Principal Component Analysis and Binary Classification of Office Room Occupancy

Gursimarjit Singh Saini
Student Id: 40220457
Github Link: https://github.com/GursimarSaini/INSE6220Project

Abstract*—Principal Component Analysis (PCA) is a fast and flexible unsupervised dimensionality reduction method that transforms a high dimensional data with correlated features to low dimensional data with uncorrelated features. This report illustrates the use of PCA when applied to the office room occupancy data set attributes to classify if the room is occupied. Determination of occupancy detection in a room can led to considerable energy savings in modern smart home/buildings. **~~Still some part remaining~~*

Keywords—Principal Component Analysis (PCA), Binary Classification,

I. INTRODUCTION

With the decreasing price of sensors and the availability of reasonable computational power for automation systems, determining occupancy is a very promising way to lowering energy usage in buildings through appropriate control of HVAC and lighting systems. Threat of climate adversity has made it important for the production of most energy efficient products [1]. The precise detection of occupancy in buildings has lately been projected to save energy in the range of 30 to 42 percent. When occupancy data was employed as an input for HVAC control algorithms, it resulted in energy savings of 37 percent without sacrificing indoor climate and between 29 and 80 percent in another [2]. When privacy matters are considered, it makes much more sense to use sensors for getting accurate occupant numbers than to use cameras. Determining building inhabitants behavior and Security are another two applications for occupancy detection.

The research [2] used data from light, temperature, humidity, and CO2 sensors to detect occupancy, as well as a digital camera to determine ground occupancy for data labelling. This data set created for occupancy detection is used for this study.

Working with a huge dataset as what used in this study is usually perplexing and laborious. To make the research easier, the approach must incorporate dimension reduction, while preserving the majority of the data variability. PCA is generally used for such tasks [3], which is described and implemented in Section 3, after giving a brief Exploratory Data Analysis in Section 2. Section 4, throws light on applicable Classifiers with brief explanations and discussion of the most promising model that helps in detection for this process in Section 5. Section 6 closes with a summary of the findings

II. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis refers to the crucial process of conducting preliminary investigations on data in order to uncover patterns, spot anomalies, test hypotheses, and validate assumptions using summary statistics and graphical representations. Here it's done in three parts, first by giving a brief introduction for the raw data set, then discussing the cleaning process and description for used data set. At last, checking distribution and outliers with Box Plots and correlation of points with Correlation Matrix.

A. Raw Data Set Description

As mentioned in the study [2], the following variables were observed in an office space with approximate dimensions of 5.85m, 3.50m, 3.53m (W D H): timestamp, temperature, humidity, light, and CO2 levels. The study collects the data using a microcontroller. It was linked to a ZigBee radio, which was used to relay the data to a recording station. A digital camera was utilized to assess whether or not the room was inhabited. Every minute, the camera time stamped an image, which was then manually examined to identify the data. The humidity ratio is another additional variable in the data model, calculated as:

$$W = 0.622 \times \frac{p_w}{p - p_w}$$

The data was collected in February in Mons, Belgium, during the winter. The room was heated by hot water radiators, which kept the temperature above 19 degrees Celsius. The models are tested for data sets with the office door open and closed in order to estimate the difference in occupancy detection accuracy provided by the models. The measurements were obtained at 14-second intervals/3-4 times every minute, and then averaged for that minute.

B. Data Cleaning

All three data sets were missing column name for their first column, which was named as "id" and then dropped in data pre-processing. For the purpose of this study, only one test data 1 and training data will be used.

C. Used Data Set Description

The description for these two datasets is summarized in **Figure 1**. No duplicate rows or NaN values were found for both of the datasets. And all the values are floating point numbers, except the column "Occupancy" which is labelled with int values, 0 and 1.

Data Set	Number of Observations	Data Class Distribution	
		0 (non occupied)	1 (occupied)
Training	8143 of 7 variables	0.79	0.21
Testing 1	2665 of 7 variables	0.64	0.36

Fig. 1. Data Set Description

The distribution of class can be seen with the bar plot in **Figure 2**. As said, the label 1 represents that the room was occupied and Class 0 for unoccupied rooms.

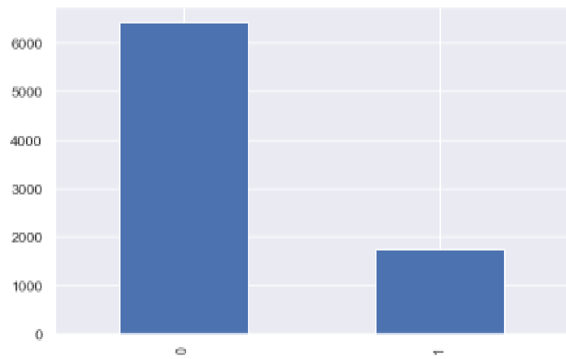

Fig. 2. Class Distribution

D. *Data Analysis*

	Temperature	Humidity	Light	CO2	HumidityRatio
count	8.143000e+03	8.143000e+03	8.143000e+03	8.143000e+03	8.143000e+03
mean	7.818326e-16	4.467615e-16	2.233807e-16	-1.884775e-16	1.116904e-16
std	1.000061e+00	1.000061e+00	1.000061e+00	1.000061e+00	1.000061e+00
min	-1.592248e+00	-1.624791e+00	-6.137261e-01	-6.165933e-01	-1.394355e+00
25%	-9.038502e-01	-1.000115e+00	-6.137261e-01	-5.330748e-01	-9.201475e-01
50%	-2.252867e-01	8.877312e-02	-6.137261e-01	-4.869407e-01	-7.243751e-02
75%	7.581387e-01	8.681862e-01	7.027469e-01	1.027265e-01	5.742528e-01
max	2.518470e+00	2.420232e+00	7.326619e+00	4.524170e+00	3.066492e+00

Fig. 3. Descriptive Statistics

Standardization is put into use to adjust each input variable independently by removing the mean, and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one [4]. After standardization, the Descriptive Statistics metrics can be seen in **Figure 3**. If it's not done, then covariances for larger number ranges will be much higher.

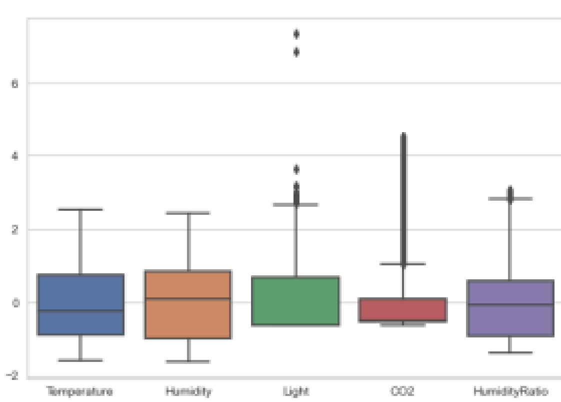

Fig. 4. Box Plot

Putting this standardize matrix in a Box Plot gives us the idea about the distribution of the data, measures of central tendency and spread. In **Figure 4**, we can see that, all the data attributes are positively skewed to an extent. Data is centered around 0, because of standardization and variability is minimum for the same reason. Outliers are present for 3 of the 5 attributes and all of them are on the skewed side of whiskers.

To understand the relationship among the attributes, Correlation Matrix and Pair Plot are used. It's evident that almost all of the parameters are positively correlated with each other's. There is no presence of any variable with negative correlation with all of the others variables. CO2 has significant correlation with rest of variables over others as seen in **Figure 5.**

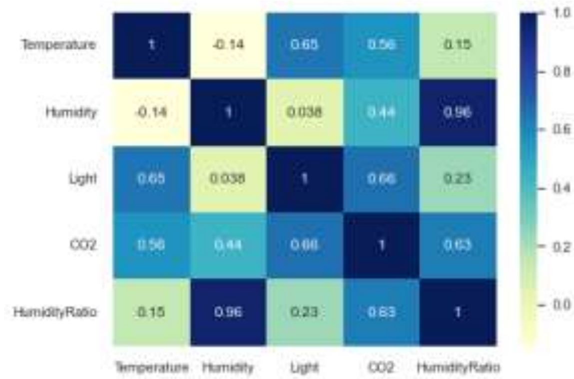

Fig. 5. Correlation Matrix

This can also be supported with the help of pairplot in **Figure** 6 The strong positive correlations are determined with increasing line. Whereas, weak correlations form clusters rather than an increasing line in pair plot.

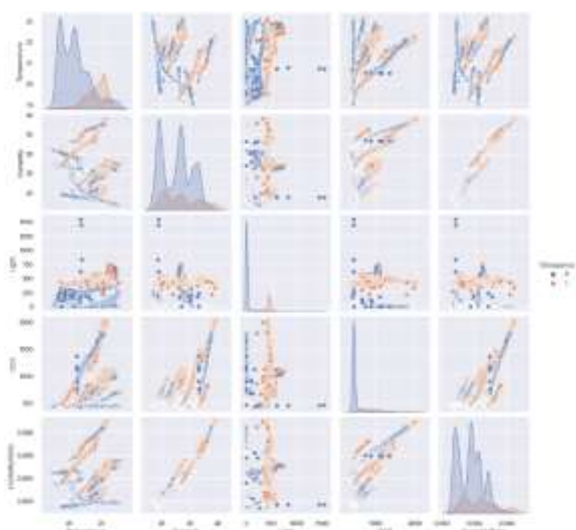

Fig. 6. Pair Plot

The multiple correlations among data set parameters, is the reason why PCA is implemented to get un-correlated data.

III. PRINCIPAL COMPONENT ANALYSIS

PCA is typically used to reduce the dimensionality of data while retaining as much of the information present in the original data as feasible. It does this by examining a data table including observations characterized by numerous dependent variables that are, in general, inter-correlated. Its purpose is to extract the key information from the data table and express this information as a set of new orthogonal variables known as principal components. Simply put, PCA is important so as to:

- Extract the most relevant information from the data table,
- Compress the size of the data set by maintaining just the most significant information,
- Simplify the data set description, and
- Simplify the data set description, and

PCA output comprises of coefficients that specify the linear combinations used to obtain the new variables (PC loadings) as well as the new variables themselves (PCs). The first PC must have the greatest potential variance. The second component is calculated with the constraint of being orthogonal to the first component and having the greatest possible inertia. The other components are calculated in the same way. [6]

A. *Implementation of PCA in steps*

We need to make sure, data should be structured in a typical matrix format, with n rows of samples and p columns of variables. There should be no missing values: each variable should have a value for each sample, which can be zero [7]. The steps needed for PCA is as follows [3].

1. Centering the dataset: For this step we subtract the mean of a variable from all of its values, so that the data stays centered on the origin of main components, because any algorithm which is based on distance computations are affected a lot if the data used isn't normalized/centralized [8].

$$Y = H \times X \tag{1}$$

	Temperature	Humidity	Light	CO2	HumidityRatio
0	23.18	27.2720	426.0	721.25	0.004793
1	23.15	27.2675	429.5	714.00	0.004783
2	23.15	27.2450	426.0	713.50	0.004779
3	23.15	27.2000	426.0	708.25	0.004772
4	23.10	27.2000	426.0	704.50	0.004757

Fig. 7. Before Standardization

For our study, standardization is preferred over centering to avoid precision error when range of variables is different.

	Temperature	Humidity	Light	CO2	HumidityRatio
0	2.518470	0.278526	1.573763	0.364948	1.091757
1	2.488967	0.277713	1.591735	0.341881	1.080555
2	2.488967	0.273645	1.573763	0.340290	1.075888
3	2.488967	0.265508	1.573763	0.323587	1.066555
4	2.439796	0.265508	1.573763	0.311655	1.049523

Fig. 8. After Standardization

1. Calculate Covariance Matrix: Covariance matrix of size $p \times p$ is produced to check if data set has correlated features and also so as eigen decomposition can be applied to the data [3].

$$S = \frac{1}{n-1} \times Y^T \times Y \tag{2}$$

	0	1	2	3	4
0	1.000123	-0.141777	0.650022	0.559963	0.151780
1	-0.141777	1.000123	0.037833	0.439077	0.955315
2	0.650022	0.037833	1.000123	0.664104	0.230449
3	0.559963	0.439077	0.664104	1.000123	0.626633
4	0.151780	0.955315	0.230449	0.626633	1.000123

Fig. 9. Covariance Matrix

2. Eigen Decomposition: Eigenvectors and eigenvalues are obtained from the Covariance matrix S with eigen decomposition. Eigenvectors give direction of Principal Components with variance of PCs dented with eigenvalues and are given by [3]:

$$S = A \times \lambda \times A^T \tag{3}$$

where A is a $p \times p$ orthogonal eigenvector matrix and is a diagonal eigenvalue matrix. In our study, we have a 5×5 eigenvector matrix and a 1×5 column matrix of eigenvalues, both given as:

$$A = \begin{bmatrix} 0.343856 & 0.535864 & -0.713374 & 0.225382 & -0.186850 \\ 0.395664 & -0.574111 & 0.009249 & 0.226966 & -0.679888 \\ 0.414149 & 0.444612 & 0.665424 & 0.433177 & 0.019235 \\ 0.550070 & 0.120106 & 0.110938 & -0.817265 & -0.052622 \\ 0.501117 & -0.413692 & -0.189517 & 0.205245 & 0.706895 \end{bmatrix}$$

Fig. 10. Eigenvectors

$$\lambda = \begin{bmatrix} 2.736860 \\ 1.699679 \\ 0.348872 \\ 0.214393 \\ 0.000809 \end{bmatrix}$$

Fig. 11. Eigenvalues

3. Principal Components: The last step yields a $n \times p$ matrix Z, with its rows giving the observed values and columns representing the PCs as given in **Fig. 14**. Given by equation [3]:

$$Z = Y \times A \tag{4}$$

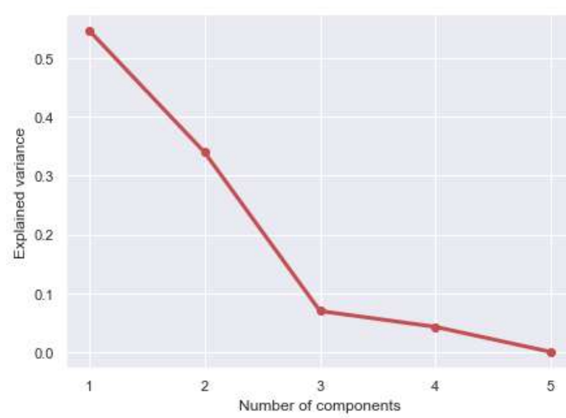

Fig. 12. Scree Plot

The variance of j^{th}PC is given as following [3]:

$$l_j = \frac{\lambda_j}{\Sigma_j^p \lambda_j} \times 100\%, for\ j = 1, \ldots, p \qquad (5)$$

where λ_j gives the variance of j^{th} PC. Both Scree/Elbow plots can be used to get an idea of how many PCs are needed to represent the variance present in the data. In this study, we found out that variance accounted for first PC is l_1= 54.7% and by 2nd PC it is l_2= 33.9% and that by 3rd PC is l_3= 6.97%. The elbow joint in the scree plot, shows a bend at PC number 3, that is also supported with Pareto Chart. So, it's safe to assume that dimensions of eigenvector or Z components can be reduced to 3.

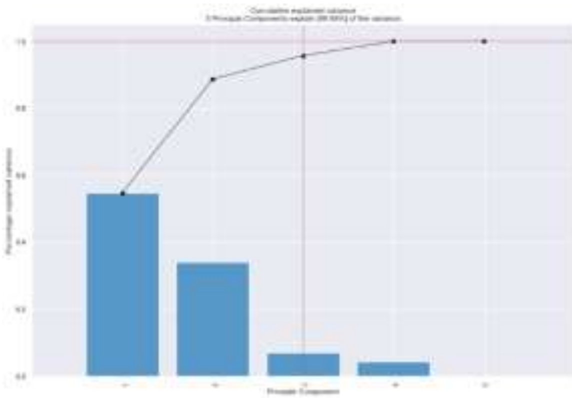

Fig. 13. Pareto Chart

The **Fig. 14** gives a subset of the PCs, 8143×5 as 8143×3 because first 3 PCs explain 99.98% of the whole variance of dataset. The first PC component Z1 is given by [3]:

	PC1	PC2	PC3
0	2.375810	1.481548	-0.913234
1	2.354485	1.476060	-0.880672
2	2.342218	1.472145	-0.891961
3	2.325134	1.478671	-0.892121
4	2.293128	1.457935	-0.855139
...	...	...	...
8138	2.879268	-0.897206	0.468321
8139	2.866476	-0.877949	0.472498
8140	2.925637	-0.874514	0.433447
8141	2.990828	-0.895826	0.435690
8142	3.011957	-0.852825	0.485749

Fig. 14. Z scores

$$Z_1 = 0.55007 \times X_4 + 0.501116 \times X_5 + 0.414149 \times X_3 + 0.395664 \times X_2 + 0.343856 \times X_1$$

X_4 (CO_2), X_5 (Humidity Ratio), X_3 (Light), X_2 (Humidity), and X_1 (Temperature), contribute most to the 1st PC, respectively. For Z_2, we got

$$Z_2 = 0.574111 \times X_2 + 0.535864 \times X_1 + 0.444612 \times X_3 - 0.413692 \times X_5 + 0.120106 \times X_4$$

For both first two principal components we don't have any attribute contributing negligibly to them. But in case of, third principal component, X_2 doesn't affect it that much, so effective PC will be:

$$Z_3 = -0.713374 \times X_1 + 0.665424 \times X_3 - 0.189517 \times X_5 + 0.110938 \times X_4$$

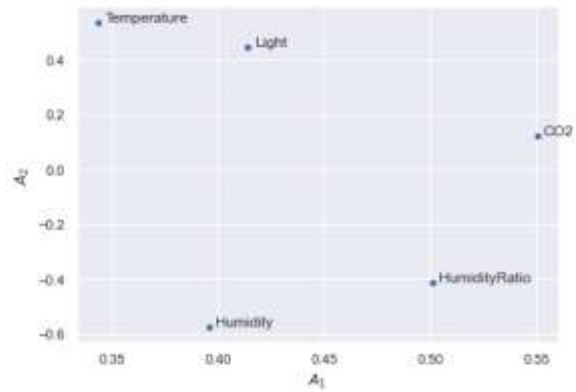

Fig. 15. PC Coefficient Plot

This same could be verified with the help of PC coefficient Plot as in **Fig. 15.** Temperature and Humidity lies in the lower range for the first PC, along with CO_2 and Humidity Ratio being the most important factors for consideration, with light being somewhere in the middle of first two and latter two. Whereas all the bottom 3 contributors of 1st PC got to be at the top for 2nd PC and CO_2 being the least.

Biplot gives the same information as of **Fig. 15.** The angles between the vectors (rows of eigenvector matrix) and axes (representing the first two PCs) gives the contribution of variables to PCs [3], i.e., the vector with smallest angle with the axe contributes most to that axe/PC. Also, each observation is scattered as a point in the plot.

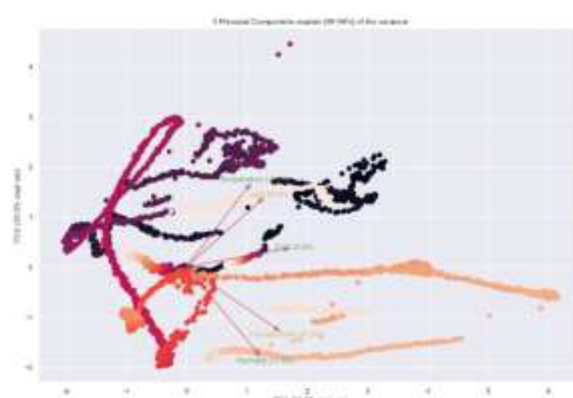

Fig. 16. 2D Biplot

This same could be represented for 3 PCs with help of 3d Biplot as shown in Figure 17.

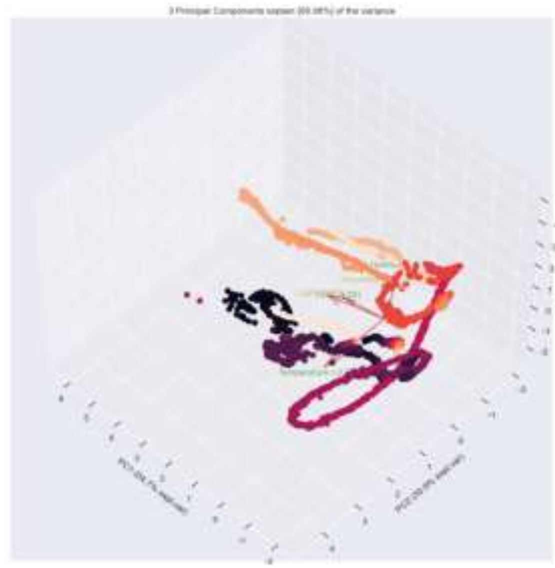

Fig. 17. 3D Biplot

dfffgfgdgg

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

B. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

C. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m2" or "webers per square meter", not "webers/m2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm3", not "cc". (*bullet list*)

D. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.

- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

IV. Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Authors and Affiliations

The template is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

1) For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

2) For papers with less than six authors: To change the default, adjust the template as follows.

a) Selection: Highlight all author and affiliation lines.

b) Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

c) Deletion: Delete the author and affiliation lines for the extra authors.

B. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named "Heading 1", "Heading 2", "Heading 3", and "Heading 4" are prescribed.

C. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I. TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

a. Sample of a Table footnote. (*Table footnote*)

Fig. 1. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

Acknowledgment *(Heading 5)*

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] Candanedo Ibarra, Luis & Feldheim, Veronique. (2015). Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Energy and Buildings. 112. 10.1016/j.enbuild.2015.11.071. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] K. Elissa, "Title of paper if known," unpublished.

[4] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[6] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.