Hong Kong CIE sky classification and prediction by accessible weather data and trees-based methods

S Lou, D H W Li1, J C Lam
Building Energy Research Group, Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR, China

E-mail: bcdanny@cityu.edu.hk

Abstract. Solar irradiance and daylight illuminance are important for solar energy and daylighting designs. Recently, the International Commission of Illuminance (CIE) adopted a range of sky conditions to represent the possible sky distributions which are crucial to the estimation of solar irradiance and daylight illuminance on vertical building facades. The important issue would be whether the sky conditions are correctly identified by the accessible variables. Previously, a number of climatic parameters including sky luminance distributions, vertical solar irradiance and sky illuminance were proposed for the CIE sky classification. However, such data are not always available. This paper proposes an approach based on the readily accessible data that systematically recorded by the local meteorological station for many years. The performance was evaluated using measured vertical solar irradiance and illuminance. The results show that the proposed approach is reliable for sky classification.

1. Introduction
Daylight data are essential to the designs of daylight-linked artificial lighting controls in energy-efficient buildings [1] by reducing the lighting power consumption and interior heat gain. For daylighting applications, the skylight is always preferred since it has less potential to cause glare. The scattered daylight, however, can be a non-uniform light source whose intensity in terms of luminance varies over the whole sky hemisphere [2]. The anisotropy of skylight can cause significant differences of the daylight via fenestrations in different orientations. It is important to predict the sky luminance distributions. Since both sky luminance and solar radiance are of similar characteristics, the same mathematical approaches for determining daylight illuminance can also be applied to solar irradiance [3].

Sky distributions can be influenced by factors including solar position, turbidity and cloud amount. A usual approach to estimate the distribution is by the corresponding sky condition, since specified sky conditions induced from climatic factors at approaching levels are featured by similar luminance (and radiance) distributions. Generally, the sky conditions can be interpreted into overcast, partly cloudy and clear skies. In 2003, the International Commission on Illumination (CIE) adopted 15 standard skies proposed by Kittler and his colleagues [4]. Each of the CIE Standard Sky has its well defined luminance distribution pattern by mathematical expressions which have been reviewed by measurements in a number of world-wide regions [5-7]. Many studies reported that the 15 CIE

1 Address for correspondence: D H W Li, Building Energy Research Group, Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR, China. E-mail: bcdanny@cityu.edu.hk.
Standard Skies provide good and comprehensive frameworks of the actual skylight luminance distributions in the world.

The usual instrument for measuring sky luminance distribution is by means of a sky scanner that records the luminance levels of 145 points of the whole sky dome. However, the sky-scanner is costly and maintenance intensive while the 145 points are not recorded simultaneously. Sky classifications based on different climatic indices were also proposed [3, 8, 9]. Sky luminance distributions particularly for the whole sky are seldom available for the sites under consideration or exist for very limited periods. It is expected that an approach based on the readily accessible horizontal solar radiations and available meteorological data would be more appropriate. This paper studies the CIE sky classification using measured data in our measuring station [10] and the Hong Kong Observatory (HKO) [11]. A sky classification approach was developed by the algorithm of Classification Tree (C_Tree). The performance of the proposed method was evaluated against measured data. The findings and design implications are discussed.

2. CIE Standard Skies

The set of 15 CIE Standard Skies [12] predicts the ratio between luminance at any point of the sky hemisphere \( L_{\alpha\phi} \) to that at the zenith point \( L_Z \). Since daylight is the visual part of solar radiation, the same mathematical models can also be used for the radiance distribution calculation. The standard formula of the ratio \( l_{\alpha\phi} \) is the product of a relative gradation expression \( \phi(Z)/\phi(0) \) and a relative scattering indicatrix function \( f(\chi)/f(Z_s) \) as expressed in Equations 1 to 3.

\[
l_{\alpha\phi} = \frac{L_{\alpha\phi}}{L_Z} = \frac{f(\chi)\phi(Z)}{f(Z_s)\phi(0)}
\]

\[
\frac{\phi(Z)}{\phi(0)} = \frac{1+a\exp(b/\cos Z)}{1+a\exp b}
\]

\[
\frac{f(\chi)}{f(Z_s)} = \frac{1+c\left[\exp(d\chi)-\exp(d\pi/2)\right]+\cos^2\chi}{1+c\left[\exp(dZ_s)-\exp(d\pi/2)\right]+\cos^2Z_s}
\]

Where \( Z \) is the zenith angle of the sky point under consideration (rad), \( Z = \pi/2 - \alpha \); \( Z_s \) is zenith angle of the sun; \( a, b, c, d \) and \( e \) are coefficients, the values of which can be adjusted to describe the 15 CIE Standard Skies from heavy overcast to clear skies; \( \chi \) is the shortest angular distance between the sky element and the solar disc, i.e. scattering angle (rad), which can be calculated as:

\[
\chi = \arccos\left(\cos Z_s \cdot \cos Z + \sin Z_s \cdot \sin Z \cdot \cos(\varphi - \varphi_s)\right)
\]

Where \( \varphi_s \) and \( \varphi \) are the azimuth angle of the sun and the sky point under consideration, respectively. Detailed angles are given in Figure 1.
The gradation function (Equation 1) describes the sky luminance variation between horizon and zenith. In terms of overcast skies, greater luminance can be found at zenith. The trend is opposite for clear skies. The indicatrix function (Equation 2) peaks at the position of sun and drops rapidly when the scattering angle increases. By adjusting the value of coefficients from $a$ to $e$, 15 groups of equations can be generated to denote the luminance distribution of a range of sky conditions, including 5 clear, 5 partly cloudy and 5 clear skies. The coefficients of the skies and their features are listed in Table A1 (Appendix A).

3. Sky classification by luminance

The most straightforward and comprehensive approach of sky classification is by comparing the luminance distribution pattern between the 15 CIE Standard Skies with measurement. The data of luminance distribution are usually acquired by a sky scanner which measures the luminance of 145 circular sky patches ($L_i$, cd m$^{-2}$, $i = 1$ to 145). As mentioned by Tregenza [13], using the measured zenith luminance ($L_Z$, cd m$^{-2}$) as the basis of a Standard Sky can cause uncertainty, because the measured $L_Z$ may be biased due to the direct sunlight when the sun is close to zenith point. It is proposed to normalize $L_i$ and the ratio of the CIE Standard Skies ($l_i$) by the corresponding horizontal sky-diffuse. The normalized distributions of the measurement ($L_{i,nor}$) and the CIE Standard Skies ($l_{i,nor}$) are illustrated in Equations 5 and 6.

$$L_{i,nor} = \frac{L_i}{V_h}$$

$$l_{i,nor} = \frac{l_i}{v_h}$$

Where $V_h$ is the horizontal illuminance from all the 145 sky luminance by measurement, $v_h$ is the horizontal illuminance from all the 145 sky luminance under individual CIE Standard Skies with $L_Z$ set as one. Detailed calculations of the $V_h$ and $v_h$ by $L_i$ and $l_i$ were illustrated by Tregenza [13]. The Standard Sky is identified as the one with the minimum root mean square error between $L_{i,nor}$ and $l_{i,nor}$, as given in Equation 7.

$$rms = \sqrt{\frac{1}{145} \sum_{i=1}^{145} (L_{i,nor} - l_{i,nor})^2}$$

Figure 2 shows the frequency of occurrence (FOC) of sky classified by the aforementioned approach. The sky luminance data were 10-minute readings recorded at the City University of Hong Kong (CityUHK) in 2004. Details of the station can be found in our previous work [10]. In total, 21,089 sets of data were employed, among which 37.4, 39.8 and 22.8% were identified as overcast, partly cloudy
and clear skies, respectively. The most representative skies were Skies 1, 8 and 13 with the FOC of 20.1%, 31.9 and 15.7%, respectively. The results are consistent with a previous study by others [14].

**Figure 2.** The frequency of occurrence of CIE Standard Skies in Hong Kong.

### 4. Sky classification by meteorological data

The classification by scanning the luminance distribution is perhaps the most comprehensive method but it has a number of limitations [3]. For instance, the long-term luminance measurements are rarely available while the meteorological features of a place are normally determined by the data recordings for 25 to 30 years [15]. To develop a long-term database of the CIE Standard Sky according to the accessible HKO measurements, the algorithm of Classification Tree was used to correlate the meteorological data sets with the sky classification results in previous section (Figure 1).

#### 4.1. The Classification Tree

The algorithm of Classification Tree (C_Tree) was first developed by Breiman et al [16]. The algorithm categorizes data sets with a number of variables ($\text{Var}_1, \text{Var}_2, \ldots$) into various groups by a sequence of binary partitions (i.e. splits). Figure 3 illustrates a split that separates a single group with $N_i$ data sets to two child groups according to the value of the selected $j$th variable ($\text{Var}_j$). The variable and its value for the split are selected to maximally reduce the size-weighted impurity of the CIE Skies in the child groups. By quantifying the impurity as the Gini index ($G$) defined in Equation 8, the reduction of size-weighted impurity can be quantified as Equation 9.

\[
G = 1 - \sum_k [p(k)]^2 \tag{8}
\]

\[
\Delta G = G_{i1}N_{i1} - G_{i+1}N_{i+1} - G_{i+2}N_{i+2} \tag{9}
\]

Where $p(k)$ is the fraction of datasets with sky type $k$ that reaches a node. $G$ is positive for a node that has more than a single type of the CIE Standard Sky. The classification by a single split repeats until the data in the node have purified a particular sky type, or the tree structure being established reaches the designated complexity.
Figure 3. The topography of a single split in Classification Tree.

The complexity of the C_Tree can be controlled by the minimum leaf size \( L_{MIN} \) [17], which stops the splitting when the size of data sets in its child node is less than the setting. A lower \( L_{MIN} \) allows the algorithm to do more detailed classifications by a more complexed C_Tree which can improve the purity of sky type in a signal group. However, the model developed with too small \( L_{MIN} \) could be complicated to interpret and even over-fitted, which may give poor performance for new data sets. It is preferably to optimize the \( L_{MIN} \) in a case by case manner. In present study, the C_Trees were trained with \( L_{MIN} \) from 10 to 1,000.

4.2. The Development Classification Tree

The development of C_Tree that correlates the types of CIE Standard Sky with the meteorological data (variables) was implemented by the Machine Learning Toolbox of MATLAB 2015b [18]. The variables summarized in Table 1 were enrolled without a pre-selection since there was little knowledge on the importance of the variables and the algorithm itself can eliminate the unnecessary ones. The clearness index \( K_t \) and the diffuse fraction \( K_d \) are defined in Equations 10 and 11.

\[
K = \frac{E_G}{E_E} \quad (10)
\]

\[
K_d = \frac{E_D}{E_G} \quad (11)
\]

Where \( E_G \) is the horizontal global irradiance, \( E_E \) is the horizontal extraterritorial irradiance and \( E_D \) is the horizontal diffuse irradiance. The irradiance data were measured in our measuring station (CityUHK) every 10 minutes while the data of other variables were recorded by the HKO in 1-hour interval. The data from HKO were converted into 10-min by a shape-preserving piecewise cubic interpolation [19] based on the values of neighbouring points in MATLAB 2015b [20].

Table 1. Input variables for the development of Classification Tree (C_Tree)

| Parameter                  | Notation | Unit | Parameter                  | Notation | Unit |
|----------------------------|----------|------|----------------------------|----------|------|
| Solar altitude angle       | \( \alpha_s \) | °     | Relative humidity          | \( rh \)  | %    |
| Clearness index            | \( K_t \) |       | Visibility                 | \( vis \) | m    |
| Diffuse fraction           | \( K \)  |       | Wind speed                 | \( v \)   | m/s  |
| Sunshine duration hour     | \( H \)   | hour | Mean sea level pressure    | \( P \)   | Pa   |
| Air temperature            | \( T_a \) | °C    | Cloud amount               | \( cld \) | %    |

To reduce the burden of computation, simplify the structure of C_Tree and improve reliability of the results, the framework given in Figure 4 were adopted to develop the approach for sky
classification by C_Tree. As given in the figure, C_Tree 0 was developed to categorize the overcast (Skies 1 to 5), partly cloudy (Skies 6 to 10) and clear (Skies 11 to 15) before 3 C_Trees 1 to 3 were established to further specify the 15 Skies.

Figure 4. The framework of proposed sky classification approach by Classification Trees (C_Tree)

To evaluate the expected performance of every C_Tree with different $L_{MIN}$, we repeated a 5-fold cross validation [21] for 1,000 times which can reduce the uncertainty of random training data selection in sensible time and RAM consumption. The results include the median and range of percentage of skies being misclassified ($\%\text{Mis}$) for each C_Tree, and the average number of splits with $L_{MIN}$ being set from 10 to 500 as shown in Figures 5 to 8 for C_Trees 0 to 3.

For the development of all C_Trees, the $\%\text{Mis}$ drops initially then rises gently as $L_{MIN}$ increases. The C_Trees are not powerful enough when $L_{MIN}$ is too small and over-fitted when $L_{MIN}$ is too large. The number of splits, however, reduces drastically when $L_{MIN}$ increases indicating that less complexed trees are formed with the setting of lower $L_{MIN}$. Although the minimum $\%\text{Mis}$ can be identified for each C_Tree, the corresponding number of splits may be overwhelming. Setting $L_{MIN}$ slightly greater can significantly lower down the number of splits and tree complexity, which makes the C_Trees more interpretable though the quality of prediction may be sacrificed slightly. By compromising the tree complexity against the performance, the $L_{MIN}$ were set 130, 150, 425 and 100 for the C_Trees 0, 1, 2 and 3, respectively. The far larger $\%\text{RMSE}$ of C_Tree 1 compared with other C_Trees (i.e. C_Tree 0, C_Tree 2 and C_Tree 3) implies that it would be quite difficult to further identify the inconspicuous differences between the classified overcast skies (i.e. Skies 1 to 5) via meteorological data, especially based on the horizontal solar irradiance. The numbers of splits were controlled in a sensible level, which were around 27, 12, 3 and 7 respectively for C_Trees 0, 1, 2 and 3. The C_Trees developed by the designated settings are given in Figures B1 to B4 in Appendix B.

Figure 5. The percentage of misclassification ($\%\text{Mis}$) of the CIE Overcast, Cloudy and Clear Skies by C_Tree 0 with different minimum leaf size ($L_{MIN}$).
Figure 6. The percentage of misclassification (%Mis) of Sky 1 to 5 by C_Tree 1 with different minimum leaf size ($L_{MIN}$)

Figure 7. The percentage of misclassification (%Mis) of Sky 6 to 10 by C_Tree 2 with different minimum leaf size ($L_{MIN}$)

Figure 8. The percentage of misclassification (%Mis) of Sky 11 to 15 by C_Tree 3 with different minimum leaf size $L_{MIN}$
5. Performance evaluation

For the development of vertical building integrated photovoltaic panels and evaluation of solar heat gains and daylighting designs, accurate vertical global solar irradiance and illuminance data are required. To evaluate whether the skies are correctly identified, the global components for the vertical planes facing the 4 cardinal directions were modelled based on the sky types classified by the two approaches [3, 22]. The global solar irradiance and illuminance incident on a vertical surface \( E_{VG} \) is the sum of direct beam \( E_{VB} \), sky-diffuse \( E_{VD} \) and ground reflected components \( E_{VR} \), which can be expressed as Equation 12.

\[
E_{VG} = E_{VB} + E_{VD} + E_{VR}
\]  

(12)

The calculation of \( E_{VB} \) is based on the solar altitude \( \alpha_s \) angle, azimuth angle \( \varphi_s \) and the surface azimuth angle \( \varphi_s \). For the estimation of \( E_{VR} \), it is assumed that the vertical plane receives half of the irradiance and illuminance being reflected isotropically from the ground. Denoting the ground reflectance \( \rho \) of 0.2 is a reasonable assumption in dense urban environment and adopted by a number of researchers [23-25], the 2 components can be calculated by Equations 13 and 14.

\[
E_{VB} = \left[ \left( E_G - E_D \right) \sin \alpha_s \right] \cos \alpha_s \cos (\varphi_s - \varphi_s)
\]  

(13)

\[
E_{VR} = \rho E_G / 2
\]  

(14)

The numerical method was used to compute \( E_{VD} \) [3]. The performance was assessed by the root mean square error with respect to the average of the simultaneously measured global solar irradiance and illuminance (%RMSE) and the results are summarized in Table 2. It is found the %RMSE of vertical irradiance and illuminance on vertical planes facing all directions are less than 20% and 25% respectively. The %RMSE of proposed approach is slightly greater than the luminance based approach and the difference is less than 2% for the four cardinal orientations. The findings show that the proposed approach can give reliable prediction of the CIE Standard Skies and the vertical illuminance and irradiance.

| %RMSE        | Vertical irradiance | Vertical illuminance |
|--------------|---------------------|----------------------|
|              | N       | E     | S   | W   | N   | E   | S   | W   |
| Luminance by sky scanner | 15%    | 17%   | 13% | 14% | 23% | 11% | 12% | 14% |
| Proposed     | 16%    | 19%   | 14% | 15% | 24% | 13% | 13% | 16% |

6. Conclusions

An approach that classifies the sky conditions of CIE Standard Skies was developed by the algorithm of Classification Tree, which correlates the readily accessible meteorological data with individual CIE Standard Skies. The CIE Skies being employed to develop the model were identified by the luminance data of a sky scanner. The solar radiation, daylight illuminance and sky luminance data were acquired in the City University of Hong Kong and the meteorological data were measured by the Hong Kong Observatory in 2004. The skies identified by the proposed approach give good estimations of irradiance and illuminance on vertical planes with the root mean square error of less than 25%. The findings show that the proposed approach is appropriate to classify the CIE Standard Sky in Hong Kong using readily accessible meteorological data. Future work will be conducted to reduce the complexity of C_Tree structure and sensitivity analysis will be implemented to introduce more important variables and remove those rarely used in the model.
Acknowledgements
The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 7004543). Siwei LOU was supported by a City University of Hong Kong studentship.

Appendix A
The coefficients and features of the 15 CIE Standard Skies are given in Table A1

| Sky | Gradation | For indicatrix | Type of sky       | General classification | Description of luminance distribution                  |
|-----|-----------|----------------|-------------------|------------------------|-------------------------------------------------------|
| 1   | 4.0       | -0.7           | 0                 | Overcast               | Steep gradation, azimuthal uniformity                  |
| 2   | 4.0       | -0.7           | 2                 | Overcast               | Steep gradation, slight brighten around sun           |
| 3   | 1.1       | -0.8           | 0                 | Overcast               | Gentle gradation, azimuthal uniformity                 |
| 4   | 1.1       | -0.8           | 0                 | Overcast               | Gentle gradation, slight brighten around sun          |
| 5   | 0         | -1             | 0                 | Overcast               | Uniform                                              |
| 6   | 0         | -1             | 2                 | Partly cloudy          | No gradation, slight brighten around sun              |
| 7   | 0         | -1             | 5                 | Partly cloudy          | No gradation, brighter circumsolar region            |
| 8   | 0         | -1             | 10                | Partly cloudy          | No gradation, distinct solar corona                  |
| 9   | -1        | -0.55          | 2                 | Partly cloudy          | Sun position shaded/obscured                         |
| 10  | -1        | -0.55          | 5                 | Partly cloudy          | Brighter circumsolar region                          |
| 11  | -1        | -0.55          | 10                | Clear                  | White-blue sky, distinct solar corona                |
| 12  | -1        | -0.32          | 10                | Clear                  | Low turbidity                                        |
| 13  | -1        | -0.32          | 16                | Clear                  | Polluted atmosphere                                  |
| 14  | -1        | -0.15          | 16                | Clear                  | Cloudless turbid sky, broad sun corona               |
| 15  | -1        | -0.15          | 24                | Clear                  | White-blue turbid sky, wide sun corona               |
Appendix B
Structure of the Classification Trees (C_Trees 0 to 3) are given in Figures B1 to B4

Figure B1. The C_Tree that determines the Overcast, Partly Cloudy and Clear Skies (i.e. C_Tree 0)

Figure B2. The C_Tree that determines the Sky 1 to 5 for the classified Overcast Skies (i.e. C_Tree 1)

Figure B3. The C_Tree that determines the Sky 6 to 10 for the classified Cloudy Skies (i.e. C_Tree 2)

Figure B4. The C_Tree that determines the Sky 6 to 11 for the classified Clear Skies (i.e. C_Tree 3)
References

[1] Li D H W, Cheung K L, Wong S L and Lam T N T 2010 An analysis of energy-efficient light fittings and lighting controls Appl. Energy 87 558–67
[2] Perez R, Seals R and Michalsky J 1993 All-weather model for sky luminance distribution—Preliminary configuration and validation Sol. Energy 50 235–45
[3] Li D H W, Chau T C and Wan K K W 2014 A review of the CIE general sky classification approaches Renew. Sust. Energ. Rev. 31 563–74
[4] Kittler R, Darula S and Pereze R 1998 A set of standard skies (Bratislava: Polygrafia SAC)
[5] Li D H W, Lau C C S and Lam J C 2003 A study of 15 sky luminance patterns against Hong Kong data Architect. Sci. Rev. 46 61–8
[6] Tregenza PR 1999 Standard skies for maritime climates Lighting Res. Technol. 3197–106
[7] Wittkopf S K and Soon L K 2007 Analysing sky luminance scans and predicting frequent sky patterns in Singapore. Lighting Res. Technol. 39 31–51
[8] Kittler R and Darula S 2002 Parametric definition of the daylight climate Renew. Energy 26 177–87
[9] Li D H W and Lau C C S and Lam J C 2004 Standard skies classification using common climatic parameters J. Sol. Energy Eng. 126 957–64
[10] Li D H W and Tang H L 2008 Standard skies classification in Hong Kong J. Atmos. Sol.-Terr. Phy. 70 1222–30
[11] Li D H W, Lou S W and Lam J C 2015 An analysis of global, direct and diffuse solar radiation Energy Procedia 75 388-93
[12] CIE 2003 Spatial Distribution of daylight - CIE Standard General Sky (Vienna: CIE Central Bureau)
[13] Tregenza P R 2004 Analysing sky luminance scans to obtain frequency distributions of CIE standard general skies Lighting Res. Technol. 36 271–81
[14] Ng E, Cheng V, Gadi A, Mu J, Lee M and Gadi A 2007 Defining standard skies for Hong Kong. Build. Environ. 42 866–76
[15] Chan A L S, Chow T T, Fong S K F and Lin J Z 2006 Generation of a typical meteorological year for Hong Kong Energy Convers. Manag. 47 87–96
[16] Breiman L, Friedman J, Olshen R A and Stone CJ 1984 Classification and regression trees (Belmont, CA: CRC press)
[17] Khoshgoftaar T M, Allen E B, Deng J 2001 Controlling overfitting in software quality models: experiments with regression trees and classification Proc. 7th Int. Software Metrics Symp. METRICS (Los Alamitos, CA: IEEE Computer Society) p 190
[18] Mathworks 2015 Machine learning toolbox user's guide (Natick, MA: The MathWorks, Inc.)
[19] Fritsch F N and Carlson R E 1980 Monotone piecewise cubic interpolation SIAM J. Numer. Anal. 17 238–46
[20] Mathworks 2015 Curve fitting toolbox user's guide (Natick, MA: The MathWorks, Inc.)
[21] Kohavi R 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection Int. joint conf. artificial intelligence (IJCAI) 2 1137-43
[22] Ho J, Ng E and Chan P 2015 Predicting the hourly Hong Kong representative sky from Typical Meteorological Year data for dynamic daylighting simulation Lighting Res. Technol. 47 730–9
[23] Li D H W, Cheung G H W, Cheung K L and Lam J C 2009 Simple method for determining daylight illuminance in a heavily obstructed environment Build. Environ. 44 1074-80
[24] Ng E 2001 A simplified daylighting design tool for high-density urban residential buildings Lighting Res. Technol 33 259–72
[25] Tregenza P R 1995 Mean daylight illuminance in rooms facing sunlit streets Build. Environ. 30 83-9