frac{\sum_{n=1}^{N}(x(n)-\tilde{x}(n))^2}{\sum_{n=1}^{N}(x(n)-\bar{x})^2}} \quad (7)$$

The normalized mean square error (NMSE) measure defined as (8):

$$NMSE = \frac{1}{\sigma^2} * \frac{\sum_{n=1}^{N}(x(n)-\tilde{x}(n))^2}{N} = \frac{MSE}{\sigma^2} \quad (8)$$

Where σ are the variance and MSE are mean square error measure.

Because our messages are text type, ie contain characters and numbers, we chosen to transform the messages in numerical signals based on the ASCII codes.

To use the identity matrix as decoding dictionary, the plaintext is necessary to be a sparse signal [18]. Because our messages had not this property, we have modified them by artificial insertion of zeros, thus obtaining sparse signals.

We used random matrix for encryption and for reconstruction we used two different algorithms, and namely,

- Basis pursuit algorithm (BP), known in the CS domain as the optimal algorithm in terms of errors [19-20] and
- Orthogonal matching pursuit algorithm (OMP) known in CS domain for its speed far superior to BP [21].

The orthogonal matching pursuit algorithm (OMP) is an iterative greedy algorithm. In this algorithm, at each step, the dictionary element which has the maximum correlation with the residual part of the signal is selected. The Basis Pursuit algorithm (BP) is a more sophisticated approach comparatively with OMP. In case of the BP algorithm, the initial sparse approximation problem is reduced to a linear programming problem.

Generically, the greedy algorithms (such OMP) have the disadvantage that there are not general guarantees of optimality. The basis pursuit algorithm, namely the convex relaxation algorithms, has the disadvantage of high computational complexity, translated into large computing time [22-26].

To synthesize ideas, we present the encryption and decryption necessary steps, namely:

- The message transformation into digital signal using extended ASCII code. This achieves a 1D digital signal.
- The segmentation of message or digital signal into segments of length 100.

- Transforming of the signals (signals with length 100) in sparse signals by inserting a predefined number of zeros. The position of the zeros is random from one segment to another.
- Encryption of sparse segments using a random matrix. Encryption is done by multiplying the signal sparse with a random matrix (Φ), resulting a lower dimension signal than initially sparse signal. The signal thus obtained is not sparse.
- Transmission of the message text is achieved by transmitting the encrypted signals (ciphertext) on an insecure line. It is important that random matrix (encryption matrix representing the secret key) is not sent with the ciphertext; it should be sent on a secure line. Another variant is use case when there is an agreement between the transmitter and receiver to generate random matrices in the same way, for example, using the same random number generator which is started from the same initial conditions.
- Decryption of the message will be achieved using a greedy algorithm (either orthogonal matching pursuit (OMP), or matching pursuit (MP), or greedy LS etc.) or convex relaxation algorithm (basis pursuit (BP)). For decryption, it is necessary to know the following: random matrix encryption Φ , the encrypted message (the ciphertext) Y, and the base for sparsity Ψ (in case of this paper, it is the identity matrix, due the fact that the message that was encrypted was a sparse signal).
- Because there is a decryption error which is very small, to return to the decrypted text, a decryption correction will be necessary. This correction consists in rounding of decrypted values to the nearest integer because the ASCII code is built from integers.

Figure 2 shows an example of plaintext and figure 3 presents a plot of the plaintext in ASCII format.

Anexa nr.1
DECLARATIE
privind veniturile realizate
Agentia Nationala de Administrare
200
Fiscala
din România
Anul Se completeaza cu X în cazul declaratiilor rectificative
A. DATE PRIVIND ACTIVITATEA DESFASURATA Cod CAEN cote forfetare de cheltuieli norma de venit Nr. Data 7. Data începerii activitatii 4. Obiectul principal de activitate 5. Sediul/Datele de identificare a bunului pentru care se cedeaza folosinta 8. Data încetarii activitatii asociere fara personalitate juridica entitati supuse regimului transparentei fiscale individual 6. Documentul de autorizare/Contractul de asociere/Închiriere/Arendare 3. Forma de organizare: 2. Determinarea venitului net: comerciale profesii libere drepturi de proprietate intelectuala cedarea folosintei bunurilor operatiuni de vânzare-cumparare de valuta la termen, pe baza de contract transferul titlurilor de valoare, altele decât partile sociale si valorile mobiliare în cazul societatilor închise activitati agricole 1. Categoria de venit Venituri: cedarea folosintei bunurilor calificata în categoria venituri din activitati independente sistem real modificarea modalitatii/formei de exercitare a activitatii

Fig. 2. The plaintext

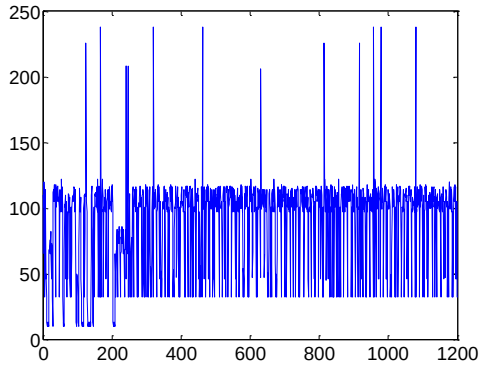


Fig. 3. The plaintext in ASCII format

We have chosen to split the signal into segments of length 100 and to insert a number of 800 by zeros for each ASCII codes segment. This means that each plaintext sequence with length 100 was transformed into a sequence with length 900. Figure 4 shows the plot of sparse plaintext.

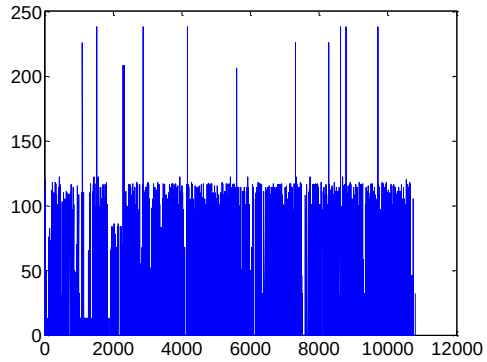


Fig. 4. The sparse plaintext

We used for encryption a random matrix of size 500x900. This random matrix represents the secret key. Figure 5 show the ciphertext obtained a random matrix for encryption. Note that the ciphertext contains positive and negative numbers and it has a different length than the plaintext.

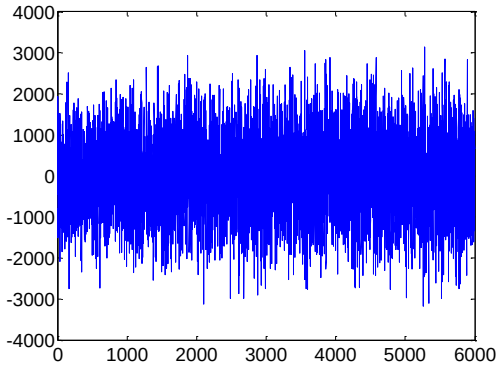


Fig. 5. The ciphertext

To decode the ciphertext we tested two known algorithms from compressed sensing domain, namely, orthogonal

matching pursuit algorithm (OMP) and basis pursuit algorithm (BP).

OMP is an iterative greedy algorithm and selects at each step the column of Φ matrix which has the maximum correlation with the current residuals. A set of iteratively selected columns is built. The residuals are iteratively updated by projecting the observation y onto the subspace spanned by the previously selected columns. This algorithm has simpler and faster implementation toward similar methods.

The Basis Pursuit (BP) algorithm consists in finding a least L1 norm solution of the underdetermined linear system $y = \Phi x = \Phi \Psi \gamma$.

The both methods can be guaranteed to have bounded approximation solution of sparse coefficients estimation for the condition that the L0 norm of sparse coefficients is smaller than a constant decided by the dictionary [1].

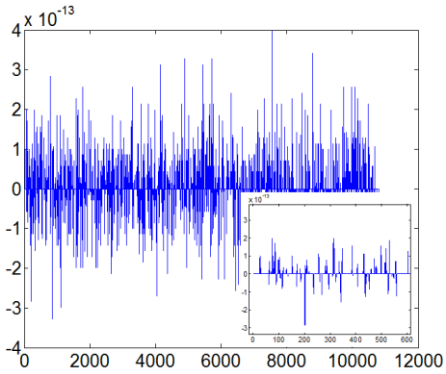


Fig. 6. Error for decoding with OMP, before decoding correction. In the bottom right corner there is a zoom for the first 600 samples

Figure 6 and figure 7 show errors for decoding with OMP, respectively BP algorithms.

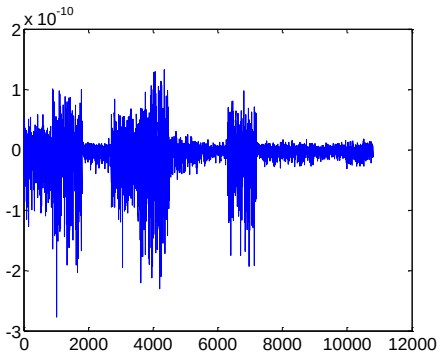


Fig. 7. Error for decoding with BP, before decoding correction

For the OMP based decoding, where the original signal (plaintext) was sparse (had null values), null values were obtained after decoding. In case of BP decoding, the algorithm approximates all values and it failed to return null values for the null values from plaintext, but it returned values very close to zero.

Because in the case of typical tax forms often it is required to encrypt only registration data and because the decryption time is higher for the completed form (data + template), we tested the proposed algorithm for encrypting data alone. In Figure 8 is an example of data belonging to the form shown in Figure 2.

12656
03.08.2010
Minerit
Str. Minei, Nr. 23, Hunedoara
02.08.2000
contract nr. 05/09.10.1999
SRL

Fig. 8. The plaintext with registration data from tax form

For a signal of dimension m with assumed sparsity s<<m, and a dictionary of N>>m atoms, computational costs for pursuits using general and fast dictionaries are:

$$complexity \quad for \quad OMP = smN(sN \log N + s^3))$$

where m stands for measurements, s stands for sparsity.

The popular basis pursuit algorithm (BP) has computational complexity

$$complexity \quad for \quad BP = O(N^3)$$

Alternatives to BP (e.g., greedy matching pursuit) also have computational complexities that depend on N.

Table 1 presents average results for 10 text messages and for 10 datasets from tax forms. The time for BP algorithm is much higher than for OMP algorithm and the errors are smaller.

TABLE I. AVERAGE RESULTS

Decoding algorithm	Time (seconds)	Error (PRD, PRDN, NMSE)
average results for 10 text messages, each with 1200 char		
Basis pursuit algorithm (BP)	867.40	PRD = 7.7521e-011 PRDN = 8.1579e-011 NMSE = 7.1476e-028
Orthogonal matching pursuit algorithm (OMP)	2.61	PRD = 1.0139e-013 PRDN = 1.0670e-013 NMSE = 1.2227e-033
average results for 10 registration data text messages, each with 103 char		
Basis pursuit algorithm (BP)	42.27	PRD = 3.9552e-011 PRDN = 4.1392e-011 NMSE = 3.2769e-028
Orthogonal matching pursuit algorithm (OMP)	0.09	PRD = 1.0979e-013 PRDN = 1.1489e-013 NMSE = 2.5248e-033

It should be noted that the OMP does not guarantee the convergence of the algorithm and for a smaller number of measurements; the results can be much worse for OMP comparatively with BP. Results depend on the number of measurements and on used decoding algorithm [24-26].

IV. CONCLUSIONS

In this paper, the perfect secrecy via compressed sensing was studied and discussed. We presented an analysis with practical results for tax forms as plaintexts. For decoding we used BP and OMP algorithms, and we presented a comparative analysis. The time for BP algorithm is much higher than for

OMP algorithm and the errors are smaller and should be noted that the OMP does not guarantee the convergence of the algorithm. According to average results from Table 1, it is more advantageous, for tax forms (or other templates that show no interest for encryption) to encrypt only the recorded data. The time required for decoding is significantly lower than the decryption for the entire form.

ACKNOWLEDGMENT

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-II-RU-TE-2014-4-0832 "Medical signal processing methods based on compressed sensing; applications and their implementation."

REFERENCES

[1] D. Donoho, "Compressed sensing", IEEE Transactions on Information Theory, vol. 52(4), pp. 1289–1306, 2006

[2] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information" IEEE Transactions on Information Theory, vol. 52(2), pp. 489–509, 2006.

[3] M. Elad, "Optimized Projections for Compressed Sensing", IEEE Transactions on Signal Processing, Vol. 52, 2007

[4] Shuhui Bu, Zhenbao Liu, Tsuyoshi Shiina, Kazuhiko Fukutani, *Matrix Compression and Compressed Sensing Reconstruction for Photoacoustic Tomography*, Elektronika ir elektrotechnika, Vol 18, No 9 (2012)

[5] C. E. Shannon, "Communication theory of secrecy systems" Bell System Technical Journal, vol. 28(4), pp. 656–715, October 1949.

[6] E. Candes, J. Romberg, "Sparsity and incoherence in compressive sampling" Inverse Problems, Vol. 23, pp. 969–985, 2007.

[7] D. Takhar, J. N. Laska, M. B. Wakin, M. F. Duarte, D. B. S. Sarvotham, K. F.Kelly, and R. G. Baraniuk, "A new camera architecture based on optical-domain compression" in Proc. IST/SPIE Symposium on Electronic Imaging: Computational Imaging, vol. 6065, 2006, pp. 129-132

[8] Y. Rachlin, D. Baron, "The Secrecy of Compressed Sensing Measurements", 46th Annual Allerton Conference on Communication, Control, and Computing, 2008

[9] A. Orsdemir, H. Oktay Altun, G. Sharma, Mark F. Bocko,"On the Security and Robustness of Encryption via Compressed Sensing" Military Communications Conference, 2008, Milcom 2008, IEEE pp.1-7

[10] G. Zhang, S. Jiao, X. Xu, "Application of Compressed Sensing for Secure Image Coding", Lecture Notes in Computer Science Volume 6221, 2010, pp 220-224.

[11] Gary C. Kessler, An Overview of Cryptography, 2015, http://www.garykessler.net/library/crypto.html

[12] Sattar B. Sadkhan Al Maliky and Nidaa A. Abbas, Multidisciplinary Perspectives in Cryptology and Information Security, IGI Global, 2014

[13] W. Stallings, Cryptography and Network Security (4th Edition), pp. 30, Prentice Hall, 2005

[14] V. Preoteasa, Cryptography and Network Security, Lecture 2: Classical Encryption Techniques, Spring 2008, Abo Akademi University

[15] D.R. Stinson, Cryptography: Theory and Practice, 2nd edition, Chapman & Hall/CRC, 2002

[16] W. Diffie, M. E. Hellman, "New directions in cryptography" IEEE Transactions on Inform. Theory, vol. IT-22, no. 6, pp. 644–654, 1976.

[17] U. Maurer, S. Wolf, "Information-theoretic key agreement: From weak to strong secrecy for free" Advances in Cryptology—EUROCRYPT, Lecture Notes in Computer Science, 2000.

[18] J. Bowley, L. Rebollo - Neira, "Sparsity and "something else": an approach to encrypted image folding", IEEE signal processing letters, 18 (3), pp. 189-192., 2011

[19] D. Donoho, Y. Tsaig, "Fast solution of L1-norm minimization problems when the solution may be sparse," Stanford University Department of Statistics Technical Report, 2006.

[20] D. Donoho, "For most large underdetermined systems of linear equations, the minimal L1 norm solution is also the sparsest solution" Communications on Pure and Applied Mathematics, Vol. 59, pp. 797–829, June 2006.

[21] J. Tropp, A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit" IEEE Trans. on Information Theory, Vol. 53, No. 12, pp. 4655–4666, December 2007.

[22] E. Candes, T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?", IEEE Transactions on Information Theory, Vol. 52, No. 12, pp. 5406–5425, December 2006.

[23] M. J. Wainwright, "Sharp thresholds for noisy and high-dimensional recovery of sparsity using L1-constrained quadratic programming (Lasso)", IEEE Transactions on Information Theory, 2009.

[24] T.T. Cai, L. Wang, "Orthogonal Matching Pursuit for Sparse Signal Recovery with Noise", IEEE Transactions on Information Theory, vol. 57, 7, 4680–4688, 2011.

[25] G. Davis, S. Mallat, M. Avellaneda, "Greedy adaptive approximation", J. Constr. Approx., 13:57-98, 1997.

[26] J.A. Tropp, "Greed is good: Algorithmic results for sparse approximation", IEEE Transactions on Information Theory, 50, pp. 2231–2242, 2004.

[27] M. N. Chavhan, S.O.Rajankar, "Study the Effects of Encryption on Compressive Sensed Data", International Journal of Engineering and Advanced Technology, Volume 2, Issue 5, pp. 179 – 182, 2013