Spectral Analysis Reviews, 2014, 2, 19-25
Published Online October 2014 in SciRes. http://www.scirp.org/journal/sar
http://dx.doi.org/10.4236/sar.2014.24025

On-Line Measurement of the Chemical Oxygen Demand in Wastewater in a Pulp and Paper Mill Using Near Infrared Spectroscopy

John Dahlbacka[1], Josefina Nyström[2], Torgny Mossing[2], Paul Geladi[2], Tom Lillhonga[1]

[1]Unit for Research and Development, Novia University of Applied Sciences, Vaasa, Finland
[2]Department of Forest Biomaterials and Technology, Swedish University of Agricultural Sciences, Umeå, Sweden
Email: John.Dahlbacka@novia.fi

Received 10 August 2014; revised 6 September 2014; accepted 8 October 2014

Copyright © 2014 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Although near infrared (NIR) spectroscopy has been evaluated for numerous applications, the number of actual on-line or even on-site industrial applications seems to be very limited. In the present paper, the attempts to produce on-line predictions of the chemical oxygen demand (COD) in wastewater from a pulp and paper mill using NIR spectroscopy are described. The task was perceived as very challenging, but with a root mean square error of prediction of 149 mg/l, roughly corresponding to 1/10 of the studied concentration interval, this attempt was deemed as successful. This result was obtained by using partial least squares model regression, interpolated reference values for calibration purposes, and by evenly distributing the calibration data in the concentration space. This work may also represent the first industrial application of on-line COD measurements in wastewater using NIR spectroscopy.

Keywords

Chemical Oxygen Demand (COD), Industrial Wastewater, Near Infrared Spectroscopy (NIR), Multivariate Calibration, On-Line Measurement

1. Introduction

Wastewater flows are characterized by constantly changing flow rates and composition [1]. Industrial production of pulp and paper generates considerable amounts of wastewater, where the contaminants may be characte-

How to cite this paper: Dahlbacka, J., *et al*. (2014) On-Line Measurement of the Chemical Oxygen Demand in Wastewater in a Pulp and Paper Mill Using Near Infrared Spectroscopy. *Spectral Analysis Reviews*, **2**, 19-25.
http://dx.doi.org/10.4236/sar.2014.24025

rized using for instance the chemical oxygen demand (COD) [2]. Activated sludge treatment of this type of wastewater from chemical pulp mills can reduce the COD by 25% - 65%, however addition of nitrogen and phosphorous may be needed in order to avoid that these elements limit the biological degradation of organic compounds [3]. In order to perform this addition efficiently, on-line information about the organic load in the activated sludge stage is essential.

One potential method of gaining on-line information of the organic load is to use near infrared (NIR) spectroscopy combined with quantitative models based on multivariate methods. Although NIR spectroscopy can be seen as an extremely powerful tool for industrial quality control and process monitoring [4], only a limited number of publications describing the use of NIR spectroscopy in wastewater applications are readily found. It is probably fair to say that there is no clear trend or typical wastewater application of NIR spectroscopy. The studies found represent quite diverse applications, for instance, quantitative measurements of oil, urea and solids [5], glycerol [6], and methanol and glycerol simultaneously [7] in wastewater from a biodiesel fuel production plant. As another example, in one of the few *in-situ* applications found [8], principal component analysis (PCA) was used instead of quantitative modelling to monitor an activated sludge plant.

However, in most treatment plants, COD is probably the most important measure, which is also reflected in the number of publications that describe the use of NIR spectroscopy for COD measurements. These include off-line measurements of only COD [9]-[11], as well as COD in combination with other parameters off-line [12] [13], and *in-situ* [14] [15]. It is an interesting observation that although the reported accuracies vary significantly in these publications, the measurement error divided by the concentration interval studied is in many cases close to 1/10. However, [15] reports on a relative error above 50% for the NIR COD measurement. This present study then complements the work already done by others by presenting a fully automated on-line and on-site measurement of COD in industrial wastewater, evaluated during an undisrupted measurement period of one month.

2. Materials and Methods

The spectra were collected with a Red Eye$^{®}$ Online sensor for suspensions and fluids (Pulp Eye AB, P. O. Box 70, 89,122 Örnsköldsvik, Sweden). The sampling system (also constructed by Pulp Eye AB), or measurement head, consisted of a filter unit followed by a flow through cell coupled with optical fibres and equipped with an automated back flush system using tap water and activated in between every measurement. The sampling system was mounted on a bypass loop of the main pipe. For every third spectrum, a new reference spectrum was collected of the tap water. Each spectrum consisted of 50 averaged scans, and the path length of the flow through cell was 1 mm. The spectra were made up by 256 wavelengths recorded between 1018 and 2032 nm at an average data resolution of 4 nm. The spectra were collected on-line at 10 minutes intervals for a period of 29 days.

During weekdays, laboratory COD reference measurements were generally performed twice a day on-site. In total 4099 spectra were collected and for 36 of these COD reference measurements were performed. The calibration models were calculated using the PLS Toolbox v. 7.0.1 (Eigenvector Research, Inc. 3905 West Eagle rock Drive Wenatchee, WA 98801, USA), together with MATLAB R2011b (The Math Works AB, Kista, Sweden) where all matrix calculations were also performed. The calibration methods used were partial least squares (PLS) regression and principal component regression (PCR). The performance of the models was assessed by, among other things, the root mean square error of calibration (*RMSEC*), the root mean square error of cross validation (*RMSECV*), and the root mean square error of prediction (*RMSEP*). A description of the regression methods and the definition of the performance parameters can be found in [16]. The main units of the wastewater treatment facility were a pre-sedimentation basin, an aerated activated sludge basin, and a post-sedimentation basin. The measurements were made on the wastewater leaving the pre-sedimentation basin.

3. Results and Discussion

Initially the measurement was performed with a transflectance probe mounted in the bypass loop. However, the probe was clogged within hours and therefore replaced with the flow through cell with an automated back flush system. This reduced the problems with clogging and fouling significantly, but at the same time reduced the potential information about suspended solids to a minimum. In order to carry out the investigation with a minimal intrusion on the daily activities in the facility, it was also decided that any quantitative calibration will have to rely on reference data obtained from the measurements routineously carried out by the plant operators. As will be discussed, this trivial experimental design posed some limitations when trying to create an accurate calibration model. It was therefore clear from the beginning that it would be a challenging task to establish a reliable calibration model for the intended application. With a PCR model built on the 36 spectra corresponding to the reference measurements and using mean centering as spectral pre-processing, 89% of the spectral variance and only 0.04% of the COD variance were explained by the first PC. The corresponding numbers for the second PC were 11% and 3.7%. Thus, 2 PCs explained essentially all of the spectral variance and almost nothing of the COD variance. The situation was very much the same in the PLS space, with spectral/COD variance explained by the first PLS component at 46/2.7% and 54/1.5% by the second. In other words, the relationship between spectral and COD variance was very weak in this data. One reason for this was that the spectra from the end of the time series displayed extreme absorbance values, apparently due to fouling of the windows of the flow through cell.

New PCR and PLS models were therefore built on the 22 first spectra corresponding to reference measurements. In this case a second order derivative (Savitzky-Golay) based on a 9 point third order polynomial was applied to minimize baseline effects. This spectral pre-treatment was followed by auto scaling instead of mean centering in an effort to enhance minor variance in the spectra potentially relatable to the COD concentration. This pre-processing was also used in all later models. The two models are summarized in **Table 1**. With this reduced data set, the PLS model explained a very reasonable amount of the COD variance (76%) when using 7 PLS components. Although both PCR and PLS were now able to explain significant amounts of the COD variance, it was clear that modelling should be performed using PLS rather than PCR. However, **Table 1** also shows that the coefficient of determination is still close to 0 for both PCR and PLS in cross validation. Predictions made with this PLS model on the full data set are shown in **Figure 1**.

According to **Figure 1**, the PLS model is capable of predicting only a noisy average COD concentration. The visually noticeable increase in the noise level after approximately 2600 spectra was assumed to be a result of fouling of the flow through cell windows. Therefore only the first 2644 spectra were considered in the later calculations and modelling. This cut resulted in 24 spectra with corresponding reference values. **Figure 2** shows every twentieth of the first 100 spectra collected. Based on this figure the wavelengths above 1840 nm were discharged. The remaining spectral information reveals very little of obvious interest by visual assessment. A new PLS model was regressed on the 24 spectra corresponding to reference values. This gave an *RMSEC* of 65 mg/l, an *RMSECV* of 222 mg/l, and a coefficient of determination in cross validation of 0.21. Thus some im-

Table 1. PCR and PLS models regressed on the first 22 spectra corresponding to reference measurements in the time series.

	PCR		PLS	
RMSEC [mg/l]	176		122	
RMSECV [mg/l]	276		280	
R^2 (calibration)	0.50		0.76	
R^2 (cross validation)	0.13		0.19	
	Cumulative variance by component #			
	PCR		PLS	
	Spectral	COD	Spectral	COD
1	96.24	15.80	96.18	16.55
2	99.05	26.38	98.97	30.63
3	99.70	33.68	99.67	38.50
4	99.86	37.12	99.83	46.10
5	99.96	43.78	99.94	50.23
6	99.97	44.68	99.97	66.17
7	99.98	49.60	99.98	75.83

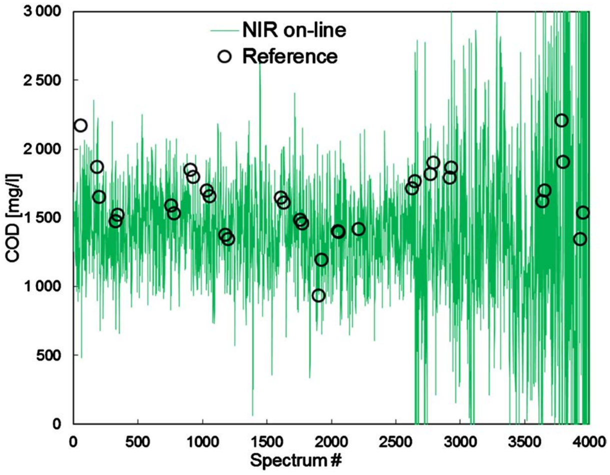


Figure 1. Predictions on the full data set using a PLS model based on the first 22 spectra corresponding to reference measurements.

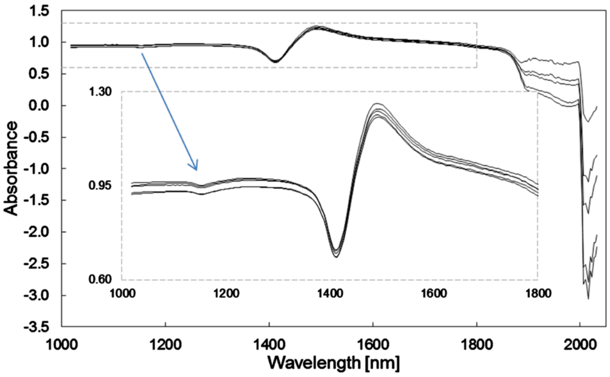


Figure 2. Absorbance spectra from the beginning of the online collected time series. The higher wavelengths (above 1840 nm) were deemed useless and only very broad bands are visible in the rest of the spectra.

provements compared to the PLS model previously accounted for was obtained, but the model was still only able to fit the regression data and showed no capacity for cross validation.

Due to the need for additional reference data, "synthetic" reference values were assigned to all the 2644 spectra by means of interpolation between the actual reference values. Here linear interpolation was used based on the simple fact that no information on the behavior of the COD concentration between the reference measurement points was available. A new PLS model based on all the 2644 spectra was thereafter regressed. For this model, an RMSEC of 75 mg/l, an RMSECV of 77 mg/l, and a coefficient of determination in cross validation of 0.86 were obtained by using 10 PLS components. It should be noted that the performance of the model is related to mainly the interpolated reference values, and should therefore be interpreted with some caution. **Figure 3** shows the predictions with this model as a time series, and it can be suggested that the measurement seem reasonable at the same time as the noise level is still considerable. These results were seen as promising, although cross validation using venetian blinds split on this type of data was assumed to produce overoptimistic validation results. The next step was therefore to split the available data into a model regression set and a fully external validation data set.

Since the use of interpolated reference values resulted in an abundance of data for regression and validation, the data set was simply split in half using the first 1322 spectra for model regression and the remaining spectra for model validation. This gave a model that in calibration and cross validation performed very similarly to the model described above. However, on the external validation data set the RMSEP was as high as 168 mg/l and

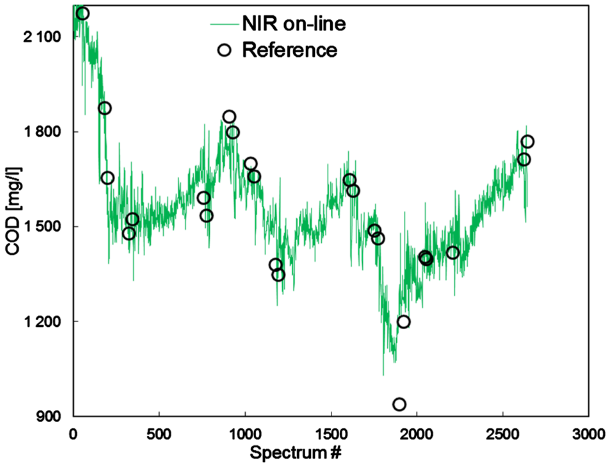


Figure 3. Predictions on the first 2644 spectra with a PLS model based on the same data set using interpolated COD values as reference values.

the coefficient of determination only 0.36. Switching the regression set for the validation set and vice versa did not improve the validation performance. One reason for this could have been the gradual fouling of the windows of the flow through cell. However, what is evident from **Figure 3** is that it is impossible to select a single continuous time series for regression and validation and at the same time cover the same COD spans in both data sets. A regression data set was therefore extracted from the 2644 spectra studied by selecting the first 322 spectra and spectrum 1897 to 2644 as regression data. In **Figure 3** this corresponds to the range covering the first 4 reference measurement points and the range from the lowest reference point to the end of the time series.

This new split gave a model with an RMSEP of 144 mg/l and a coefficient of determination of 0.43. Considering the fact that these performance parameters were based on the interpolated reference values, no further attempt was made to optimize these. Instead the focus was set on finding a model that could predict the high and low COD concentrations well, rather than being accurate around the average concentration. This was attempted by reweighting the information in the model regression data set. The data was split into 20 concentration intervals (matlab: hist). In this split, 5 intervals contained 125 spectra or more, and 6 intervals contained 12 or less. The data was thereafter reweighted by reducing the maximal number of spectra in each interval to 20. This was done by generating a random sequence of indices to remove within each concentration interval (matlab: randperm). In this way the number of spectra in the calibration data set was reduced from 1070 to 334.

After reweighting the regression data the data set was further refined by a stepwise removal of spectra with high absolute cross validation residuals (a model was regressed, high residual samples removed, and a new model regressed, etc.). This further reduced the calibration data to 274 spectra. For this model an RMSEP of 182 mg/l and a coefficient of determination of 0.35 were obtained. Based on these parameters, the reduction of the calibration data set apparently deteriorated the model performance. However, these values were obtained for the interpolated reference values instead of real reference values, and the objective was to obtain a model that predicted changes rather than average concentrations. To evaluate how this objective was met the standard deviation of the predictions of the validation data before and after the reduction of the calibration data set was computed. The standard deviation for the predictions with the model regressed on the original regression data set was determined to 176 mg/l and the corresponding value for the reduced data set was 224 mg/l. Thus, according to this somewhat unconventional performance parameter, the reduction or refinement of the calibration data set resulted in an improved model.

The predictions by this last model of the regression and validation data are shown in **Figure 4**. The predictions of the validation data were perceived as reasonable, although still impaired by a relatively high noise level. However, since this was a time series measurement, noise reduction is not confined to the spectral level only. Filtering the model predictions is also a straightforward method. **Figure 4** therefore also contains filtered predictions, obtained from a second order digital Butterworth filter as described in [17]. The ratio of the cut-off frequency to the sampling frequency was determined by minimizing the square sum of errors between filtered predictions and (real) reference values in the second half of the regression data. According to **Figure 4** it is evident

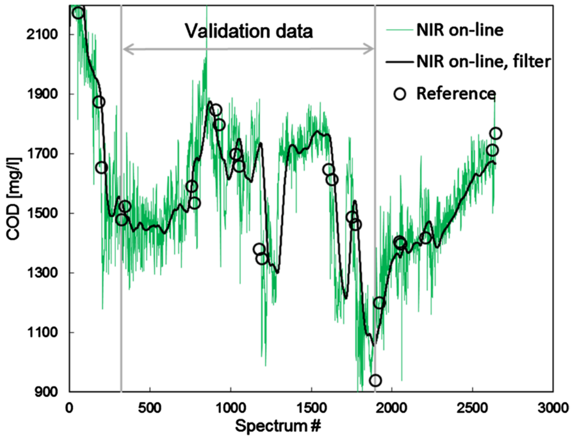


Figure 4. Predictions of the first 2644 spectra with a PLS model based on a separate model regression data set, extracted from the beginning and the end of the time series, and reduced by equalling the spread in the concentration and removing the spectra with high cross validation residuals. A second order Butterworth digital filter was used to obtain the filtered predictions.

that this filter reduced the noise level very significantly, but at the same time a phase shift was introduced. Whether or not this phase shift is of importance can be debated, but on the validation data, and computed on spectra corresponding to actual reference measurements, an RMSEP of 201 mg/l and an coefficient of determination of 0.35 were obtained for the raw model predictions and the corresponding values after the filtering were 149 mg/l and 0.65 respectively. Based on **Figure 4** it could also perhaps be argued that a large portion of this RMSEP could be contributed to time shifts between the reference and the on-line measurements.

4. Conclusion

The starting point for this attempt to create a quantitative model for the COD concentration in wastewater from a pulp and paper mill was basically a data set of 36 spectra and their corresponding reference measurements. On this data, essentially no relation between the COD concentration and the spectral features could be established, at least when only mean centering was used as spectral pre-processing. By using more advanced pre-processing options and removing the highest wavelengths, a relation could be modelled within the calibration data, but cross validation results were still not very promising. However, by increasing the amount of calibration data available by means of interpolated reference values, also the cross validation results started to look promising. The use of interpolated reference values in calibration, in combination with reweighting and refining the calibration data set, resulted in a model with very reasonable validation results. By further adding a filter to the predictions, a very appealing time series behavior was obtained. Unfortunately, if this behavior depicts the true changes in the COD concentration, much more frequent samplings would have been necessary in order to fully validate this. However, since this was an industrial installation and not a study performed in a laboratory, obtaining additional measurements is very difficult. On the other hand, the validation was still made against 14 reference measurements and this should not be an alarmingly low number.

Acknowledgements

This work was made possible through the Mare Purum project, funded by the European territorial cooperation programme Botnia-Atlantica. The authors also want to thank Hans Thorén at Svenska Cellulosa Aktiebolaget SCA, Elias Sundvall, Öjvind Sundvall, and Thomas Storsjö at PulpEye AB, as well as Anders Jonsson and Robin Norman at ProcessIT Innovations for this fruitful cooperation.

References

[1] Henze M., Harremoës P., LaCour Jansen J. and Arvin E. (2002) Wastewater Treatment: Biological and Chemical

Processes. Springer, Berlin.

[2] Pokhrel, D. and Viraraghavan, T. (2004) Treatment of Pulp and Paper Mill Wastewater—A Review. *Science of the Total Environment*, **333**, 37-58. http://dx.doi.org/10.1016/j.scitotenv.2004.05.017

[3] Diez, M.C., Castillo, G., Aguilar, L., Vidal, G. and Mora, M.L. (2002) Operational Factors and Nutrient Effects on Activated Sludge Treatment of *Pinus radiata* Kraft Mill Wastewater. *Bioresource Technology*, **83**, 131-138. http://dx.doi.org/10.1016/S0960-8524(01)00204-8

[4] Siesler, H.W., Ozaki, Y., Kawata, S. and Heise, H.M. (2007) Frontmatter. In: Siesler, H.W., Ozaki, Y., Kawata, S. and Heise, H.M., Eds., *Near-Infrared Spectroscopy*: *Principles*, *Instruments*, *Applications*. Wiley-VCH Verlag GmbH, Weinheim, Germany. http://dx.doi.org/10.1002/9783527612666.fmatter

[5] Suehara, K.I., Owari, K., Kohda, J., Nakano, Y. and Yano, T. (2007) Rapid and Simple Determination of Oil and Urea Concentrations and Solids Content to Monitor Biodegradation Conditions of Wastewater Discharged from a Biodiesel Fuel Production Plant. *Journal of Near Infrared Spectroscopy*, **15**, 89-96. http://dx.doi.org/10.1255/jnirs.721

[6] Kohda, J., Ooshita, K., Nakano, Y. and Yano, T. (2008) Measurement of the Glycerol Concentration during the Microbial Treatment of the Wastewater from the Biodiesel Fuel Production Plant Using Near Infrared Spectroscopy. *Journal of Near Infrared Spectroscopy*, **16**, 199-204. http://dx.doi.org/10.1255/jnirs.804

[7] Kawai, S., Kohda, J., Nakano, Y. and Yano, T. (2009) Predicting Methanol and Glycerol Concentrations in Microbial Treated Wastewater Discharged from a Biodiesel Fuel Production Process Using Near Infrared Spectroscopy. *Journal of Near Infrared Spectroscopy*, **17**, 51-58. http://dx.doi.org/10.1255/jnirs.825

[8] Dias, A.M.A., Moita, I., Páscoa, R., Alves, M.M., Lopes, J.A. and Ferreira, E.C. (2008) Activated Sludge Process Monitoring through *in Situ* Near-Infrared Spectral Analysis. *Water Science and Technology*, **57**, 1643-1650. http://dx.doi.org/10.2166/wst.2008.147

[9] Pan, T., Chen, W.W., Chen, Z.H. and Xie, J. (2011) Waveband Selection for NIR Spectroscopy Analysis of Wastewater COD. *Key Engineering Materials*, **480**, 393-396. http://dx.doi.org/10.4028/www.scientific.net/KEM.480-481.393

[10] Pan, T. and Chen, Z.H. (2012) Rapid Determination of Chemical Oxygen Demand in Sugar Refinery Wastewater by Short-Wave Near-Infrared Spectroscopy. *Advanced Materials Research*, **549**, 167-171. http://dx.doi.org/10.4028/www.scientific.net/AMR.549.167

[11] Pan, T., Chen, W.W., Huang, W.J. and Qu, R.T. (2012) Model Optimization for Near-Infrared Spectroscopy Analysis of Chemical Oxygen Demand of Wastewater. *Key Engineering Materials*, **500**, 832-837. http://dx.doi.org/10.4028/www.scientific.net/KEM.500.832

[12] Melendez-Pastor, I., Almendro-Candel, M.B., Navarro-Pedreño, J., Gómez, I., Lillo, M.G. and Hernández, E.I. (2013) Monitoring Urban Wastewaters' Characteristics by Visible and Short Wave Near-Infrared Spectroscopy. *Water*, **5**, 2026-2036. http://dx.doi.org/10.3390/w5042026

[13] Yang, Q., Liu, Z. and Yang, J. (2009) Simultaneous Determination of Chemical Oxygen Demand (COD) and Biological Oxygen Demand (BOD_5) in Wastewater by Near-Infrared Spectrometry. *Journal of Water Resource and Protection*, **4**, 286-289. http://dx.doi.org/10.4236/jwarp.2009.14035

[14] Páscoa, R.N., Lopes, J.A. and Lima, J.L. (2008) *In Situ* Near Infrared Monitoring of Activated Dairy Sludge Wastewater Treatment Processes. *Journal of Near Infrared Spectroscopy*, **16**, 409-419. http://dx.doi.org/10.1255/jnirs.803

[15] Sarraguça, M.C., Paulo, A., Alves, M.M., Dias, A.M., Lopes, J.A. and Ferreira, E.C. (2009) Quantitative Monitoring of an Activated Sludge Reactor Using On-Line UV-Visible and Near-Infrared Spectroscopy. *Analytical and Bioanalytical Chemistry*, **395**, 1159-1166. http://dx.doi.org/10.1007/s00216-009-3042-z

[16] Næs, T., Isaksson, T., Fearn, T. and Davies, T. (2002) A User Friendly Guide to Multivariate Calibration and Classification. NIR Publications, Chichester, UK.

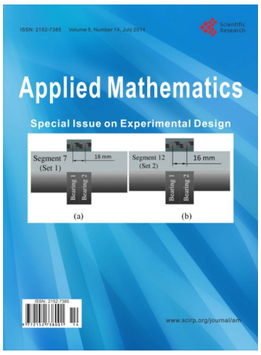
[17] Winter, D.A. (2009) Biomechanics and Motor Control of Human Movement. John Wiley & Sons, Hoboken, New Jersey. http://dx.doi.org/10.1002/9780470549148

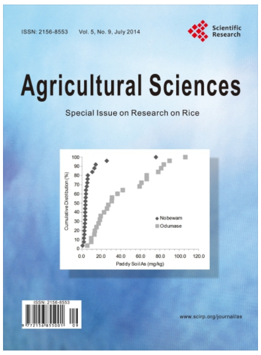
Scientific Research

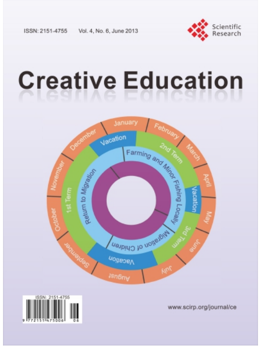
www.scirp.org

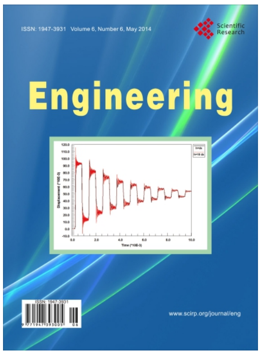
Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

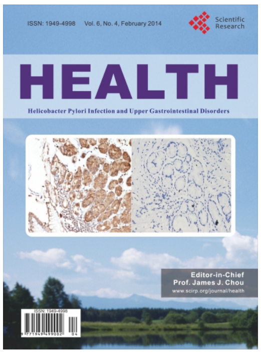
Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.










HEALTH
Helicobacter Pylori Infection and Upper Gastrointestinal Disorders
Editor-in-Chief
Prof. James J. Chou

Journal of Cancer Therapy
Special Issue on Bone Tumor Research