A Causal Analysis by Structural Equation Modeling of Sleep Monitoring Sensor Data

Nobuhiro Oishi, Naoki Yamamoto, Akio Ishida, and Jun Murakami
National Institute of Technology, 2659-2 Koshi-shi, Kumamoto, Japan
Email: oishi@kumamoto-nct.ac.jp

***Abstract*—In this paper, Structural Equation Modeling (SEM) is used to analyze the causal relationship between the level of sleep and the environmental data. The data used for the analysis was obtained by a care support device used in an elderly care facility. By applying the stepwise selection method to this data, we were able to find four observation variables that affect the level of sleep. And the latent variables are determined by scree plot. We proposed a causal model in which four observed variables and two latent variables affect the level of sleep. Statistical analysis environment R and the lavaan package were used for SEM analysis in this paper. From this model, it was found that the indoor environment and the vital signs affect sleep, and that heart rate should be reduced to obtain deep sleep.**

***Index Terms*—sleep monitoring sensor, causal analysis, structural equation modeling, sleep level, R language**

I. Introduction

In recent years, we have entered an era of big data that can easily acquire large amounts of data [1]. Available data includes, for example, POS data with ID, online shopping purchase data, website browsing statistics, SNS transmission statistics, physical quantity data, position information, biological information by various sensors mounted on autonomous cars and wearable devices, various information obtained from fixed camera images, and so on. By statistically analyzing these information, it has become possible to clearly show the relationship between observed variables that has been vague until now. We have analyzed multi-dimensional data by the tensor decomposition method using R language [2], and this time we performed a causal analysis of time series data in this language. This paper describes this analysis.

Commonly used data analysis methods include multiple regression analysis without latent variables, and Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA) with latent variables [3]. In the relationship of observed information, there may be a causal relationship between observed variables. There are Graphical Modeling (GM) and Structural Equation Model (SEM) as analysis methods to clarify the causal relationship in this case [4].

We applied SEM to the data obtained from the sleep monitoring sensor to model the relationship between the observed variables and analyze the causal relationship. In Japan, there is a serious shortage of care workers in elderly facilities, which causes problems such as excessive labor burden on workers and reduced service to residents [5]. For example, if a resident leaves the bed without going to bed at night, the caregiver on duty will be busy responding, reducing the time needed to care for other residents. With regard to this problem, we thought that if the resident could get enough sleep, the workload of the care worker could be reduced.

Therefore, the purpose of this study was to find conditions for good sleep by analyzing the causal relationship between environmental data and depth of sleep obtained from care support devices used in elderly care facilities. As a result, it is expected that residents will have a healthy life by ensuring sufficient sleep, and care workers will be able to reduce the labor burden because the number of visits to the resident will be appropriate. In this paper, we constructed a model that expresses the causal relationship between environmental variables obtained from sleep monitoring sensors and sleep levels.

This data analysis was performed using the statistical analysis language R [6]. SEM was adopted as the method of causal analysis, and it was implemented using the lavaan package for R language [7]. The model description grammar used in this package is simple and versatile enough to represent various models [8].

II. Selecting Data for Causal Analysis

We obtained sensor data for several people in nursing homes using a sleep monitoring sensor system "Mamoruno" [9] manufactured by ASD Inc. and analyzed this data. By attaching this device to a bed in a nursing facility, the temperature, humidity, atmospheric pressure, illuminance, biological information (heart rate, respiratory rate, sleep level), and bed in or out state (getting up, lying on, leaving) are automatically measured and uploaded to the cloud. There are two measurement modes, one-minute interval mode and five-minute interval mode, and data for one month at either interval is stored on the cloud.

The data used in this analysis are seven observation valuables: room temperature (*tm*), humidity (*hm*),

Manuscript received March 27, 2020; revised August 23, 2020.

doi: 10.18178/ijeee.8.3.58-62

barometric pressure (*at*), illuminance (*il*), respiratory rate one minute before (*rr1*), heart rate one minute before (*hr1*), and sleep level (*ss*), which were measured at five-minute intervals, where the names in parentheses are observation variable names. The variable *ss* takes four values from level 1 to level 4 and the other observed variables are positive real numbers.

The time series data used in the analysis is 2926 points (about 10 days) excluding those with missing values. Only the analysis results for an average person among the data obtained from three persons are shown this time. However, it should be noted here that the same configuration models were obtained for the other two people.

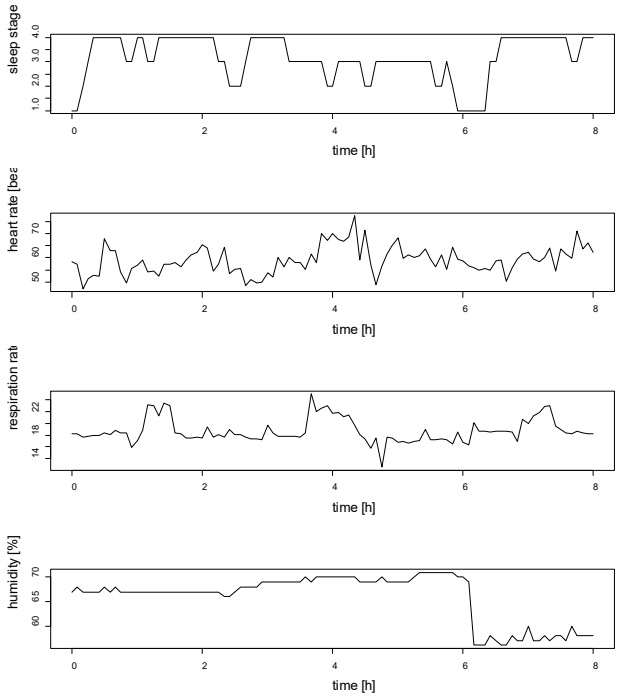

Figure 1. Time series data of several observed variables.

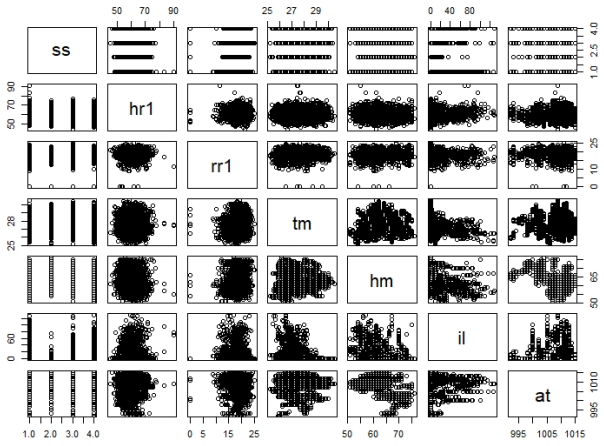

Figure 2. Scatter plot matrix representation of observed data.

Fig. 1 shows excerpts of 480 minutes (8 hours) of time series data for several observed variables. It can be seen from the figure that the values of *ss*, *rr1* and *hr1* change greatly, but *hm* has little change and switches from a high value to a low value. Although *rr1* has a value of 0, it may be a sign of sleep apnea syndrome, so it is not treated as an outlier. Next, in order to see the relationship between the observed values of seven variables, the correlation diagram matrix is shown in Fig. 2. From this figure, there is no strong correlation between each variable, and it is difficult to evaluate the correlation between *ss* and each variable.

As mentioned above, no clear relationship was found between sleep level ss and the other six observed variables, but first we performed causal analysis using all seven variables. However, as a result, a model suitable for the data could not be built. For reference, AGFI $= 0.936$ and RMSE $= 0.103$ in terms of the index described later. Therefore, we decided to reduce the variables used in the model. Stepwise selection method [10] was used to select the variables, and the observed variables to be adopted were determined while monitoring the Akaike's Information Criterion (AIC). This criterion is calculated by the following formula [11]:

$$\mathrm{AIC} = \mathrm{n} \cdot \log\left(\frac{\mathrm{RSS}}{\mathrm{n}}\right) + 2\mathrm{k} \tag{1}$$

where *n* is the number of samples, *k* is the number of explanatory variables, and *RSS* is the residual sum of squares of fitted model.

TABLE I. TRACE OF VARIABLE SELECTION

Method	Selected variables	AIC
Variable reduction method (see top to bottom)	*hr1, rr1, tm, hm, il, at*	775.55
	hr1, rr1, tm, hm, il	774.22
	hr1, rr1, hm, il	773.84
Variable addition / reduction method (see bottom to top)	*il, hr1, rr1*	780.23
	il, hr1	817.61
	il	829.55
	none	990.38

Actually, the variable selection was performed using the step function of R language, with the objective variable as *ss*. The results show that AIC is minimized when *ss* is expressed using four variables *il*, *hm*, *hr1*, and *rr1*, as shown in Table I. Based on this result, we decided to conduct analysis using five observation variables including the objective variable.

III. INTRODUCTION OF LATENT VARIABLES INTO MODELS

To represent the construct behind the above five variables, we decided to introduce latent variables into the causal model. In order to determine the number of these latent variables, the eigenvalues of the correlation

coefficient matrix between the observed variables were calculated and represented in the scree plot as shown in Fig. 3. From this figure, the number of latent variables is set to 2 because the attenuation after the second eigenvalue is gradual. This number of latent variables also meets the Kaiser-Guttman rule [12], which give the number of eigenvalues with a value greater than 1 (including nearly 1 here).

From the above, we introduced two latent variables f1 and f2, and assumed that each represents the indoor environment and vital signs as structural concepts. In addition, the observed variables to be linked to latent variables are illuminance (il) and humidity (hm) for f1, and heart rate (hr1) and respiration rate (rr1) for f2.

From the above, we introduced two latent variables *f1* and *f2*, and assumed that each represents the indoor environment and vital signs as structural concepts. In addition, the observed variables to be linked to latent variables are illuminance (*il*) and humidity (*hm*) for *f1*, and heart rate (*hr1*) and respiration rate (*rr1*) for *f2*.

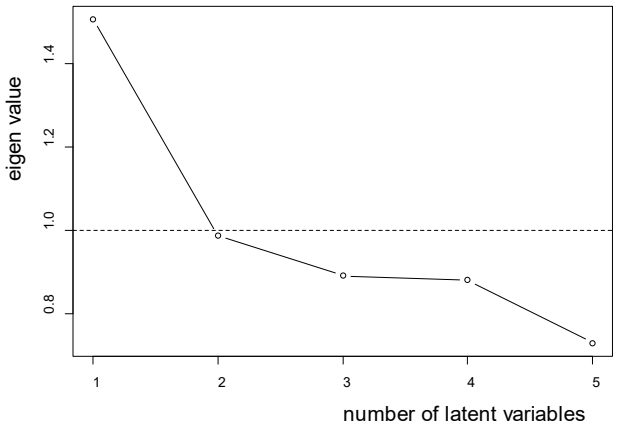

Figure 3. Scree plot of eigenvalues for correlation matrix.

IV. Causal Model Creation

There are Multiple Indicator models, Multiple Indicator Multiple Causes (MIMIC) models, Partial Least Squares (PLS) models, etc. to express causal relationships [13]. In this paper, we would like to analyze how latent variables and observed variables affect sleep level (*ss*), so the MIMIC model and PLS model are considered appropriate. Therefore, we propose the model shown in Fig. 4 using the features of these models. Such a diagram is called a path diagram.

In this model, it is assumed that *f1* and *f2* are orthogonal to each other because the indoor environment (*f1*) is unlikely to affect vital signs (*f2*) and vice versa. The measurement equation and the structural equation for this model are described as follows:

$$\begin{cases} \mathrm{f1} = \alpha_1 \cdot \mathrm{il} + \alpha_2 \cdot \mathrm{hm} \\ \mathrm{f2} = \alpha_3 \cdot \mathrm{hr1} + \alpha_4 \cdot \mathrm{rr1} \\ \mathrm{ss} = \gamma_1 \cdot \mathrm{f1} + \gamma_2 \cdot \mathrm{f2} + \mathrm{e_1} \end{cases} \quad (2)$$

In Eq. (2), the equations that describe how the observed variables affect the latent variable are called the structural equation (the first and second equations), and the equation that describes the opposite effect is called the measurement equation (the third equation). By solving this equation, the path coefficients (γ_1, γ_2 and α_1 *to* α_4) and the error variable (e_1) can be obtained. Note that *f1* and *f2* are endogenous variables but not accompanied by error variables. This is because in the PLS model, in general, the latent variable that is an endogenous variable placed on the left side of the equation is defined by the exogenous variables on the right side. That is to say, the indoor environment (*f1*) is regarded as a construct defined by illuminance (*il*) and humidity (*hm*), and similarly, vital sign (*f2*) are regarded as a construct defined by heart rate (*hr1*) and respiratory rate (*rr1*). Hence, neither *f1* nor *f2* is accompanied by an error variable.

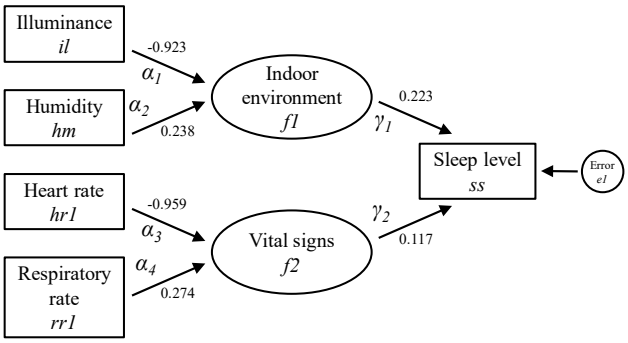

Figure 4. Path diagram of proposal model.

```
# SEM analysis of proposed model
# standardized data of selected variables should be set to "data.dat"
# evaluation measure of fitness : GFI, AGFI, RMSEA, CFI, SRMR, AIC, BIC

library(lavaan)

# model : proposed model in Fig. 4
model <- '
        f1 ~ il + hm
        f2 ~ hr1 + rr1
        f1 + f2 =~ ss
        f1 ~~ 0*f1
        f2 ~~ 0*f2
        f1 ~~ 0*f2
'
fit <- lavaan::sem(model, data=data.dat, orthogonal=T, fixed.x=T, std.lv=F)
summary(fit, standardized=T)
fitMeasures(fit, fit.measures = c("gfi","agfi","rmsea","cfi","srmr","aic","bic"))
```

Figure 5. Lavaan script describing proposed model.

Equation (2) was described in lavaan syntax, and parameters were estimated using the sem function in the lavaan package. Maximum Likelihood (ML) estimation was used as the estimation method. Note that the observation data used for the analysis was standardized one. This R script is shown in Fig. 5. Table II shows the standardized parameter estimates. In this table, the path coefficients α_1 and α_3 have negative values. This means that when the illuminance il becomes strong (bright), the value of the indoor environment f1 decreases, and conversely, the value of f1 increases when il becomes weak (dark). Similarly, when the heart rate hr1 increases (early), the vital sign value f2 decreases, and when hr1

decreases (slow), f2 increases. Usually, in a path diagram, the larger the pass coefficient value, the greater the causality. Fig. 4 also shows the obtained pass coefficient values.

TABLE II. ESTIMATED STANDARDIZED VALUES OF PARAMETERS

α_1	α_2	α_3	α_4	γ_1	γ_2	Variance of *e1*
-0.923	0.238	-0.959	0.274	0.223	0.117	0.928

TABLE III. GOODNESS-OF-FIT INDICES VALUES OF PROPOSED MODEL

GFI	AGFI	CFI	RMSEA	SRMR
1.000	1.000	1.000	0.000	0.000

Table III shows the goodness-of-fit indices [14] of the proposed model. The GFI (Goodness of Fit Index), AGFI (Adjusted GFI) and CFI (Comparative Fit Index) are indices that evaluate the accuracy of the analysis from the viewpoint of the explanatory rate of the model with respect to the variance of the observed variables. Each of these values takes a value between 0 and 1 and is considered as a good fit if it is greater than 0.95. These indices are all 1 in the table because the degree of freedom of the proposed model is 0. The RMSEA (Root Mean Square Error of Approximation) and SRMR (Standardized Root Mean Square Residual) are the evaluation indices based on the magnitude of deviation between actual data and the predicted value obtained from the model. If each index values is less than 0.05, the model is considered a good fit. In the table, these values are 0 because the degree of freedom is 0 as same as below.

V. INTERPRETATION OF ANALYSIS RESULTS

Based on the path coefficients obtained as a result of the analysis, the path diagram of the proposed model in Figure 4 is interpreted. The indoor environment *f1* is defined by illuminance *il* and humidity *hm*, and their pass coefficient values are -0.923 and 0.238 respectively, indicating that the indoor environment is highly dependent on illuminance. The vital sign *f2* is defined by heart rate *hr1* and respiration rate *rr1*, and these values are -0.595 and 0.274, so it can be seen that the vital sign depends on heart rate. Since the influence on the sleep level *ss* is 0.223 from the indoor environment *f1* and 0.117 from the vital sign *f2*, it can be said that the sleep level depends on the indoor environment and the vital sign, and the influence of the former is greater than the latter in this person. The overall effect of the influence of each observation variable *il*, *hm*, *hr1* and *rr1* on *ss* is estimated as following equation:

$$\begin{cases} \text{total effect of il to ss} = -0.923 \times 0.223 = -0.206 \\ \text{total effect of hm to ss} = -0.053 \\ \text{total effect of hr1 to ss} = -0.959 \times 0.117 = -0.112 \\ \text{total effect of rr1 to ss} = 0.274 \times 0.117 = 0.032 \end{cases} \quad (3)$$

From this estimation result, it can be seen that the illuminance *il* has the most impact on sleep levels *ss*, heart rate *hr1* has the next largest impact, and both are negative values. In other words, darkening the room will give the best sleep level and lowering the heart rate is the second most effective factor.

VI. CONCLUSIONS

By reducing the number of observed variables from the monitoring sensor that affect sleep levels using the stepwise selection method, we were able to analyze the causality between them by structural equation modeling. From the analysis results, it was found that the indoor environment and the vital signs have a significant effect on sleep, and that taking deep sleep requires darkening the room and lowering heart rate. We analyzed the data for three people, including the other two, and confirmed that the model with the same configuration as that presented in Section IV can be applicable with high goodness-of-fit indices to everyone. Although the detailed values of the pass coefficients vary depending on each data, these values were almost the same tendency for at least two persons as the person analyzed in this paper. In the future, we would like to analyze more people and confirm the general validity of this model including gender and age differences.

In addition, this time, by using the lavaan of R language, we were able to confirm that it is easy to describe structural equations and that causal analysis can be carried out relatively easily. In data analysis using R language, we believe that causal analysis using the lavaan package should spread. Furthermore, the model proposed this time cannot reveal the process of deepening sleep levels. Therefore, in the future, we would like to clarify this process by using, for example, a growth curve model.

ACKNOWLEDGMENT

This work was supported by the JSPS KAKENHI (Grants-in Aid for Scientific Research) under Grant No. JP18K11596.

REFERENCES

[1] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mob. Netw. Appl.*, vol. 19, no. 2, pp. 1-39, 2014.
[2] A. Ishida, U. Aibara, J. Murakami, N. Yamamoto, S. Saito, T. Izumi, and N. Kano, "Analysis of rehabilitation data by multi-dimensional principal component analysis method using the statistical software R," *Adv. Mater. Res.*, vol. 823, pp. 650-656, 2013.
[3] A. G. Yong and S. Pearce, "A beginner's guide to factor analysis," *Tutor. Quant. Methods Psychol.*, vol. 9, no. 2, pp. 79-94, 2013.
[4] M. Drton and M. H. Maathuis, "Structure learning in graphical modeling," *Annu. Rev. Statist. Appl.*, vol. 4, pp. 365-393, 2017.
[5] C. A. Morgan, "Demographic crisis in Japan: Why Japan might open its doors to foreign home health-care aides," *Pac. Rim L. & Policy J.*, vol. 10, no. 3, pp. 749-779, 2001.
[6] R Core Team. (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. [Online]. Available: https://www.R-project.org/
[7] Y. Rosseel, "Lavaan: An R package for structural equation modeling," *J. Statist. Soft.*, vol. 48, no. 2, pp. 1-36, 2012.

[8] J. B. Grace. (2013). Basic Lavaan Syntax Guide. [Online]. Available: http://www.structuralequations.com/resources/Basic_lavaan_Syntax_Guide_Aug1_2013.pdf
[9] ASD Inc. (2019). Mamoruno. [Online]. Available: http://mamoruno.miel.care/
[10] J. M. Wagner and D. G. Shimshak, "Stepwise selection of variables in data envelopment analysis: Procedures and managerial perspectives," *Eur. J. Oper. Res.*, vol. 180, pp. 57-67, 2007.
[11] S. Hu. (Feb. 2012). Akaike Information Criterion. [Online]. Available: http://www.researchgate.net/profile/Shuhua_Hu/publication/267201163_Akaike_Information_Criterion/links/599f662aa6fdccf5941f894b/Akaike-Information-Criterion.pdf
[12] H. F. Golino and S. Epskamp, "Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research," *PloS One*, vol. 2, no. 6, p. e0174035, 2017.
[13] M. Tenenhaus, V. E. Vinzi, Y. M. Chatelin, and C. Lauro, "PLS path modeling," *Comput. Stat. & Data Anal.*, vol. 48, no. 1, pp. 159-205, 2005.
[14] K. Schermelleh-Engel, H. Moosbrugger, and H. Müller, "Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures," *Methods Psych. Res.*, vol. 8, pp. 23-74, 2003.

Nobuhiro Oishi received the M.E. from Toyohashi University of Technology, Japan, in 1988 and Ph. D. in engineering from Kyushu Institute of Technology, Japan, in 2004.
He is currently a professor and the dean of the Department of Information, Communication and Electronic Engineering, Kumamoto College, National Institute of Technology, Japan. His research area of interests includes numerical calculation, statistical analysis, and causal analysis of multidimensional data.
Prof. Oishi is a member of the Physical Society of Japan (JPS).

Naoki Yamamoto received the Ph.D. in engineering from Kyushu Institute of Technology, Japan, in 2001.
He is currently a professor of the Department of Human-Oriented Information Systems Engineering, Kumamoto College, National Institute of Technology, Japan. His research interests are in the area of multidimensional data analysis and numerical calculation.
Prof. Yamamoto is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and Kyushu Society for Engineering Education (KSEE).

Akio Ishida received the M.S. and Ph.D. in science from Kumamoto University, Japan, in 2010 and 2014.
He is currently an assistant professor of the Faculty of Liberal Arts, Kumamoto College, National Institute of Technology, Japan. His research interests are in the area of multi-dimensional data analysis and numerical calculation.

Jun Murakami received the Ph.D. in engineering from Toyohashi University of Technology, Japan, in 2000.
He is currently a professor of the Department of Human-Oriented Information Systems Engineering, the Director of Library, Kumamoto College, National Institute of Technology, Japan. His research area of interests includes statistical analysis, numerical calculation, and digital signal processing.
Prof. Murakami is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Information Processing Society of Japan (IPSJ), and the Human Interface Society Japan (HISJ).